Coevolution of transcription factors and their binding sites in sequence space

by

Roshan Prizak
Klosterneuburg, Austria

March, 2019

A thesis presented to the Graduate School of the
Institute of Science and Technology Austria, Klosterneuburg, Austria
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Institute of Science and Technology

The thesis of Roshan Prizak,

titled

Coevolution of transcription factors and their binding sites in sequence space,

is approved by:

Supervisor: Gašper Tkačik, IST Austria, Klosterneuburg, Austria

Signature: _____

Co-supervisor: Nick Barton, IST Austria, Klosterneuburg, Austria

Signature: _____

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

Signature: _____

Committee Member: Michael Lässig, University of Cologne, Cologne, Germany

Signature: _____

Defense Chair: Beatriz Vicoso, IST Austria, Klosterneuburg, Austria

Signature: _____

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

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.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Roshan Prizak
March, 2019

To the three people I love the most.

Ammu,
for her friendship and her patience.

Amma,
for her unconditional support and her strength.

Pappa,
for his kindness and inspiration. We miss you.

Abstract

Transcription factors, by binding to specific sequences on the DNA, control the precise spatio-temporal expression of genes inside a cell. However, this specificity is limited, leading to frequent incorrect binding of transcription factors that might have deleterious consequences on the cell. By constructing a biophysical model of TF-DNA binding in the context of gene regulation, I will first explore how regulatory constraints can strongly shape the distribution of a population in sequence space. Then, by directly linking this to a picture of multiple types of transcription factors performing their functions simultaneously inside the cell, I will explore the extent of regulatory crosstalk – incorrect binding interactions between transcription factors and binding sites that lead to erroneous regulatory states – and understand the constraints this places on the design of regulatory systems. I will then develop a generic theoretical framework to investigate the coevolution of multiple transcription factors and multiple binding sites, in the context of a gene regulatory network that performs a certain function. As a particular tractable version of this problem, I will consider the evolution of two transcription factors when they transmit upstream signals to downstream target genes. Specifically, I will describe the evolutionary steady states and the evolutionary pathways involved, along with their timescales, of a system that initially undergoes a transcription factor duplication event. To connect this important theoretical model to the prominent biological event of transcription factor duplication giving rise to paralogous families, I will then describe a bioinformatics analysis of C2H2 Zn-finger transcription factors, a major family in humans, and focus on the patterns of evolution that paralogs have undergone in their various protein domains in the recent past.

About the Author

Roshan Prizak completed B. Tech and M. Tech in Electrical Engineering, with a specialization in Communication and Signal Processing at the Indian Institute of Technology Bombay in Mumbai, before joining IST Austria in September 2013. His main research interests include the biophysics and evolution of gene regulation, and its connection to the spatial organization of DNA inside the nucleus. During undergraduate studies, he worked on various research projects as part of internships, some of which were published in journals – social organization of the Asian elephant using graph theory at JNCASR Bangalore, opinion formation model investigating cyclic dominance on graphs with Dr. Thilo Gross and Dr. Güven Demirel (MPIPKS, Dresden), which has been published in European Physical Journal B, quantitative genetics of maternal, and grand-maternal inheritance in the Hoyle lab at University of Surrey, results of which have been published in Ecology and Evolution, and Functional Ecology, respectively. During his PhD studies, Roshan has also presented his research results in the $[BC]^2$ Basel Computation Biology conference in Basel in 2017 and published his work in two Nature Communication papers.

List of Publications

The thesis Chapter 2 is based on the third publication listed below - [Friedlander *et al.*, 2016], and Chapter 4 is based on the second publication listed below - [Friedlander *et al.*, 2017]. Tamar Friedlander has made important contributions to both these chapters. She performed some of the calculations in Section 2.2, Section 2.3, and Section 4.9, and also the calculation in Eq. 4.18.

1. Carballo-Pacheco M*, Desponds J*, Gavrilchenko T*, Mayer A*, **Prizak R***, Reddy G*, Nemenman I, Mora T. *Receptor crosstalk improves concentration sensing of multiple ligands.* **arXiv preprint** arXiv:1810.04589 (2018) (* - alphabetical)

2. Friedlander T*, **Prizak R***, Barton NH, Tkačik G - *Evolution of new regulatory functions on biophysically realistic fitness landscapes.* **Nature Communications**, 8(1):216 (2017). (* - equal contribution)

3. Friedlander T, **Prizak R**, Guet CC, Barton NH, Tkačik G - *Intrinsic limits to gene regulation by global crosstalk.* **Nature Communications**, 7:12307 (2016).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNA Deoxyribonucleic acid

RNA Ribonucleic acid

mRNA Messenger RNA

TF Transcription factor

DBD DNA-binding domain

BS Binding site

CRE Cis-regulatory element

GRN Gene regulatory network

RNAP RNA Polymerase II

TSS Transcription start site

PWM Position Weight Matrix

PCM Position Count Matrix

ZNF Zn-finger TF without KRAB domain

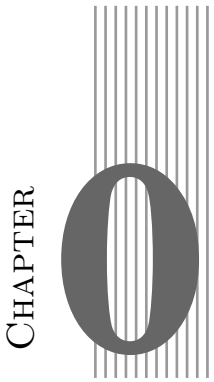
KZNF Zn-finger TF with KRAB domain

TE Transposable element

ChIP Chromatin Immunoprecipitation

SMART Simple Modular Architecture Research Tool

PAML Phylogenetic Analysis by Maximum Likelihood



Introduction

Biological systems use information stored in DNA to ensure that various necessary cellular processes run in the correct spatiotemporal context. This information contains not just the "how-to-build" instructions to manufacture the set of necessary proteins from individual amino acids, but also "regulatory" information about when, where and how much of each protein to produce [Jacob and Monod, 1961; Britten and Davidson, 1969; François and Hakim, 2004]. While the how-to-build instructions are mostly contained in the coding sequences of genes (some RNAs also act as building blocks), regulatory information is contained in both non-coding DNA and a few proteins involved in gene regulation. Such an important set of proteins called transcription factors (TFs) are one of the primary molecular actors in the decoding of this regulatory information from non-coding DNA [Lambert *et al.*, 2018; Vaquerizas *et al.*, 2009; Mitchell and Tjian, 1989]. TFs bind to specific regions - binding sites (BSs) in the cis-regulatory elements (CREs) on non-coding DNA - and based on their binding activities, control the spatiotemporal expression of nearby target genes to ensure that the required proteins are produced by the cell in various cellular contexts. This fundamental process, called transcription, is a crucial step in the conversion of information on the DNA into proteins, which in turn have the capability to perform various functions. Some of these proteins are in turn themselves transcription factors and hence form an inter-connected network of genes called gene regulatory networks (GRNs) that together read out genetic information from DNA to assist in cellular programs [Sauka-Spengler and Bronner-Fraser, 2008; Olson, 2006; Zhou *et al.*, 2007]. It is important to remember that such a network picture is a caricature, but it offers a useful conceptual framework to understand transcription factor based gene regulation.

In a GRN, different TFs and other target genes, which sometimes themselves code for TFs, are associated with the nodes of the network. The network is then defined by

connections between pairs of nodes, called edges, symbolizing an interaction between the genes associated with the nodes; for instance, a TF node activating another gene. The basic molecular interaction that forms these edges is the binding of a transcription factor protein to the binding site in the CRE corresponding to a target gene. In prokaryotes, such binding sites are located in close proximity to the target gene's transcription start site (TSS) on the linear DNA, and this stretch of DNA close to the gene's TSS is called a promoter. The mechanism of activation is usually via a direct interaction between an activator TF and RNA Polymerase (RNAP), and repression is achieved by a physical exclusion of RNAP or other potential activator TFs through the competitive binding of a repressor TF. On the other hand, the mechanisms behind transcription regulation and the role of TF-DNA binding in eukaryotes are vastly more complex and not well understood [Coulon *et al.*, 2013]. Apart from the gene-proximal promoters, eukaryotes also have distal regulatory elements called enhancers that are sometimes located as far as a few Mbp (mega basepairs) away from the TSS [Blackwood and Kadonaga, 1998; Pennacchio *et al.*, 2013; Maston *et al.*, 2006]. Both promoters and enhancers contain a large number of binding sites of various types of transcription factors, and it is believed that TFs, via their joint binding activities on enhancers, influence transcription in a combinatorial fashion [Spitz and Furlong, 2012]. Further, the question of how the binding activities of TFs on these distal enhancers combines forces with those on proximal promoters and thereby influences transcription, is largely open. It is especially challenging to explain how these regulatory elements interact at the typical long genomic distances of a few kilo basepairs to a few mega basepairs, bringing into focus the dynamics of these regulatory elements in the 3D space of the nucleus [Chen *et al.*, 2018; Lim *et al.*, 2018]. This general question of how the 3D spatial organization of DNA, and chromatin at large, inside the nucleus, interacts with transcriptional machinery is opening new avenues of research. Using a combination of new experimental methods like advanced imaging [Chen *et al.*, 2018; Lim *et al.*, 2018], chromosome conformation capture techniques [Lieberman-Aiden *et al.*, 2009] and concepts from physics like non-equilibrium statistical physics and phase separation [Hnisz *et al.*, 2017; Strom *et al.*, 2017], important headways are being made into answering these fundamental questions [Boehning *et al.*, 2018; Cho *et al.*, 2018].

While the physical principles inform us about how gene regulatory systems work, to understand why a particular gene regulatory network performs a particular function, we need to embed these biophysical models into an evolutionary framework to properly define such questions [Levo and Segal, 2014; Necsulea and Kaessmann, 2014; Villar *et al.*, 2014]. Cells and organisms have evolved different types of gene regulatory networks to respond in a myriad of situations. Constraints emerging from a combination of the underlying biophysics of transcription and the particular information processing needs of the cell in a signalling context dictate which network evolves in that context. Such ideas are now being put on a solid quantitative footing [Hillenbrand *et al.*, 2016; Tkačik and Bialek, 2016]. Insights from such an approach to understand the evolution of gene regulatory networks are more generally translatable to other biological systems involving molecular recognition between interaction partners. A few ubiquitous examples of molecular recognition that permeate life are, nucleotide interactions in DNA replica-

tion and repair, ligand-receptor interaction in signal sensing [Mora, 2015], protein-protein interactions [Chothia and Janin, 1975; Jones and Thornton, 1996], immune system recognition events [Akira *et al.*, 2006] and molecular self-assembly [Murugan *et al.*, 2015]. This is an important insight, given the biological cell is an extremely crowded and dynamic environment with many atoms and molecules of multiple types - sugars, proteins, ions, lipids, acids, etc. These molecules are moving around inside the cell and constantly interacting with each other, but however, only a few of these interactions – precise biochemical reactions – are successful in affecting the state of the cell. How do we reconcile the picture of a crowded cell with these precise schemes of biochemical reactions?

The functional fidelity of such systems, and therefore, ultimately, the fitness of the cells and organisms, depends on the specificity of such molecular recognition interactions embedded as precise biochemical reactions in a background of numerous interactions in the crowded cell. This specificity stems from the specific interactions between the underlying monomer units involving the recognition partners - via hydrogen bonds, electrostatic interactions, entropic forces, and other such molecular forces. These forces are not perfectly specific, giving rise to interactions between non-cognate monomer unit pairs. As cells have to process multiple molecular recognition events in parallel, a large number of non-cognate partners are typically present. This large excess of non-cognate partners and the limited specificity of monomer interactions inevitably result in a large number of incorrect recognition events between the wrong pairs of partners. This is called crosstalk [Friedlander *et al.*, 2016; Jacob and Monod, 1961; Britten and Davidson, 1969].

Crosstalk is a systemic property of biological systems whose quantification requires a bridging of the microscopic molecular picture of molecular recognition with a global systemic view of the molecular interaction network. Crosstalk is typically considered deleterious as incorrect recognition events can result in loss/alteration of important biological information. For instance, a wrong nucleotide base insertion during DNA replication can result in a mutation, a wrong antibody recognition event can result in autoimmune diseases, and a wrong TF-BS binding can result in erroneous activation of genes [Hahn *et al.*, 2003]. In this context, one important question is how the different possible underlying microscopic molecular pictures vary in the amount of crosstalk they produce, as organisms might have evolved molecular mechanisms that mitigate crosstalk. It has been suggested before that molecular mechanisms like kinetic proofreading and cooperativity result in reduced crosstalk [Bird, 1995; Todeschini *et al.*, 2014], but a rigorous quantitative framework has been lacking to properly investigate such claims.

To answer such questions in the case of transcriptional regulation, we developed a joint framework combining the biophysics of transcription and evolutionary dynamics. The basic interaction motif in gene regulatory networks is a TF-BS binding pair, which is my focus of investigation for a major part of this thesis. Transcription factor proteins are equipped with a DNA binding domain (DBD) – the amino acids which interact with nucleotides on the DNA via hydrogen bonds and electrostatic interactions. Specificity in TF-DNA binding arises from the specificity of the interaction between the amino

acids and the nucleotides, and hence, TFs, depending on the amino acid sequence of their DBD, have distinct DNA sequence binding profiles [Lambert *et al.*, 2018]. We use an equilibrium thermodynamic model of TF-BS binding based in sequence space that captures these biophysical properties of TF-DNA interaction [Ackers *et al.*, 1982; Von Hippel and Berg, 1986; Lynch and Hagner, 2015], as the biophysical framework for asking questions about crosstalk and evolution.

The framework we develop to understand TF-BS coevolution can be modified to understand other molecular recognition systems mentioned previously. A major determinant of the patterns of crosstalk and coevolutionary dynamics is the specificity per site, and how it is distributed over the set of interacting sites. Further, the molecular mechanisms behind the transfer of information from molecular recognition to downstream target cellular process take a crucial role in the models. In this thesis, we assume that TF-BS binding is enough to trigger transcription, in an equilibrium assumption. Non-equilibrium models - for instance, kinetic proofreading models, and richer models accommodating more states for the machinery behind molecular recognition and the associated cellular process - in the case of transcriptional regulation, models with different enhancer and promoter states, multi-step promoter architectures [Rieckh and Tkačik, 2014] etc., offer a possibility of investigating a broader range of biophysical models. This is the genotype-phenotype map. Finally, coevolutionary dynamics are also influenced by the phenotype-fitness map, which captures how the survival of the cell (or organism) depends on the set of phenotypes associated with the molecular process in question. Completing the picture, the genotype-phenotype-fitness map, or the fitness landscape, of a molecular-recognition system and the associated cellular process, captures both selective as well as mutational constraints via a combination of biophysical and evolutionary models, and the differences in these genotype-phenotype-fitness maps between various molecular recognition system result in different patterns of their evolution.

In Chapter 1, I will introduce in detail the biophysical framework of TF-DNA binding in sequence space that we use in answering the various questions of the thesis. We use a grand-canonical ensemble framework to describe the bound/unbound states of BSs inside the nucleus, by treating different TF species with different chemical potentials. Under this equilibrium assumption, we compute the equilibrium probability of a BS sequence being bound by a TF molecule under various TF concentrations. Then, I will also show how regulatory constraints on DNA sequence in the form of presence or absence of strong binding sites for various TFs, shape the structure of the sequence space.

In Chapter 2, I will quantify crosstalk in transcriptional regulation and investigate its dependence on various underlying parameters like the number of TFs, number of target genes, TFs' degree of specificity, binding site length and other biophysical parameters. I will also show how various molecular mechanistic strategies of transcription like combinatorial regulation, use of activators and repressors, result in different levels of crosstalk. To connect with data, I will also present a basic bioinformatics strategy to compute crosstalk in real organisms.

Then in Chapter 3, I will introduce a generic theoretical model that merges the biophys-

ical framework of TF-DNA binding in sequence space developed in Chapter 1 with an evolutionary framework. In this model, we also consider upstream signal statistics as well to investigate transcription in a broader signal-response framework. Such a setup allows us to ask evolutionary questions of TF-DNA systems.

In Chapter 4, I will describe a specific case of the generic TF-BS coevolution (from Chapter 3) that has only two TFs regulating a set of target genes. I will investigate the evolutionary dynamics following a TF duplication event and describe the steady states, evolutionary pathways and the corresponding timescales the system takes to acquire specialized TFs. I will also show that the fitness landscape is rugged in the presence of multiple target genes, and consequently results in slow evolutionary dynamics. Then I will show that "promiscuity-promoting" mutations, a novel mutation type that has been observed in a few experimental studies on TFs and protein-protein interactions, help the system escape the ruggedness of the fitness landscape, and result in fast specialization.

In Chapter 5, I will describe a preliminary bioinformatics analysis of the evolution of Zn-finger TFs in the human genome, one of the largest families of transcription factor in animals. I will illustrate how various genomic features of Zn-finger TFs might be related to each other, and probe their relationship between paralog TFs. After showing that computation of a single dN/dS ratio (depicting presence or absence of positive selection) across all sites of the protein damps any signal that might be present, I will describe various site-specific models of dN/dS computation. These inform us that Zn-finger TFs have undergone positive selection primarily at the amino acids responsible for their binding specificity to DNA. Further, I illustrate a possible coevolution between Zn-finger TFs with a KRAB domain and transposable elements, and show that KRAB-domain containing Zn-finger TFs might have undergone adaptation to bind new transposable elements.

Finally, in Chapter 6, I will summarize the main findings of this thesis in a concise manner. If you are in a hurry, and would like get an overall picture and a gist of the main results, I would advise you to read this chapter after introduction, and depending on further interest, refer back to individual chapters and go through them in detail.

Chapter 1

TF-DNA binding in sequence space

1.1 Introduction

Transcription factor proteins contain at least one DNA binding domain (DBD), comprising the amino acids which interact with nucleotides on the DNA via hydrogen bonds and other molecular forces like electrostatic and van der Waals interactions. It is these contacts that primarily confer sequence-specific binding properties to the TF. Specificity may also, to some extent, depend on the whole sequence or other amino acids, via protein conformation or a shape-based TF-DNA interaction. For a majority of the TFs, depending on the amino acid composition of the DBD, each TF has a specific preferred binding site (BS) sequence, called the consensus sequence, that the TF binds with the highest affinity [Levo and Segal, 2014; Najafabadi *et al.*, 2015]. This sequence-specific binding of TFs is what enables correct spatiotemporal expression patterns in the cell. By evolving TFs and "matching" BS sequences at the appropriate locations along the genome, the cell ensures that the right genes are activated in various contexts. While the biophysical mechanism behind this complex process of how TF binding is related to the expression of the target gene is still largely unclear, studies in prokaryotes have shown that a model assuming equilibrium binding of TF to the BS, and subsequent recruitment of RNAP to transcribe the target gene, predicts the expression levels very well. On the other hand, the scenario is much more complex in eukaryotes, and much less is known about the concerted action of various TFs in regulating the transcriptional status of a gene.

Inside a cell, there are molecules belonging to various different TF types, each involved in various aspects of the cell's transcriptional programs. In such a picture, an important aspect of such a regulatory system that requires multiple TFs to simultaneously find the correct "addresses" in the form of BS sequences, is the chance of finding an incorrect

address. Given that the cell has many transcription factors that recognize a whole range of short sequences (of about lengths between 5bp and 20bp typically), and that a large portion of the DNA does not contain functional BS sequences, this becomes a crucial problem that the cell has to address. The cell should make sure that not only incorrect activation of off-target genes is minimized, but should also make sure that BS sequences do not frequently arise by chance in the vast chunk of non-regulatory sequence that makes up the genome. This is related not just in preventing unnecessary sequestration of TF molecules on nonspecific sites, but is also related to genomic stability by heterochromatin maintenance. Some TFs can trigger the opening of heterochromatin, leading to a cascade of transposable element expression, for instance. Such a scenario would lead to combinatorial action of the heterochromatin-modifying TFs and the specific activating TFs, with possibly sharper induction curves.

In this chapter, after introducing the biophysical framework of TF-DNA binding that I use in this thesis, I will describe how constraints on a DNA sequence to be not regulatory (not contain any BSs) shapes the sequence space. The biophysical model will set the background for the various questions tackled in the subsequent chapters, and the exploration of non-regulatory sequence space will set the tone for tackling the related problem of transcriptional crosstalk, the topic of the next chapter.

1.2 Biophysical model

In our biophysical model of transcription, we consider TF-DNA binding to be at thermodynamic equilibrium and use a grand-canonical ensemble framework to describe the bound/unbound states of the BSs [Ackers *et al.*, 1982; Von Hippel and Berg, 1986; Lynch and Hagner, 2015]. Even though this equilibrium assumption is more suited to prokaryotes than eukaryotes, we consider both prokaryotic and eukaryotic TFs in our models. We compute the equilibrium probability of a BS sequence being bound by a TF molecule, which is a quantity that we will keep coming back to in various chapters of the thesis. This probability of a BS being bound is related to the level of expression of the target genes. We describe the TF by its consensus sequence and specify the affinity of the TF to different BS sequences by their binding energies as

$$E = \sum_{j=1}^{L} E^{(j)}, \tag{1.1}$$

where the TF binds to sequences with length L and $E^{(j)}$ is the contribution from the j^{th} position in the sequence. Such a definition assumes that different positions in the sequence contribute linearly towards the total binding energy. Given a TF with consensus sequence s^*, and a BS with sequence s, we have $E^{(j)} = \mathbf{PWM}_{s(j),j}$ where $\mathbf{PWM}$ is the position-weight matrix of the TF. This is a matrix of size $4 \times L$, with entries describing the energetic contribution of different nucleotides at each position of the binding site sequence. We have $\mathbf{PWM}_{s^*(j),j} = 0, \forall j$ to ensure that that the contribution of the

correct nucleotide (that in the consensus sequence) at each position is zero towards the binding energy. In the constant mismatch model, we assume that all other entries equal ϵ: $\mathbf{PWM}_{s(j)\neq s^*(j),j} = \epsilon, \forall j$. In this model, the binding energy between a TF with consensus sequence s^* and BS sequence s is given by $E = \epsilon d(s, s^*)$, where $d(s, s^*)$ is the Hamming distance between the sequences s and s^* [Ackers *et al.*, 1982; Von Hippel and Berg, 1986; Lynch and Hagner, 2015; Berg and von Hippel, 1987].

1.3 Grand-canonical ensemble

We consider the following situation of TFs inside the nucleus (though I refer to nucleus, the arguments apply to prokaryotes as well). In a nucleus of volume V, we have TFs from one species with copy number C, and each TF molecule occupying a volume v. Each TF binds to a particular BS (of length L) on the DNA with energy E, which depends on the TF consensus sequence s^*, the BS sequence s, and the position-weight matrix **PWM** of the TF. Each TF also binds to a random sequence of length L on the rest of the genome (of size G base pairs) with a nonspecific binding energy E_{ns}. An unbound BS has an energy of E_u.

We want to obtain μ, the chemical potential of this TF species, to use in a grand-canonical ensemble treatment to compute p_{on}, the equilibrium probability that the BS is bound by a TF molecule. We have different components in this system - TF molecules, BS sequence, random sequences on the rest of the DNA, and free solution of the nucleus. In the canonical treatment, we consider the whole nucleus, including all of the mentioned components, as the system, while in the grand-canonical ensemble, we consider the BS + any bound TF as system, and the rest of DNA + the nuclear free solution as the reservoir. We obtain the chemical potential as,

$$\mu(C) = \ln \frac{Z(C-1)}{Z(C)}, \tag{1.2}$$

where $Z(C)$ is the canonical partition for the reservoir. Intuitively, $\ln[Z(C)]$ corresponds to the free energy of the reservoir and this way of finding the chemical potential agrees with the relation between μ and free energy: chemical potential is defined as the rate of change of the free energy of a system with respect to the change in the number of molecules. In the reservoir, the C TF molecules are distributed between being bound nonspecifically to the rest of the genome and floating around in the free solution.

$$Z(C) = \sum_{C_1=0}^{C} \binom{G}{C_1} e^{(-C_1 E_{ns})} e^{-(G-C_1)E_u} \binom{N}{C-C_1} e^{-(C-C_1)E_{sol}}, \tag{1.3}$$

where C_1 molecules are bound somewhere on the DNA and the rest are in solution, $N = (V/v)$ is the number of boxes (of size v) comprising the free solution and E_{sol} is the

energy of a TF in free solution. Typically, $N \gg G \gg C$ which allows us to approximate using $\binom{N}{i} \approx \frac{N^i}{i!}$ when N is large and $N \gg i$,

$$Z(C) \approx e^{-GE_u} \sum_{C_1=0}^{C} \frac{G^{C_1}}{C_1!} e^{-C_1(E_{ns}-E_u)} \frac{N^{C-C_1}}{(C-C_1)!} e^{-(C-C_1)E_{sol}} \tag{1.4}$$

$$= \frac{e^{-GE_u}}{C!} \sum_{C_1=0}^{C} \binom{C}{C_1} \Big[Ge^{-(E_{ns}-E_u)}\Big]^{C_1} \Big[Ne^{-E_{sol}}\Big]^{(C-C_1)} \tag{1.5}$$

$$= e^{-GE_u} \frac{1}{C!} \Big[Ge^{-(E_{ns}-E_u)} + Ne^{-E_{sol}}\Big]^{C}. \tag{1.6}$$

Now, using Eq. 1.2, we have, $\mu = \ln C - \ln\Big[Ge^{-(E_{ns}-E_u)} + Ne^{-E_{sol}}\Big]$. Further, if the rest of the genome and free solution are macroscopic subsystems of the reservoir, their local chemical potentials are equilibrated. From this, one can show that the chemical potential μ obtained above can also be obtained by just considering the copy number of TFs that is bound to the DNA C_b as

$$\mu = \mu_b = \ln C_b - \ln\Big[Ge^{-(E_{ns}-E_u)}\Big], \qquad \text{where} \tag{1.7}$$

$$C_b = \frac{Ge^{-(E_{ns}-E_u)}}{Ge^{-(E_{ns}-E_u)} + Ne^{-E_{sol}}} C. \tag{1.8}$$

Also, we have, with C_f as the number of TF molecules free in the solution,

$$\mu = \mu_f = \ln C_f - \ln\Big[Ne^{-E_{sol}}\Big], \qquad \text{where}$$

$$C_f = \frac{Ne^{-E_{sol}}}{Ge^{-(E_{ns}-E_u)} + Ne^{-E_{sol}}} C.$$

Some of the random sequences on the rest of the DNA might be very similar to s^* by chance, and hence a TF molecule could bind to it in a specific configuration that depends on the specific random sequence. Hence, there are two alternative TF-DNA binding configurations - specific binding, depending on the sequence, and non-specific binding which is sequence-independent. Here, I will illustrate how to include specific binding to rest of the DNA as well into the computation of the chemical potential. We have,

$$Z(C) = \sum_{C_1=0}^{C} \binom{G}{C_1} \Big[\sum_{C_2=0}^{C_1} \binom{C_1}{C_2} e^{-(C_1-C_2)E_{ns}} \prod_{i=1}^{C_2} e^{-E(s_i)} \Big] e^{-(G-C_1)E_u} \binom{N}{C-C_1} e^{-(C-C_1)E_{sol}}, \tag{1.9}$$

where C_2 TF molecules (out of C_1 that are bound to the DNA) bind random sequences, $\{s_i\}$, on the DNA in a specific configuration, with energies $\{E(s_i)\}$. Assuming that random DNA sequences come from some underlying distribution P_{rand}, we can write the

following via a mean-field assumption:

$$Z(C) = \sum_{C_1=0}^{C} \binom{G}{C_1} \Big[\sum_{C_2=0}^{C_1} \binom{C_1}{C_2} e^{-(C_1-C_2)E_{ns}} \big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}}^{C_2} \Big] \\ e^{-(G-C_1)E_u} \binom{N}{C-C_1} e^{-(C-C_1)E_{sol}} \tag{1.10}$$

$$= \sum_{C_1=0}^{C} \binom{G}{C_1} \Big[e^{-E_{ns}} + \big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}} \Big]^{C_1} e^{-(G-C_1)E_u} \binom{N}{C-C_1} e^{-(C-C_1)E_{sol}} \tag{1.11}$$

Making the same approximations as before, one obtains

$$\mu = \ln C - \ln \Big[G e^{E_u} \Big(e^{-E_{ns}} + \big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}} \Big) + N e^{-E_{sol}} \Big] \tag{1.12}$$

$$= \ln C - \mu_0, \tag{1.13}$$

where we define $\mu_0 = \ln \Big[G e^{E_u} \Big(e^{-E_{ns}} + \big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}} \Big) + N e^{-E_{sol}} \Big]$. Assuming that the underlying distribution of DNA sequences is uniform over the nucleotide alphabet $\{A, C, G, T\}$, and making the constant mismatch penalty assumption, we have

$$\big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}} = \Big(\frac{1 + 3e^{-\epsilon}}{4} \Big)^L. \tag{1.14}$$

Saturating energy landscape

We have assumed that binding energy between a TF and a binding site increases linearly with the number of mismatches between the TF consensus sequence and the binding site sequence. It is often seen in experimental measurements that the binding energy increases till a certain threshold mismatch, after which it saturates to a constant value. Hence, an alternative model of TF-DNA binding is that of a saturating energy landscape in which binding energy increases linearly with mismatch k, as ϵk, till a threshold mismatch k_{ns}, after which it is constant at E_{ns}. A biological picture corresponding to this model is one in which the TF-DNA complex takes up two alternative binding configurations - specific or nonspecific - depending on which is the more favourable one energetically. In this case, we have,

$$Z(C) = \sum_{C_1=0}^{C} \binom{G}{C_1} \Big[P_{NS} e^{-E_{ns}} + (1 - P_{NS}) \big\langle e^{-E(s_i)} \big\rangle_{P_{\text{rand}}|P_{NS}} \Big]^{C_1} \\ e^{-(G-C_1)E_u} \binom{N}{C-C_1} e^{-(C-C_1)E_{sol}}, \tag{1.15}$$

where $P_{NS} = P_{\text{rand}}(E(s_i) \geqslant E_{ns})$ is the probability that a random DNA sequence of length L has specific binding energy greater than or equal to the nonspecific binding energy. In the constant mismatch energy model and for a uniformly distributed DNA

sequence, the following expressions follow from above. First, let's define $k_{ns} := \lceil E_{ns}/\epsilon \rceil$. We have,

$$P_{NS} = \sum_{i=k_{ns}}^{L} \binom{L}{i} \frac{3^i}{4^L}, \tag{1.16}$$

$$\left\langle e^{-E(s_i)} \right\rangle_{P_{\text{rand}}|P_{NS}} = \frac{1}{P_{NS}} \sum_{i=0}^{k_{ns}-1} \binom{L}{i} \frac{3^i}{4^L} e^{-\epsilon i}. \tag{1.17}$$

Probability of bound BS

Now, given the chemical potential of the TF species is μ, we can write, for the equilibrium probability that the BS is bound,

$$p_b = \frac{1}{Z} e^{-(E-\mu)} = \frac{1}{1+\exp(E-\mu)} = \frac{C}{C+\exp(E-\mu_0)}, \tag{1.18}$$

where $Z = 1 + e^{-(E-\mu)}$ is the partition function of the BS (+any bound TF) as the system. If the BS sequence s has k mismatches with the TF consensus sequence s^*, in the constant mismatch penalty model, we have,

$$p_b = \frac{1}{1+\exp(\epsilon(k-k^*)}, \tag{1.19}$$

where we define $k^* := \mu/\epsilon$, as a critical mismatch threshold. If the BS sequence has more than k^* mismatches with the TF consensus sequence, then its probability of binding is less than half: $p_b < 1/2$.

If there are multiple TF species (Q in total) present in the nucleus, with copy numbers $\{C_i\}$, chemical potentials $\{\mu^{(i)}\}$, binding energies $\{E_i\}$ and mismatches with BS sequence $\{k_i\}$ respectively, we have,

$$Z = 1 + \sum_{i=1}^{Q} e^{-(E_i-\mu_i)} = 1 + \sum_{i=1}^{Q} C_i e^{-(E_i-\mu_0^{(i)})} = 1 + \sum_{i=1}^{Q} e^{-\epsilon(k_i-k_i^*)}, \tag{1.20}$$

where k_i^* is the threshold mismatch of the i^{th} TF. Now the probability that the BS is bound by a molecule of the x^{th} TF is,

$$p_b^{(i)} = \frac{C_x e^{-(E_x-\mu_0^{(x)})}}{1 + \sum_{i=1}^{Q} C_i e^{-(E_i-\mu_0^{(i)})}} = \frac{e^{-\epsilon(k_x-k_x^*)}}{1 + \sum_{i=1}^{Q} e^{-\epsilon(k_i-k_i^*)}}. \tag{1.21}$$

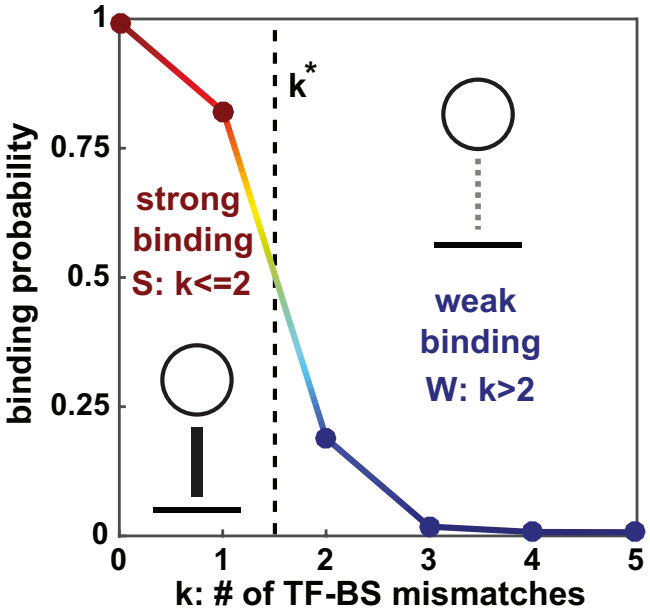


Figure 1.1: **Binding probability depends on the mismatches between TF consensus sequence and BS sequence.** Binding probability, the probability that a TF binds a DNA sequence, under the equilibrium assumption and the constant mismatch penalty assumption, depends on the number of mismatches, k, between the TF consensus sequence and the DNA sequence. If the mismatches k are smaller than a threshold, k^*, then the binding is strong and the DNA sequence is a BS, else it is weak and the sequence is not a BS.

1.4 Regulatory constraints on non-coding sequence

As seen in Fig. 1.1, the probability that a particular BS (with a specified sequence) is bound by a particular TF depends on whether the TF-BS mismatch is greater than or lesser than the critical mismatch threshold. In real organisms, there exist long stretches of regulatory elements like promoters and enhancers with very specific sequence features that shape their function. On the other hand, other stretches of DNA are required to not have any binding site sequences for TFs that are expressed in particular cells. In this chapter, I will explore the question of how do such regulatory constraints on the sequence space shape its structure? I will describe a first approach to this problem using a motif-based analysis.

The following framework is used. Given a length G of genomic sequences, σ, under the absence of any constraints there are 4^G sequences possible, with each sequence equally probable: $P_0(\sigma) = 1/4^G$, is a uniform distribution over the sequence space. When we introduce equality constraints via $f(\sigma) = a$, or inequality constraints via $g(\sigma) \leqslant b$) on the sequences, their distribution, $P(\sigma)$, changes to satisfy these constraints. In our analysis, we consider regulatory constraints as the existence or non-existence of motifs, meaning the similarity of binding site sequences to the consensus sequences of a few specified TFs.

Hence, the functions f and g, given a critical mismatch threshold, count the number of BS sequences (subsequences of σ) that are similar or dissimilar to a set of TF consensus sequences.

We are effectively asking what sequences are possible when certain motifs (sequences) have to be present (within the critical mismatch threshold) and certain other motifs have to be absent. First, let us consider the problem of understanding the space of non-regulatory sequences (no motifs should be present) considering only 1 motif of length L: no BSs for a specific TF.

1.5 One motif: non-regulatory sequence

Given a motif s^* of length L, we'll add in constraints on σ via the non-existence of motifs similar to s^* in any of σ's subsequences.

> For a sequence s^* (motif) of length L and an integer $k^* \geqslant 0$, what is the number of sequences σ of length G, such that for all subsequences s of σ, we have $d(s, s^*) > k^*$?

Here, $d(s, s^*)$ is the mismatch or the Hamming distance between s and s^*. To answer this question, we treat σ as a list of overlapping subsequences, $(s_1, s_2, s_3, \dots)$, of length L, each of which is a potential binding site. There are $X = G - L + 1$ such subsequences. Let us assume that the mismatch $d(s_i, s^*)$ between s_i and s^* is k_i. We want to calculate $P(k_1 > k^*, k_2 > k^*, \dots)$, the probability that every subsequence has a mismatch greater than k^*. We write this as $\gamma_{all} = P(\{k_i > k^*\})$ for ease. First, let us consider the computation of $P(\{k_i\})$, the overall joint probability of all the mismatches together, which, using the chain rule, can be decomposed as

$$P(\{k_i\}) = P(k_1) \prod_{j=2}^{X} P(k_j | k_{j-1}, k_{j-2}, \dots). \tag{1.22}$$

In the terms in the above equation, the distribution of k_j depends only on the mismatches of sites $i < j$. However, as the binding sites are of length L, the distribution of k_j depends only on the previous $L - 1$ mismatches: $k_{j-1}, k_{j-2}, \dots, k_{j-L+1}$. The other subsequences $s_i, i \leqslant j - L$ of length L do not overlap with s_j. Also, note that in general, k_j depends decreasingly lesser on k_{j-a} with increasing a as the two subsequences share a smaller overlap. Hence, we can write

$$P(\{k_i\}) = P(k_1) \prod_{j=2}^{X} P(k_j | \{k_{j-a}\}_{a<j, a<L}), \tag{1.23}$$

where $\{k_{j-a}\}_{a<j,a<L}$ represents a maximum of previous $L-1$ mismatches from site j. The central term in this calculation is $P(k_j | k_{j-1}, k_{j-2}, \dots, k_{j-L+1})$. Now, we can calculate $\gamma(Q = 1, L, G, k^*)$, the probability that a DNA sequence of length G does not contain

subsequences that are similar (by a maximum of k^* mismatches) to one given motif of length L as

$$\gamma_{all} = P(\{k_i > k^*\}) \tag{1.24}$$

$$= P(k_1 > k^*)\prod_{j=2}^{X} P(k_j > k^* | k_{j-1} > k^*, k_{j-2} > k^*, \ldots, k_{j-L+1} > k^*). \tag{1.25}$$

This dependence on the previous $L-1$ mismatches is hard and cumbersome to compute, and so we will assume that adjacent mismatch correlations can capture this mismatch dependency sufficiently and verify its validity in Fig. 1.2. Specifically, we will assume that $P(k_j > k^* | k_{j-1} > k^*, k_{j-2} > k^*, \ldots, k_{j-L+1} > k^*) \approx P(k_j > k^* | k_{j-1} > k^*)$. This is the first major approximation we undertake.

$$\gamma_{adj} = P(\{k_i > k^*\}) \quad = P(k_1 > k^*) \textstyle\prod_{j=2}^{X} P(k_j > k^* | k_{j-1} > k^*) \tag{1.26}$$

$$= P(k_1 > k^*) P(k_2 > k^* | k_1 > k^*)^{X-1}. \tag{1.27}$$

The second set of terms capture the mismatch correlations between adjacent sides. As there are $X-1$ adjacency mismatch terms from site 2 to site X, the exponent $X-1$ comes about. Thus, neglecting adjacent mismatch correlations, we have,

$$\gamma_{none} = \quad P(k_1 > k^*)^X. \tag{1.28}$$

We use γ_{adj} in further computations to capture adjacent mismatch dependency. The validity of this approximation depends crucially on the properties of the motif, and would not work well for motifs that are highly repetitive – say AAAAA, as seen in Fig. 1.2. However, most motifs, when randomly picked from the sequence space of some length L would not contain repetitive substrings, and hence these series of approximations hold well.

1.6 Many motifs: non-regulatory sequence

Next, we consider the case of the non-existence of matches to Q different motifs.

> For Q sequence $\{s_m^*\}$, $m = 1, 2, \ldots, Q$ (motifs) of length L and an integer $k^* \geqslant 0$, what is the number of sequences σ of length G, such that for all subsequences s of σ, we have $d(s, s_m^*) > k^* \ \forall \ m = 1, 2, \ldots, Q$?

For this, we want to calculate $P(\{k_{im} > k^*\}_{i \leqslant G-L+1, m \leqslant Q})$, where i indexes over binding sites on the sequence and m indexes over motifs. For simplicity, we write this as $P(\{k_{im} > k^*\})$. While in the previous scenario of 1 motif, we had to capture the mismatch correlations between adjacent sites, in the case of Q motifs, we have to capture also the mismatch correlations of a site with different motifs. Hence, the arrangement of the motifs in the sequence space needs to be accounted for. For instance, if two motifs are exactly similar, then their mismatches are completely correlated, and if the motifs are

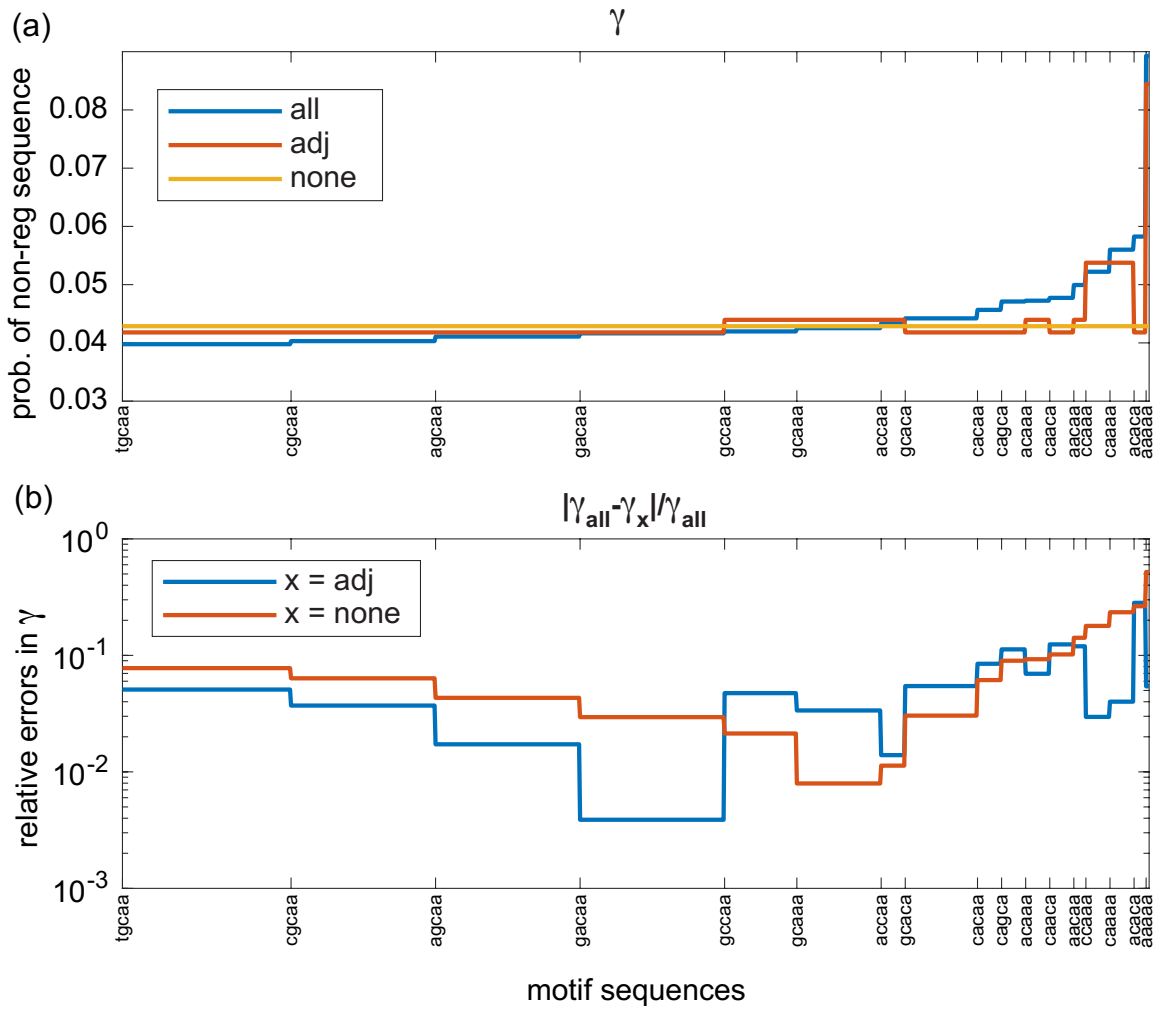


Figure 1.2: **Adjacent mismatch correlations capture the mismatch correlation structure sufficiently well.** **(a)** We show the probability that a sequence of length $G = 105$ is nonregulatory, with $L = 5$ and $k^* = 1$, when it is constrained to not contain sequences similar to one particular motif. We change that motif on the x-axis and order the motifs by this probability. We show the actual probability (Eq. 1.25), and also the probability under the assumption of only adjacent mismatch correlation (Eq. 1.26) and neglecting mismatch correlations (Eq. 1.28). **(b)** We show the relative errors in γ due to the approximations in the mismatch correlation structure. Highly repetitive motifs result in larger errors, as neighbouring mismatches are correlated for such motifs. A separate analysis (not shown) reveals that most motifs found empirically are not repetitive and hence this approximation holds.

quite different from each other, then their mismatches are uncorrelated. Decomposing like before, we have,

$$P(\{k_{im}\}) \;=\; P(\{k_{1m}\})\prod_{j=2}^{X} P(\{k_{jm}\}|\{k_{j-1,m}, k_{j-2,m}, \ldots, k_{j-L,m}\}). \tag{1.29}$$

Again, assuming that adjacent mismatch correlations can capture these mismatch dependencies sufficiently, thereby resulting in $P(\{k_{jm}\}|\{k_{j-1,m}, k_{j-2,m}, \ldots, k_{j-L,m}\}) \approx P(\{k_{j,m}\}|\{k_{j-1,m}\})$. Hence, we have,

$$P(\{k_{im}\}) \;=\; P(\{k_{1m}\})\prod_{j=2}^{X} P(\{k_{jm}\}|\{k_{j-1,m}\}). \tag{1.30}$$

We can calculate $\gamma(Q, L, G, k^*)$, the probability that a DNA sequence of length G does not contain subsequences that are similar to Q given well-separated motifs of length L, by a maximum of k^* mismatches as

$$\begin{aligned}
\gamma &= P(\{k_{im} > k^*\}) && (1.31)\\
&= P(\{k_{1m} > k^*\})\prod_{j=2}^{X} P(\{k_{jm} > k^*\}|\{k_{j-1,m} > k^*\}) && (1.32)\\
&= P(\{k_{1m} > k^*\})\prod_{j=2}^{X} \frac{P(\{k_{jm} > k^*\}, \{k_{j-1,m} > k^*\})}{P(\{k_{j-1,m} > k^*\})} && (1.33)\\
&= \frac{P(\{k_{2m} > k^*, k_{1m} > k^*\})^{X-1}}{P(\{k_{1m} > k^*\})^{X-2}}. && (1.34)
\end{aligned}$$

We need to decompose $P(\{k_{1m} > k^*, k_{1m} > k^*\})$ and $P(\{k_{1m} > k^*\})$ into terms corresponding to different motifs. To achieve this, we need to specify the arrangement of these motifs in sequence space. Our basic strategy is a mean-field-like formulation of the arrangement of "motif balls" in the sequence space, and consider their individual positioning and their pairwise overlaps. These balls have a "radius" k^* around each motif, and contain DNA sequences that are the BSs for the specified TF, and are hence forbidden to be present in a sequence that is non-regulatory. For intuition, I use the term "radius" but it is important to remember that we are working in a discrete space of sequences.

As an instructive guide, let us consider the decomposition of $P(\{k_{1m} > k^*\})$. We define $B_{1m} : k_{1m} \leqslant k^*$ as an event that corresponds to such "motif balls" described above. $P(\{k_{1m} > k^*\})$ now corresponds to exactly the (fractional) volume of the sequence space (of size 4^L) outside the set of all motif balls. If the balls were not at all overlapping (if TF consensus sequences are sufficiently distinct from each other), it is pretty straightforward to compute this as we just have to sum up the volumes of these balls. In the case that the balls overlap, which can happen if for a pair of motifs s^*_m, s^*_n, $d(s^*_m, s^*_n) \leqslant 2k^*$, we have to correct for this overlap when computing the volume outside the balls. We formalize this approach in the following manner, and illustrate the approach in Fig. 1.3.

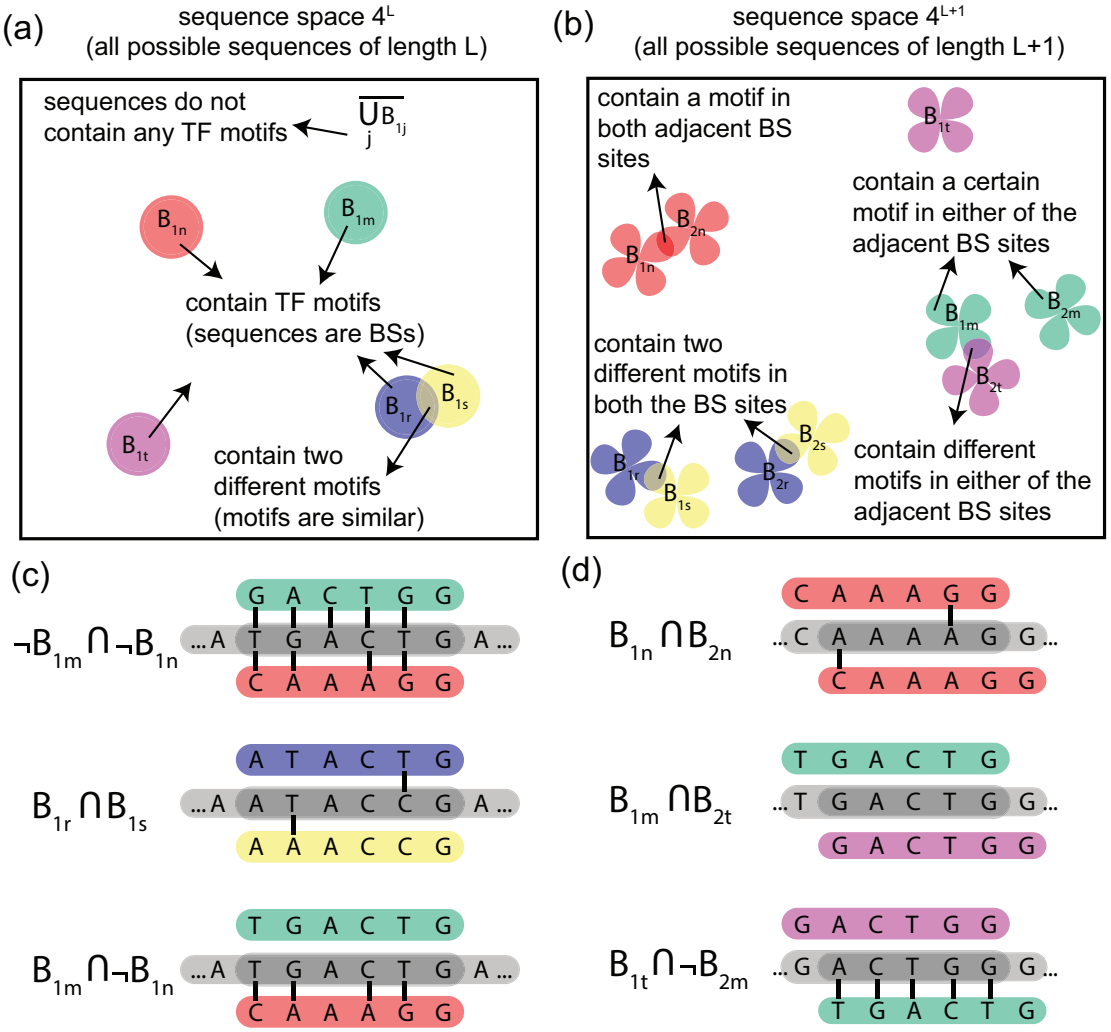


Figure 1.3: **Arrangement of motifs in sequence space.** The probability of obtaining a non-regulatory sequence depends on the total number of motifs to be avoided via the arrangement of these motifs in sequence space, bringing into picture how similar or dissimilar different pairs of motifs are. In **a** and **b**, we illustrate how various "motif balls" or "4−finned fans", which are DNA sequences that are similar to TF motif sequences (Hamming distance between DNA sequence and TF consensus sequence not greater than k^*), are arranged in sequence spaces of size L and $L+1$ respectively. **(a)** This picture corresponds to the computation of $P(\{k_{1m} > k^*\}$. For a sequence of length L, the size of sequence space is 4^L, and corresponding to each motif x is a ball of radius k^*, B_{1x}. Sometimes these balls can overlap (blue and yellow) if two motifs, here r and s, have similar sequences. We consider this overlap only in the second-order approximation. (Caption continued in the next page.)

Figure 1.3: (Continued from previous page.) **(b)** This picture is used in the computation of $P(\{k_{2m} > k^*, k_{1m} > k^*\})$. For a sequence of size $L+1$, the sequence space is of size 4^{L+1} and corresponding to each motif, there are two 4−finned fans, each of which corresponds to either of the two adjacent length L sites numbered 1 and 2. For each fan, there are 4 fins because the nucleotide in the other $(L+1)^{th}$ position (to the extreme left or extreme right) can contain any of the four possible nucleotides. Various 4−finned fans can overlap as illustrated. The fans B_{1n} and B_{2n} corresponding to adjacent sites for the same motif n can overlap sometimes (red), for instance, if the motif has continuous repeats of the same nucleotide. This is accounted for by $P(B_1, B_2)$ in Eq. 1.41 in both the first-order and second-order approximations. Similar to **s**, the fans corresponding to two motifs r and s can overlap (blue and yellow) in the same site if the motifs are similar. The fans B_{1m} and B_{2t}, corresponding to one motif with the first site and another motif with the second site, can overlap if motif sequences are similar after shifting one of them by 1bp. We consider these only in the second-order approximation. In both **a** and **b**, we do not show overlap of 3 or more balls or 4−finned fans, which we also neglect in our computations. In **c** and **d**, we show specific examples of sequences corresponding to various scenarios of **a** and **b** respectively. **(c)** *Top:* $\neg B_{1m} \cap \neg B_{1n}$ – length L sequence that is not similar to motifs m (green) and n (red) both, *Middle:* $B_{1r} \cap B_{1s}$ – length L sequence that is similar to motifs r (blue) and s (yellow) both, *Bottom:* $B_{1m} \cap \neg B_{1n}$ – length L sequence that is similar to motif m (green) but not similar to motif n (red). **(d)** *Top:* $B_{1n} \cap \neg B_{2n}$ – length $L+1$ sequence that is similar to a motif n (red) in both the sites 1 and 2, *Middle:* $B_{1m} \cap B_{2t}$ – length $L+1$ sequence that is similar to motif m (green) in site 1 and similar to motif t (purple) in site 2, *Bottom:* $B_{1t} \cap \neg B_{1m}$ – length $L+1$ sequence that is similar to motif t (purple) in site 1 but not similar to motif m (green) in site 2 (compare with *Middle* of **d**).

1.7 Dependence on motif arrangement

As described above, when we have multiple motifs, the arrangement of these motifs in the sequence space becomes important, and we have to account for motif similarity which can result in mismatch correlations. A generic term that comes about in the expressions of γ is $P(\{\neg E_x\})$ for a set of x that corresponds to different motifs and/or different sites, with each E_x being events pertaining to whether various mismatches exceed the mismatch threshold k^* or not. Note that "¬" denotes NOT, "," and "∩" denote intersection. One can expand $P(\{\neg E_x\})$ as,

$$P(\{\neg E_x\}) = \underbrace{\underbrace{1 - \sum_i P(E_x)}_{\text{first-order}} + \sum_{x \neq y} P(E_x, E_y)}_{\text{second-order}} - \underbrace{\sum_{x \neq y \neq z} P(E_x, E_y, E_z) + \ldots}_{\text{higher-order terms}} \tag{1.35}$$

where each successive correction term corresponds to considering successively higher order overlap among the events $\{E_x\}$.

In the computation of $P(\{k_{1m} > k^*\})$, we have $\neg E_x := k_{1m} > k^*$, letting us identify $E_x = B_x$, with $x = 1m$, corresponding to various motifs. In the computation of $P(\{k_{1m} > k^*, k_{1m} > k^*\})$, we have $\neg E_x := k_{1m} > k^*, k_{1m} > k^*$, letting us identify $E_x = \neg(\neg B_{1m} \cap \neg B_{2m})$, with $x = m$ going over various motifs.

1.7.1 First order approximation

In the first-order approximation, we assume that the motifs, $\{s^*_m\}$, are fairly independent of each other: $d(s^*_m, s^*_n) > 2k^* \ \forall \ m, n$. In this case, the events $B_{jm} : k_{jm} \leqslant k^*$ and $B_{jn} : k_{jn} \leqslant k^*$ can be assumed to be independent, and terms of the form $P(B_{jm}, B_{jn})$ and higher can be neglected. Hence, we can decompose the joint distribution of all mismatches into distributions containing mismatches for each motif separately, resulting in,

$$\begin{aligned} P(\{k_{1m} > k^*\}) &= 1 - \sum_{m \leqslant Q} P(k_{1m} \leqslant k^*) && (1.36)\\ &= 1 - QP(k_1 \leqslant k^*) && (1.37)\\ &= 1 - QP(B_1), && (1.38) \end{aligned}$$

where we have abused notation to define $B_1 : k_1 \leqslant k^*$ as the motif ball around any particular motif. Similarly, we have, for adjacent sites,

$$\begin{aligned} P(\{k_{2m} > k^*, k_{1m} > k^*\}) = & \quad 1 - Q\Big[1 - P(k_1 > k^*, k_2 > k^*)\Big] && (1.39)\\ = & \quad 1 - Q + QP(\neg B_1, \neg B_2) && (1.40)\\ = & \quad 1 - Q\Big[P(B_1) + P(B_2) - P(B_1, B_2)\Big], && (1.41) \end{aligned}$$

where k_1 and k_2 are mismatches of any motif to adjacent binding sites, and $B_2 : k_2 \leqslant k^*$ is defined for any particular motif like B_1 by abuse of notation. Combining these, we have,

$$\begin{aligned} \gamma^I_{adj} &= \frac{\Big(1 - Q + QP(k_2 > k^*, k_1 > k^*)\Big)^{X-1}}{\Big(1 - QP(k_1 \leqslant k^*)\Big)^{X-2}} && (1.42)\\ &= \frac{\Big(1 - Q + QP(\neg B_1, \neg B_2)\Big)^{X-1}}{\Big(1 - QP(B_1)\Big)^{X-2}}. && (1.43) \end{aligned}$$

The condition on motif arrangement for the first-order approximation to strictly hold for adjacent mismatches is not as straightforward. Even if motif similarity is avoided (B_{1m} and B_{1n} overlap avoided) by making sure they are sufficiently dissimilar, B_{1m} and B_{2n} can overlap for two different motifs m and n that are similar when one motif is shifted by 1bp. Also, in this case, the relevant sequence is of size $L + 1$ with mismatches belonging to the two adjacent sites of length L. For each motif m, there are 8 balls now, 4 corresponding to $B_{2m} : k_{2m} \leqslant k^*$ and 4 corresponding to $B_{1m} : k_{1m} \leqslant k^*$. There are 8 balls because

the $L+1$ sequence is either $*s_m$ or s_m* where $*$ is any of $\{A, C, G, T\}$ in the first or last position. Hence, the "shape" of the event for each motif is now two 4-finned fans (instead of balls), with the two fans corresponding to the $L+1$ sequence with a motif in one of the adjacent sites, and the four fins of each fan corresponding to the 4 possible nucleotides in the other single position.

1.7.2 Second order approximation

Next, we consider the overlap between balls (and 4-finned fans) and use a mean-field approach to correct the first order approximation. We assume that all higher-order overlap is negligible: the probability of three or more balls overlapping can be neglected. With the overlap taken into account, we have,

$$P(\{k_{1m} > k^*\}) \quad = 1 - QP(B_1) + \frac{Q(Q-1)}{2}\langle P(B_{1m}, B_{1n})\rangle, \tag{1.44}$$

where the second-order correction term accounts for overlap between balls of different motifs. We have,

$$\langle P(B_{1m}, B_{1n})\rangle = \sum_{z=0}^{L} \alpha_z(L)\Gamma(k^*, k^*, z, L), \tag{1.45}$$

where $\alpha_z(L)$ is the probability that a pair of motifs (of length L) from our motif ensemble are separated by Hamming distance z, and $\Gamma(i, j, z, L) = P(k_{1m} \leqslant i, k_{1n} \leqslant j | d_{mn} = z)$ is the probability that mismatches k_{1m} and k_{1n} for a sequence of length L with two motifs m and n, are less than i and j respectively, when the consensus sequences of the motifs are separated by z. This can be computed as in Eq 4.18.

The computation of $P(\{k_{2m} > k^*, k_{1m} > k^*\})$ is a bit more involved as this involves various kinds of overlaps between the 4-finned fans corresponding to the two adjacent sites of the various motifs. We have,

$$P(\{k_{2m} > k^*, k_{1m} > k^*\}) = 1 - Q + QP(\neg B_1, \neg B_2) + \frac{Q(Q-1)}{2}\langle P_{\text{II}}\rangle, \tag{1.46}$$

where $\langle P_{\text{II}}\rangle = \langle P(\neg(\neg B_{1m} \cap \neg B_{2m}) \cap \neg(\neg B_{1n} \cap \neg B_{2n}))\rangle$ is the mean second-order correction term that considers the joint overlap between motifs and adjacent sites. We can expand this as,

$$\begin{aligned}\langle P_{\text{II}}\rangle = \langle & P(B_{1m}, \neg B_{2m}, B_{1n}, \neg B_{2n}) + P(\neg B_{1m}, B_{2m}, \neg B_{1n}, B_{2n}) \\ & + P(B_{1m}, \neg B_{2m}, \neg B_{1n}, B_{2n}) + P(\neg B_{1m}, B_{2m}, B_{1n}, \neg B_{2n}) \\ & + P(B_{1m}, B_{2m}, B_{1n}, \neg B_{2n}) + P(B_{1m}, B_{2m}, \neg B_{1n}, B_{2n}) \\ & + P(B_{1m}, \neg B_{2m}, B_{1n}, B_{2n}) + P(\neg B_{1m}, B_{2m}, B_{1n}, B_{2n}) \\ & + P(B_{1m}, B_{2m}, B_{1n}, B_{2n})\rangle.\end{aligned} \tag{1.47}$$

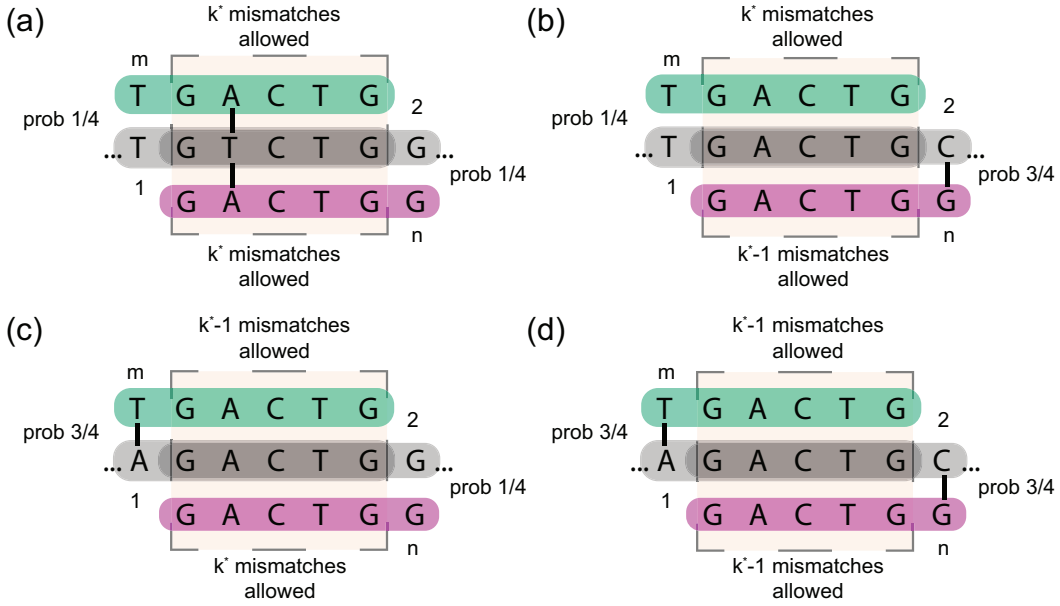


Figure 1.4: **Terms in the calculation of Eq. 1.50.** The calculation of $\langle P(B_{1m}, B_{2n})\rangle$ in Eq. 1.50, involves calculating the probability that there is a sequence of length $L+1$ such that for two motifs m (green) and n (purple), the first site is similar to motif m while the second site is similar to motif n. This occurs if the motifs m and n are similar when one of them (here, n) is shifted (here, to the right) by 1bp. The calculation involves 4 different terms corresponding to the whether motif m and motif n have a mismatch with the first and the last position of the $L+1$ length DNA sequence respectively. **(a)** Motif m does not have a mismatch with the first nucleotide (probability 1/4) and motif n does not have a mismatch with the last nucleotide (probability 1/4), resulting in a join probability of 1/16. The remaining $L-1$ positions can contain up to k^* mismatches with both motifs. **(b)** Motif m does not have a mismatch with the first nucleotide (probability 1/4) and motif n has a mismatch with the last nucleotide (probability 3/4), resulting in a join probability of 3/16. The remaining $L-1$ positions can contain up to k^* mismatches with motif m and up to k^*-1 mismatches with motif n. **(c)** Motif m has a mismatch with the first nucleotide (probability 3/4) and motif n does not have a mismatch with the last nucleotide (probability 1/4), resulting in a join probability of 3/16. The remaining $L-1$ positions can contain up to k^*-1 mismatches with motif m and up to k^* mismatches with motif n. **(d)** Motif m has a mismatch with the first nucleotide (probability 3/4) and motif n has a mismatch with the last nucleotide (probability 3/4), resulting in a join probability of 9/16. The remaining $L-1$ positions can contain up to k^*-1 mismatches with both motifs m and n.

We assume that k^* is small enough and that the motif ensemble is sufficiently random that intersections of more than two 4-finned fans simultaneously is negligible. This means that the terms in lines 3, 4, 5 of Eq. 1.47 can be neglected, giving us,

$$\langle P_{\text{II}}\rangle = \langle P(B_{1m}, B_{1n})\rangle + \langle P(B_{2m}, B_{2n})\rangle + \langle P(B_{1m}, B_{2n})\rangle + \langle P(B_{2m}, B_{1n})\rangle. \tag{1.48}$$

The first two terms contain the overlap of the 4-finned fans of the two motifs to the same binding site, while the last two terms contain the overlap of the 4-finned fans of the two

motifs to adjacent binding sites. As there is symmetry between motifs m and n, without loss of generality, the second order term can be written as,

$$\begin{aligned}\langle P_{\mathrm{II}}\rangle &= 2\langle P(B_{1m}, B_{1n})\rangle + 2\langle P(B_{1m}, B_{2n})\rangle \\ &= 2\langle P(B_{1m}, B_{2n})\rangle + 2\sum_{z=0}^{L}\alpha_z(L)\Gamma(k^*, k^*, z, L),\end{aligned} \tag{1.49}$$

where $\alpha_z(L)$ and $\Gamma(i, j, z, L) = P(k_{1m} \leqslant i, k_{1n} \leqslant j | d_{mn} = z)$ are as defined before. The computation of $\langle P(B_{1m}, B_{2n})\rangle$ involves accounting for various scenarios as depicted in Fig. 1.4, resulting in,

$$\begin{aligned}\langle P(B_{1m}, B_{2n})\rangle = &\frac{1}{16}\sum_{z=0}^{L-1}\alpha_z(L-1)\Gamma(k^*, k^*, z, L-1) \\ &+ \frac{3}{16}\sum_{z=0}^{L-1}\alpha_z(L-1)\Gamma(k^*, k^*-1, z, L-1) \\ &+ \frac{3}{16}\sum_{z=0}^{L-1}\alpha_z(L-1)\Gamma(k^*-1, k^*, z, L-1) \\ &+ \frac{9}{16}\sum_{z=0}^{L-1}\alpha_z(L-1)\Gamma(k^*-1, k^*-1, z, L-1).\end{aligned} \tag{1.50}$$

1.8 Results

Equipped with expressions for γ^I_{adj} and $\gamma^{II}_{\mathrm{adj}}$, we now explore the fraction of non-regulatory sequences of different lengths under different constraints. First, in Fig. 1.5, we plot the ratio of the fraction of non-regulatory sequences according to the first-order and second-order approximations, $\gamma^I_{\mathrm{adj}}/\gamma^{II}_{\mathrm{adj}}$. The second-order approximation starts to differ from the first-order approximation when the number of motifs Q increases, as the overlap among balls and 4−finned fans also increases as a consequence.

In Fig. 1.6, we show how the fraction of non-regulatory sequences, γ_{adj}, depend on two important parameters in our model – the number of TF motifs to be avoided, Q, and the length of DNA sequence in question, G, for both first-order, γ^I_{adj}, and second-order, γ^{II}_{adj}, approximations. From these we can infer that the fraction of non-regulatory sequences decreases as the length of DNA sequence and the number of motifs to be avoided increase. These have been computed using $L = 6$ and $k^* = 1$, which are typical numbers for a eukaryote. At $G = 100$ and $Q = 10$, when 10 motifs of length 6 have to be avoided by at least two mismatches, in a sequence of length 100, we find that $\gamma^I_{\mathrm{adj}} \approx 0.0115$, and $\gamma^{II}_{\mathrm{adj}} \approx 0.0155$, indicating that only between 1% and 2% of sequences of length 100 are non-regulatory, the rest of the $\sim 98\%$ containing one BS somewhere and hence would have the potential to act as regulatory sequences. While the total number of such sequences is very large (2% of 4^{100}), assuming that these non-regulatory sequences

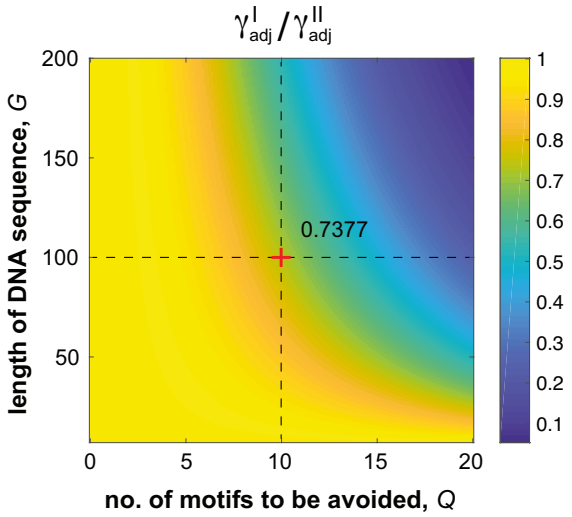


Figure 1.5: **First-order and second-order approximations differ for longer DNA sequences and more TF motifs.**. The ratio of the fraction of non-regulatory sequences from the first-order approximation to the second-order approximation is plotted against various total lengths, G, on the y-axis, and the number of TF motifs to be avoided, Q, on the x-axis. The approximations result in similar values for small Q and short sequences – low G. As the number of motifs increase, motif overlap becomes more prevalent and hence, the second-order approximation starts to deviate from the first-order approximation. For $G = 100$ and $Q = 10$, we find $\gamma^{I}_{\text{adj}}/\gamma^{II}_{\text{adj}} \approx 0.7377$, parameters used: $L = 6, k^* = 1$. We assume that motifs and DNA sequences are uniformly random sequences of their respective lengths.

are spread homogeneously across sequence space, it means that non-regulatory sequences are very hard to find. Increasing the length of the DNA sequence, G, or increasing the number of motifs, Q, that are to be avoided, only lowers the fraction of non-regulatory sequences exponentially. However, doubling the motif length, to $L = 12$, increases the fraction of non-regulatory sequences to $\gamma^{I}_{\text{adj}} \approx \gamma^{II}_{\text{adj}} \approx 0.9981$. Hence, avoiding larger motifs is easier. The fraction of $G = 1$kbp sequences that avoid $Q = 10$ motifs of length $L = 6$ is practically vanishing, while avoiding $Q = 10$ motifs of length $L = 12$ is achievable, with the fraction of non-regulatory sequences being 0.979. With $G = 10$kbp, fraction of non-regulatory sequences decreases to 0.807, and to 0.117 for $G = 100$kbp, and is vanishingly small for $G = 1$Mbp.

In summary, constraints requiring DNA sequences to be non-regulatory can significantly shape the sequence space, with the fraction of non-regulatory sequences becoming smaller as the length of the DNA sequence increases, the number of motifs to be avoided increases, and the length of the motifs to be avoided decreases. While this approach informs us of such broad dependancies of regulatory sequence space on important quantities like total DNA sequence length and number of motifs, it lacks a rigorous biophysical backbone of the kind described in the beginning of this chapter, that directly connects with the functional consequence of such constraints. In the next chapter, I will incorporate the biophysical model of TF-DNA binding described in this chapter into a broader biophysical

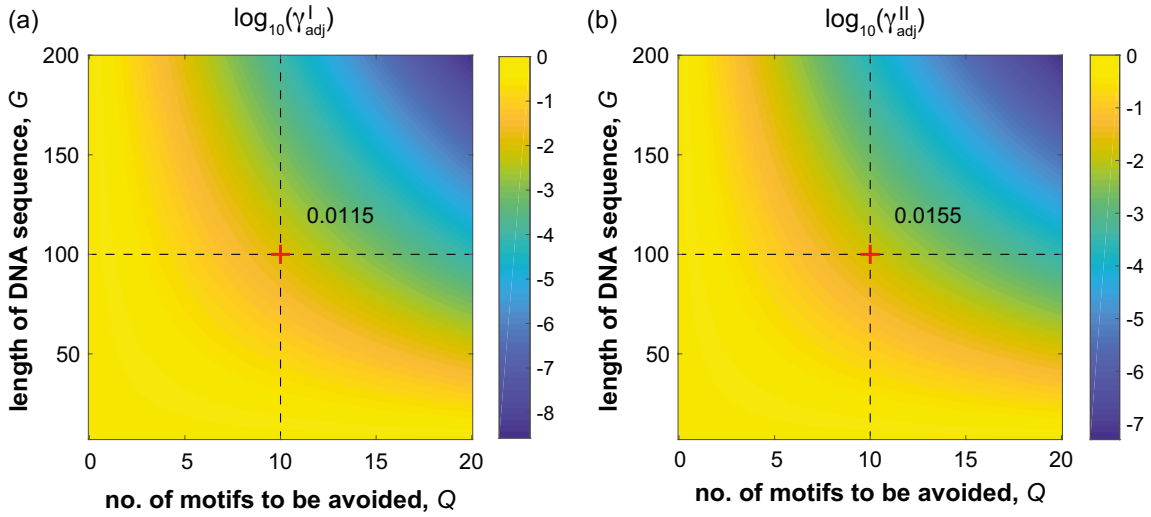


Figure 1.6: **Fraction of non-regulatory sequences of various lengths under varying number of TF motifs to be avoided.**. The fraction of non-regulatory sequences of various total lengths, G, on the y-axis, against the number of TF motifs to be avoided, Q, on the x-axis, plotted on a log-scale for both **(a)** first-order, γ^I_{adj}, and **(b)** second-order, γ^{II}_{adj}, approximations. As the number of motifs increases, the fraction of non-regulatory sequences decreases, with the largest change occurring for longer DNA sequences (larger G). Also, at a fixed Q, the fraction of longer non-regulatory sequences is smaller. For $G = 100$ and $Q = 10$, we find $\gamma^I_{\text{adj}} \approx 0.0115$, and $\gamma^{II}_{\text{adj}} \approx 0.0155$, parameters used: $L = 6, k^* = 1$.

model of global transcriptional regulation to quantify crosstalk – incorrect interactions between TFs and DNA that might lead to erroneous regulatory states – one of the direct consequences of the cell failing to shape its sequence space according to the type of non-regulatory constraints described here.

Chapter 2

Crosstalk in gene regulation

The work presented in this chapter was performed in collaboration with Tamar Friedlander and has been published in Nature Communications (see [Friedlander *et al.*, 2016]); parts of the publication that I worked on are explained in this chapter. Tamar Friedlander did some of the calculations in Section 2.2 and Section 2.3.

2.1 Introduction

As discussed in Chapter 0, the specificity of molecular recognition events is crucial to the functioning of a cell and ensuring that cellular processes run in their right spatiotemporal contexts. Cells are typically crowded with molecular components, leading to a large number of non-cognate partners. It is not trivial to see if the limited specificity of the underlying molecular forces behind molecular recognition leads to a large number of incorrect recognition events that might hamper the information processing efficiency of cellular processes. This is called crosstalk, encompassing all potentially disruptive processes due to reactions between non-cognate components. Further, it remains to be seen if such crosstalk is strong enough to exert selective pressure on the design of molecular recognition systems. Such an evolutionary pressure might lead biological systems to evolve specific molecular mechanisms and strategies that can overcome the deleterious effects of crosstalk. One paradigmatic example is the aminoacyl transfer RNA synthetase [Yamane and Hopfield, 1977], which uses kinetic proofreading [Hopfield, 1974] to load appropriate amino acids onto matching tRNAs. Other examples from ligand sensing [Mora, 2015], protein-protein interactions [Swain and Siggia, 2002; Skerker *et al.*, 2008; Johnson and Hummer, 2011; Zhang *et al.*, 2008; Ouldridge and ten Wolde, 2014; Rowland and Deeds, 2014], recognition events in the immune system [McKeithan, 1995;

Lalanne and François, 2013] and molecular self-assembly [Murugan *et al.*, 2015] indicate that biology places a large premium on the reduction of unintended crosstalk.

A key step in transcriptional regulation involves the sequence-specific binding of TFs to binding sites in regulatory elements near genes. This is another example of molecular recognition, the specificity of which arises from hydrogen bonds formed between amino acids on the DNA-binding domains of TFs and nucleotide bases on the DNA. Depending on the amino acid sequence of its DBD, each TF preferentially binds to a small number of DNA sequences. But a large body of evidence shows that this binding specificity is limited, and that TFs bind other non-cognate targets as well [Von Hippel *et al.*, 1974; Wunderlich and Mirny, 2009; Johnson *et al.*, 2005; Maerkl and Quake, 2007; Rockel *et al.*, 2012]. Such additional binding targets have been discussed as sequestering TFs at non-functional sites, and thereby reducing the free TF concentration [Burger *et al.*, 2010; Sheinman and Kafri, 2012]. But such off-targets can sometimes be embedded in the regulatory elements of other genes, leading to an interference with various gene regulatory programs. Given that multiple TFs (from different genes) are typically co-expressed in a spatiotemporal window, each molecule has a small probability of erroneously regulating some subset of all genes. Hence, crosstalk is a global systemic property that has to be understood by considering the whole ensemble of TFs and genes, that can form disruptive causal links between various gene regulatory programs.

The other feature of crosstalk is its combinatorial explosion as the regulatory system grows in complexity and the number of regulatory components increases. The number of potential non-cognate interactions grows much faster than the number of cognate interactions, making the problem biologically relevant and theoretically interesting. But studies so far have largely considered a simpler setting of a single TF, and computed its binding probabilities to cognate versus non-cognate sites [Gerland *et al.*, 2002; Sengupta *et al.*, 2002; Bintu *et al.*, 2005; Lynch and Hagner, 2015]. They have not included the crucial (mis)regulation effects of TFs on non-cognate regulatory targets and genes. They have largely focussed on the question of how reliable gene regulation is achieved by (cognate) TFs [Todeschini *et al.*, 2014], whereas the complementary question of how to prevent erroneous regulation by non-cognate TFs has remained largely unexplored (but see [Bird, 1995]).

In this chapter, I will describe a new quantitative framework for regulatory crosstalk that captures its global nature by simultaneously treating multiple TFs and multiple regulatory binding sites. We explicitly account for differential activation of genes depending on regulatory conditions, an aspect that has been missed in previous studies of molecular recognition [Hopfield, 1974]. In particular, the ability of the regulatory system to prevent spurious gene activation despite crosstalk interference will emerge as an important consideration. TF-DNA interactions are assumed to be in thermodynamic equilibrium [Bintu *et al.*, 2005; Phillips, 2015], an assumption that holds well for prokaryotic systems [Ackers *et al.*, 1982; Kinney *et al.*, 2010]. Such an assumption underlies the majority of modelling and bioinformatic applications, and puts strong constraints on models of crosstalk. In this work, we explore the consequences of this equilibrium assumption on regulatory crosstalk,

serving as an instructive platform for non-equilibrium studies [Cepeda-Humerez *et al.*, 2015].

We construct a biophysical model, based on an equilibrium assumption for TF-DNA interactions, for crosstalk in transcriptional regulation. By basing the model in sequence space (TF consensus sequences and BS sequences), we account globally for all cross-interactions between TFs and their binding sites. We construct the model using many parameters that have a direct biological meaning, and by tuning them, we identify how they influence crosstalk levels. Some of these parameters, like the TF concentrations, are empirically known to belong to a broad range but they change dynamically. To overcome this, we investigated the variation of crosstalk with respect to these parameters and show the existence of a "crosstalk floor" - a lower bound on crosstalk. Such a threshold cannot be bettered by the cell even if were to optimally adjust those parameters using specific molecular mechanisms. For instance, even if the cell adjusts TF concentrations by different feedback mechanisms and compensates for TF molecule sequestration, it cannot decrease crosstalk below a certain fixed threshold.

Using this model, we ask the following fundamental questions related to crosstalk and gene regulatory systems.

1. How does crosstalk depend on the number of (co-expressed) genes and the parameters underlying the biophysical model of TF-DNA interactions, such as binding site length and binding energy?

2. In the context of crosstalk, what are the similarities and differences between the regulatory strategies of prokaryotes and those of eukaryotes?

3. Are complex regulatory strategies, such as combinatorial regulation, or regulation by activators and repressors, capable of lowering crosstalk [Todeschini *et al.*, 2014]?

Studies have shown that many biophysical constraints, such as programmability [Gerland *et al.*, 2002], response speed [Mangan and Alon, 2003], noise in gene expression and dynamic range of regulation [Tkačik and Walczak, 2011; Dubuis *et al.*, 2013; Friedlander and Brenner, 2011; Friedlander and Brenner, 2008], robustness [von Dassow *et al.*, 2000] and evolvability of the regulatory sequences [Payne and Wagner, 2014; Stern and Orgogozo, 2009], shape the design of genetic regulatory networks. These constraints could be understood at the level of individual genetic regulatory elements and do not require a systemic formulation. But crosstalk is different. Although it arises from biophysical limits to molecular recognition locally at the level of single genetic regulatory element, it is only on a global scale that its cumulative effect emerges. At the local level, crosstalk can be reduced by increasing the concentration of cognate TFs or by introducing multiple binding sites in the promoter. It is only when we self-consistently consider that these same cognate TFs act as non-cognate TFs for other genes, or that new binding sites in the promoter drastically increase the number of non-cognate binding configurations, that crosstalk constraints become clear.

2.2 Basic model

In our basic biophysical model of gene regulation, we assume that each of the M genes in the genome of a cell is regulated by a dedicated activator TF type. For each gene, a molecule of its dedicated TF type binds to a single binding site of length L basepairs in the gene's regulatory region, and only upon this TF-BS binding, the expression of the gene is activated. Hence, in this basic model, there are M distinct TF types, resulting in one cognate TF for each gene. It is important to remember that this is not a realistic picture of a gene regulatory network – there are usually multiple TFs per gene, often acting combinatorially, and the question of who regulates the regulators. We relax some of the assumptions later in this chapter, and in the associated publication [Friedlander *et al.*, 2016], and consider combinatorial regulation, cooperative regulation etc. The question of who regulates the regulators is taken care of by assuming that upstream signals are perfectly sensed by the cell, both by transcriptional mechanisms and other mechanisms not relevant to the model, to "activate" the necessary regulators in any environmental context adopting a mean-field like approach over the various possible environmental states – thereby sampling the various "states" of the gene regulatory network inside the cell. We leave the formulation of a full model with upstream environmental signals and temporal evolution of GRNs, to future research. But see Chapter 3 and Chapter 4 for first steps in this direction.

As described in Chapter 1, the probability of a BS being bound by a TF molecule depends on the mismatch between the BS sequence and the TF consensus sequence [Ackers *et al.*, 1982; Von Hippel and Berg, 1986; Lynch and Hagner, 2015]. Each TF forms a perfect match with the BS sequence of its target gene and hence, the probability that a molecule of the cognate TF type binds is high for each BS as long as the TF is present in the cell. However, because non-cognate TFs also have some mismatch with BS sequences, and which can sometimes be very low due to the limitations of sequence space, each BS could also be occasionally bound by a non-cognate TF molecule. Hence, every TF can bind, apart from its cognate BS, other non-cognate BSs, but often only with a low probability.

To capture the ability of gene regulation to differentially activate subsets of genes in a pattern appropriate to the environmental conditions (signals, cell type or time), we assume that, at any point in time, that only a subset of these M TF types are present in the cell. We consider many such timepoints or environments, each with different subsets of size $Q \leqslant M$ activating TF present in the cell in nonzero copy numbers. The optimal gene regulatory state in each of these environments would be to activate only those corresponding Q target genes for which activating TFs are present, while keeping the remaining $M - Q$ genes inactive. Over different environmental conditions that the cell faces, different sets of Q out of M genes get activated by the corresponding Q cognate TFs.

But how does the cell know which correct set of TFs to express in each particular en-

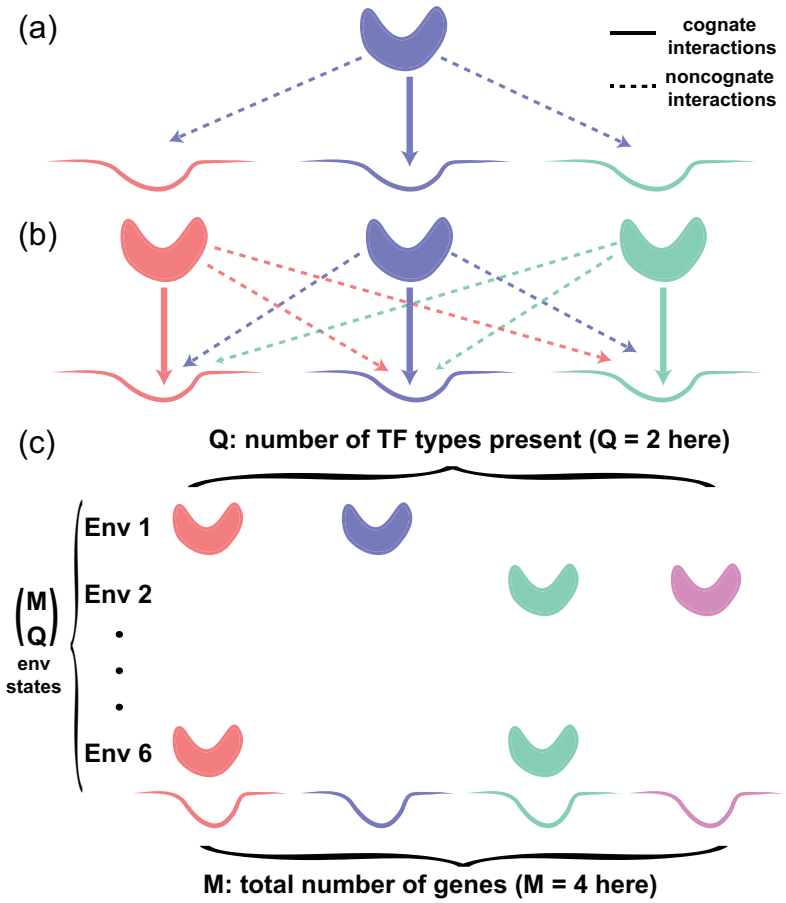


Figure 2.1: **Crosstalk in gene regulation.** **(a)** A TF preferentially binds to its cognate binding site, but can also bind non-cognate sites, potentially causing crosstalk - an erroneous activation or repression of a gene. **(b)** In a global setting where many TFs regulate many genes, the number of possible non-cognate interactions grows quickly with the number of TFs; in addition, it may become difficult to keep TF recognition sequences sufficiently distinct from each other. **(c)** Cells respond to changing environments by attempting to activate subsets of their genes. In this example, the total number of genes is $M = 4$ and different environments (here, 6 in total) call for activation of different subsets with $Q = 2$ genes. To control the expression in every environment, TFs for Q required genes are present, whereas the TFs for the remaining $M - Q$ genes are absent. Because of crosstalk, TFs can bind non-cognate sites, generating a pattern of gene expression that can differ from the one required.

vironment, and in what copy numbers? Cells have seemingly evolved mechanisms to translate upstream signals into ideal TF concentrations, accounting for such sequestration effects of TFs from random binding to DNA and elsewhere [Burger *et al.*, 2010; Sheinman and Kafri, 2012; Weinert *et al.*, 2014]. Even in the presence of such a perfect adaptive tuning of TF concentrations, we will show that a residual level of crosstalk that represents a lower bound or an intrinsic limit, is inevitable. The mechanisms behind such an adaptive tuning of TF concentrations probably involve complex regulatory dynamics with feedback loops, but we do not need to specify the details of these mechanisms as we are interested in characterizing and quantifying intrinsic limits to crosstalk, which cannot be overcome even in the perfectly evolved cell.

We use the grand-canonical ensemble framework from Chapter 1 to describe TF binding to various binding sites. In this sequence-based framework, the strength of TF-BS binding, which determines the gene regulatory state of the target gene, depends on the mismatch between the BS sequence and the consensus sequence of the TF. Molecules of the TFs can also bind other sequences on the DNA, either in a sequence-specific configuration or sequence-independent non-specific configuration. As described before, this has two kinds of effects: sequestration of TF molecules from free solution, decreasing their free concentration, and by binding to non-cognate BS sequences, incorrectly regulate the expression of other target genes.

In the grand-canonical ensemble framework of TF-BS binding, which accounts for all possible pairs of interactions between TFs and BSs, we compute crosstalk, X, as the average fraction of all genes in erroneous regulatory states. Crosstalk ranges between zero, which corresponds to no erroneous regulation, and one, which corresponds to the state with every gene being mis-regulated. We define three different kinds of erroneous regulatory states: (a) genes that should be expressed in a certain environment but are not, because their cognate TFs have not bound to their BSs, (b) genes that should not be expressed in a certain environment but are incorrectly activated by the binding of non-cognate TFs to the genes' BSs, and (c) genes that should be expressed, but are activated due to the incorrect binding of non-cognate (instead of cognate) TFs. We consider the third state – "activation out-of-context" – to be an erroneous regulatory state because activation by non-cognate TFs, even when the gene is required, might deviate the level of the gene's expression. However, we relax this assumption and not consider such states as an error in Section 2.6.

We quantify these three types of erroneous states using the following 2 contributions to crosstalk:

1. For a gene i that should be expressed and whose cognate TF is therefore present, the possible erroneous states are (a) activation out of context: its binding site is bound by a non-cognate TF, and (b) gene is not expressed: its binding site is unbound. The gene is an erroneous state with probability

$$x_1^i(\{C_j\}) = \frac{e^{-\mu_0} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}{C_i + e^{-\mu_0} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}, \tag{2.1}$$

where C_j is the concentration of the jth TF, d_{ij} is the number of mismatches between the jth TF consensus sequence and the binding site of gene i, ϵ the energy per mismatch and μ_0 the chemical potential contribution from sequestration on the DNA and in free solution.

2. For a gene i that should not be expressed and whose cognate TF is therefore absent, the only erroneous state is erroneous activation: its binding site being bound by a non-cognate regulator rather than remaining unbound. This happens with probability

$$x_2^i(\{C_j\}) = \frac{\sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}{e^{-\mu_0} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}. \tag{2.2}$$

These errors x_1 and x_2 depend on the likelihood of TFs to bind non-cognate sites, which is determined by the specific set of pairwise distances d_{ij} between TF consensus sequences and BS sequences. When all sequences are considered together, making a particular sequence less similar to the rest of the sequences can only happen at the cost of making the rest of the sequences more similar among themselves. The errors depend on the arrangement of sequences in the sequence space, with the errors of one particular gene i depending on the distances d_{ij} $\forall j \neq i$ with the rest of the BS sequences. An important quantity is the binding similarity measure S_i between the binding site sequence of gene i and all others, defined as:

$$\sum_{j\neq i} C_j e^{-\epsilon d_{ij}} := CS_i(\epsilon, L). \tag{2.3}$$

To make the analysis easier, we assume a fully symmetric setup such that x_1^i and x_2^i are independent of i. We assume that the set of distances d_{ij} $\forall j \neq i$ of i with the rest of the sequences, are distributed according to some probability density $P(d)$, independent of i. Such a mean-field-like assumption is reasonable for $Q \gg 1$ when the sequences are randomly distributed in sequence space. With this assumption, we now have:

$$S_i(\epsilon, L) \approx \sum_d P(d) e^{-\epsilon d}, \tag{2.4}$$

where $P(d)$ is the distribution of distances between BS sequences and C is the total concentration of all TFs.

The average similarity S depends only on the binding site sequences, but it carries no functional meaning in the absence of any TF, when $C = 0$. It is important to note that this quantity, S, is not arbitrary, and in fact emerges from equations Eq. 2.1 and 2.2. An analogous measure has been previously introduced and measured in olfaction and immune recognition [Lancet *et al.*, 1993], of the probability of receptors to bind an arbitrary ligand from a large repertoire. In our model, S_i is proportional to the probability of the i-th TF to bind any non-cognate binding site. Similarity is highest, $S = 1$, if all sites are identical, and it is its lowest, $S \approx 0$, if the sites are maximally separated from each other. Short binding sites (small L) and weaker binding energy E result in larger S making the sites less distinguishable (Fig. 2.2). As similarity increases, non-cognate interactions increase and hence, crosstalk also increases.

Binding site similarity $S(\epsilon, L)$ of Eq. 2.4 could be experimentally measured by probing the average TF-binding affinity to a large repertoire of known binding sites. Alternatively, S can be estimated from bioinformatic data, which we explore in Sec. 2.7. Under certain assumptions about how binding sites are organized in sequence space, S can be also computed theoretically. For instance, if the binding sites were random sequences of length L, the following analytical expression for S can be derived:

$$S(\epsilon, L) = \sum_d P(d) e^{-\epsilon d} \tag{2.5}$$

$$= \sum_{d=0}^{L} \binom{L}{d} \frac{3^d}{4^L} e^{-\epsilon d} \tag{2.6}$$

$$= \left(\frac{1}{4} + \frac{3}{4} e^{-\epsilon}\right). \tag{2.7}$$

This expression for $S(\epsilon, L)$ is same as that for $\langle e^{-E(s_i)} \rangle_{P_{\text{rand}}}$ from Chapter 1. In Section 2.7, I will explain how we studied more realistic models of BS sequences' organization in sequence space. These different variations of sequence arrangement in sequence space change only the value of S. Hence the same crosstalk formalism can be applied, and we use S directly as a parameter to compute the various quantities of interest.

Further, to factor in different environmental states in the computation of crosstalk, we assume that different subsets of Q TFs are equally likely to occur. The overall crosstalk, X, defined as the average fraction of all genes in erroneous regulatory states, is defined as

$$X(Q, M, x_1, x_2) = x_1 \frac{Q}{M} + x_2 \frac{M - Q}{M}. \tag{2.8}$$

where Q is the number of TFs present (genes that need to be expressed), M is the total number of genes that can be potentially activated, and x_1 and x_2 are the errors under the mean-field assumption of BS sequences.

The expressions for x_1 and x_2 read:

$$x_1 = \frac{e^{-\mu_0} + CS}{\frac{C}{Q} + e^{-\mu_0} + CS} \tag{2.9a}$$

$$x_2 = \frac{CS}{e^{-\mu_0} + CS}. \tag{2.9b}$$

As concentration of the TFs varies between the extremes of zero or being very large, x_1 and x_2, and hence the level of overall crosstalk vary (see Table 2.1).

1. $C = 0$: When the TF concentration is zero, the only erroneous states are those in which genes which should be expressed are not because their cognate TFs do not bind to the corresponding BSs. Hence, we have $x_1 = 1$ and $x_2 = 0$, with the total error being the fraction of genes that need to activated, $X(C = 0) = Q/M$.

2. $C \to \infty$: In this case, when the TF concentration is very large, no BS is left unbound, so all genes are always activated. We have, $x_1 = SQ/(1 + SQ)$ and $x_2 \approx 1$. The error x_1 depends on S, how similar the binding sites are, with the total

crosstalk being $X(C \to \infty) = 1 - \frac{Q/M}{1+SQ}$. If $SQ \ll 1$, crosstalk can be approximated as $X \approx 1 - \frac{Q}{M}(1 - SQ)$.

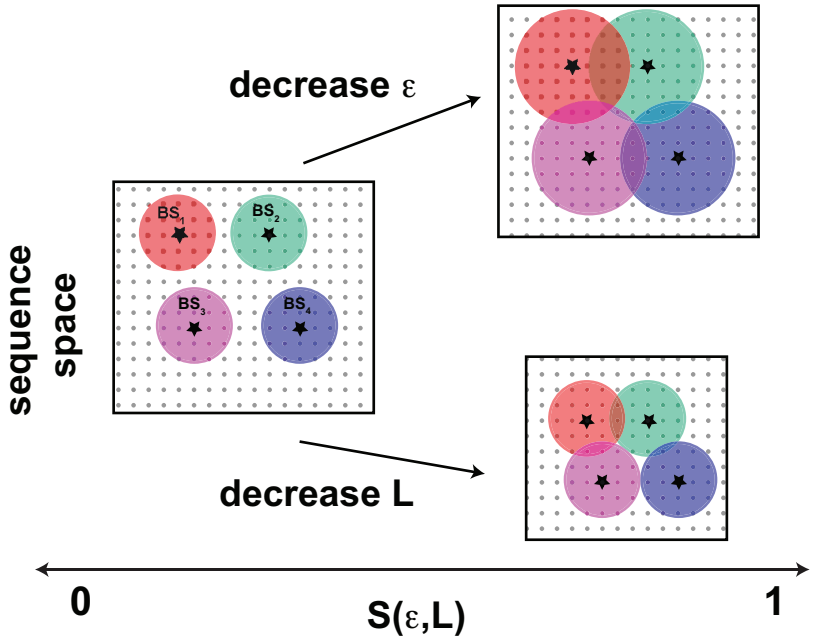


Figure 2.2: **Binding site similarity S is a basic determinant of crosstalk.** Binding site similarity, $S(\epsilon, L)$, determines the likelihood that a TF will bind non-cognate sites, if recognition sequences are of length L and the energy per mismatch is ϵ. A schematic diagram of sequence space packing by different TFs: sequences (dots) in a coloured circle are likely to be bound by the TF whose consensus is the circle's centre star. Smaller L contracts the sequence space and makes crosstalk (circle overlap) more likely (larger S); crosstalk is increased (larger S) also by smaller ϵ, which expands the circle radius.

2.3 Basic crosstalk model exhibits three regulatory regimes

The major determinants of crosstalk are the number of genes typically co-activated, Q, the total number of regulated genes, M, the binding site similarity S, and the total concentration of TFs, C. The first three are easy to estimate and specify but it is harder to determine an appropriate value for C. Not only is just a limited amount of data available, but also concentrations dynamically change in various conditions differentially for various TFs. So, instead of specifying a concentration *a priori*, we compute the optimal concentration, C^*, that minimizes crosstalk.

Such an optimal TF concentration, C^*, arises out of a trade-off between the Q genes that need to be active (x_1), for which a higher C is favored, and the $M - Q$ genes that need to be inactive (x_2), for which the opposite holds (Fig. 2.3). There is an asymmetry between the two kinds of errors: while x_2 can be completely suppressed by having no

	x_1	x_2	crosstalk, X
	$\frac{e^{-E_a}+CS}{\frac{C}{Q}+e^{-E_a}+CS}$	$\frac{CS}{e^{-E_a}+CS}$	$\frac{Q}{M}x_1 + \frac{M-Q}{M}x_2$
$C = 0$	1	0	Q/M
$C = \infty$	$\frac{SQ}{1+SQ}$	1	$1 - \frac{Q/M}{1+SQ}$
optimal C; only activators	$\frac{1+QZ}{1+Z/S+QZ}$	$\frac{QZ}{1+QZ}$	$\frac{Q}{M}\frac{1+QZ}{1+Z/S+QZ} + \frac{M-Q}{M}\frac{QZ}{1+QZ}$

Table 2.1: **Crosstalk errors in the basic model.** Per-gene errors of the two types: x_1 is the error of a site whose cognate TF exists and the site should therefore be bound, but is either unbound or bound by a non-cognate factor. x_2 is the error of a site whose cognate factor does not exist, and the site should therefore be unbound, but is bound by a non-cognate factor. The last column shows the total crosstalk, averaged over all M sites.

TFs, $C = 0$, the opposite does not hold as x_1 cannot be completely eliminated even for infinitely high C because there is always a residual cross-activation. Also, both x_1 and x_2 increase with the similarity S as higher similarity among BS sequences means that there is more frequent cross-binding.

The minimal crosstalk, $X^* = X(C^*)$, at the optimal concentration, is the value beyond which the cell cannot improve by tuning the TF concentrations, and can be computed analytically using the mean-field-like approximation.

Taking the derivative of X and solving for its zeros,

$$\frac{\partial}{\partial C}X(Q, M, x_1, x_2)\Big|_{C^*} = 0,$$

we find two potential extrema

$$C^*_{1,2} = \frac{Qe^{-\mu_0}\left(S(SMQ - Q(SQ+2) + M) \pm \sqrt{S(M-Q)}\right)}{S\left(-M(SQ+1)^2 + SQ^2(SQ+3) + Q\right)},$$

but only one of them can yield non-negative concentration values (and is consistently a minimum):

$$C^* = \frac{Qe^{-\mu_0}\left(S(SMQ - Q(SQ+2) + M) - \sqrt{S(M-Q)}\right)}{S\left(-M(SQ+1)^2 + SQ^2(SQ+3) + Q\right)}. \tag{2.10}$$

For small S the leading terms in the optimal concentration are

$$C^* = \frac{e^{-\mu_0}Q}{\sqrt{S(M-Q)}} - \frac{e^{-\mu_0}Q(M-2Q)}{M-Q} - \frac{e^{-\mu_0}Q^2(2M-3Q)\sqrt{S}^{3/2}}{M-Q} + O[S]. \tag{2.11}$$

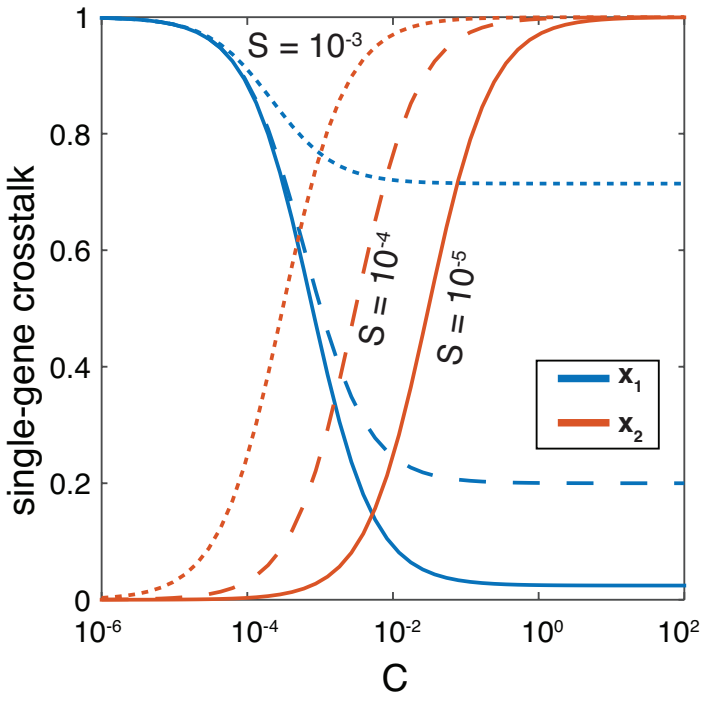


Figure 2.3: **Optimal TF concentration C^*.** x_1 crosstalk component (genes that should be active) decreases with TF concentration C, whereas x_2 crosstalk component (genes that should remain inactive) shows the opposite trend. Curves of x_1 and x_2 (crosstalk of a single gene) vs. C are illustrated for various values of S. While x_2 can be fully eliminated if $C = 0$, x_1 has a residual component which depends on S even for infinite C. Both crosstalk types increase with the similarity between the binding sites S (compare curves with various S values).

Substituting Eq. (2.10) back into Eq. (2.8) yields the minimal achievable crosstalk:

$$X^* = \frac{Q}{M}\left(-S(M-Q) + 2\sqrt{S(M-Q)}\right). \tag{2.12}$$

For constant number of co-activated genes Q, X^* increases to leading order like the square root of S

$$X^* = \frac{2Q\sqrt{M-Q}}{M}\sqrt{S} + O[S]. \tag{2.13}$$

Substituting C^* into the single gene crosstalk expressions Eqs. (2.1)-(2.2), we obtain the minimal per-gene crosstalk

$$x_1^* = \sqrt{S(M-Q)} \tag{2.14a}$$

$$x_2^* = SQ\left(\frac{1}{\sqrt{S(M-Q)}} - 1\right). \tag{2.14b}$$

With these expressions at hand, we investigate how optimal crosstalk, X^*, varies with the number of co-expressed genes, Q, and the binding site similarity, S, for a fixed number

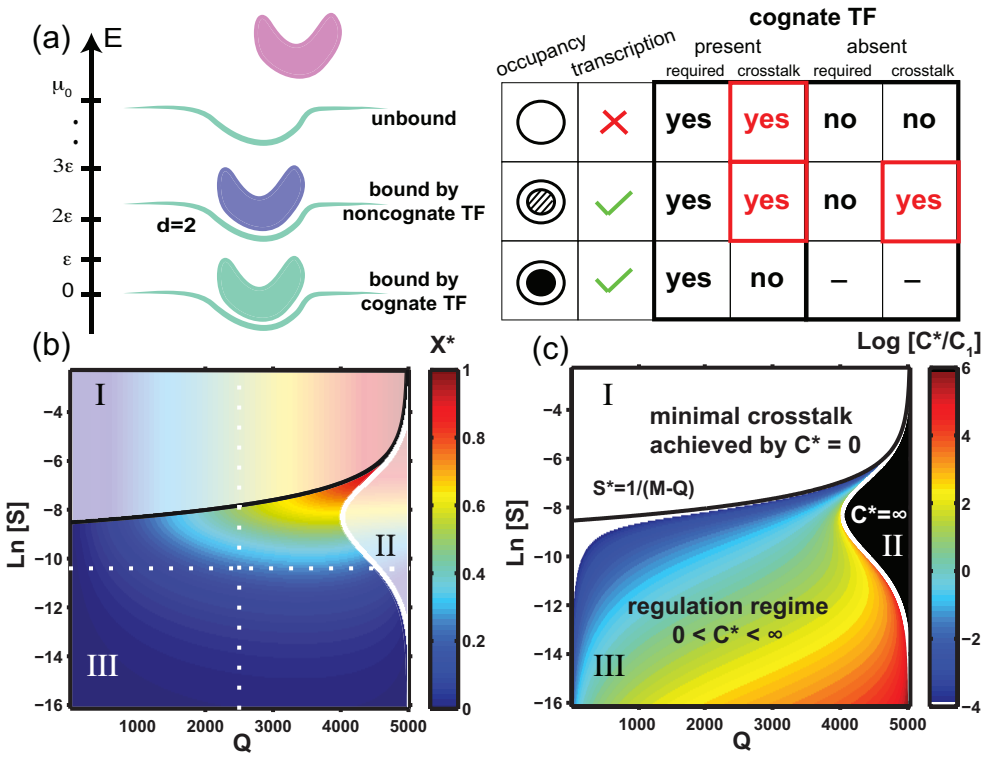


Figure 2.4: **Basic model with one activator binding site per gene exhibits three distinct regulatory regimes.** **(a)** Each binding site can be in either of the three possible states with different corresponding energies: bound by a cognate factor ($E = 0$, green molecule), bound by a non-cognate factor with d- mismatches ($E = \epsilon d$, here a blue molecule with $d = 2$), or unbound ($E = E_a$, pink molecule). The table shows which of these states lead to transcription and which of these outcomes is considered as crosstalk when the cognate TF is present and the gene is required to be active (left), or if it is absent and the gene is required to be inactive (right). **(b)** Minimal crosstalk X^*, shown in colour, as a function of the number of co-activated genes Q and binding site similarity, S. Three different regulatory regimes are separated by black and white boundary lines, identical between **b** and **c**. Dotted lines refer to the "baseline parameters" ($Q = 2500, M = 5000, \ln(S) = -10.5$ - represents $L = 10, \epsilon = 2$ with $d_{\min} = 2$) that we use in all subsequent figures if not specified differently. **(c)** Optimal TF concentration, C^*, that minimizes the crosstalk, relative to C_1, the optimal concentration at baseline parameters. For high binding site similarity (large S), the crosstalk is minimized at $C^* = 0$ (white region, I: "no regulation regime"). For $Q \to M$ and intermediate S, the crosstalk is minimized at $C^* \to \infty$ (black region, II: "constitutive regime"). In a large, biologically plausible intermediate regime, crosstalk is minimized at a finite non-zero TF concentration (colour, III: "regulation regime").

of total genes, $M = 5,000$. Various mathematical constraints on the expressions for optimal concentration and optimal crosstalk result in three distinct regulatory regimes in the (Q, S) plane.

1. **Region I**: This regime, called the "no regulation regime", is characterized by a vanishing optimal TF concentration, $C^* = 0$. This occurs for larger values of S when BS sequences are very similar, and regulation is so non-specific that there is always significant non-cognate TF-BS binding for any non-zero TF concentration. This regime occurs for $S > 1/(M - Q)$, mathematically arising of the constraint that optimal crosstalk from Eq. 2.14a should be in the range $[0, 1]$; the threshold similarity above which this regime occurs decreases as the number of genes to be silent increases. While this regime is dysfunctional and biologically implausible, it offers the insight that a fundamental limit to the similarity, S, of BS sequences is set by the typical number of genes to be inactive, $M-Q$, in each environment, as evident from the condition, $S > 1/(M - Q)$, that defines this regime. This highlights how gene regulatory requirements, which in this case is to keep undesired gene activation levels low in the presence of crosstalk, can constrain regulatory systems and put fundamental limits on their global design.

2. **Region II**: As the number of co-activated genes, Q, in each environment increases, the optimal concentration, C^*, also increases, and finally formally diverges, $C^* \to \infty$. This arises out of the constraint on optimal concentration from Eq. 2.10 to be non-negative. The boundary curve in the (Q, S) plane can be obtained as the two of the roots of the 4^{th} order equation in Q: $S(M+SMQ-2Q-SQ^2)-\sqrt{S(M-Q)} = 0$. In contrast to Region I, this corresponds to a biologically plausible scenario of constitutively expressing all the genes rather than relying on transcriptional regulation. Organisms like obligatory parasites, which live in nearly constant environments, might adopt this strategy.

3. **Region III**: This regime, called the "regulation regime", corresponds to the biological picture of transcriptional regulation, occurs in a broad region of the (Q, S) plane, and is chracterized by a finite positive C^*, which minimizes crosstalk, given by the expression in Eq 2.12. The boundary between the first and third region is at $S^* = \frac{1}{M-Q}$ and the boundary between the second and the third is at $S^* = \frac{-2M+3Q\pm\sqrt{Q(5Q-4M)}}{2Q(M-Q)}$. Hence, the second region (where $C^* = \infty$) only applies for $Q > \frac{4M}{5}$.

Optimal crosstalk, X^*, is independent of the energy difference, μ_0, between the cognate and the unbound state. Increasing this energy difference only lowers the optimal concentration, C^*, while leaving the crosstalk unchanged. Optimal crosstalk depends both on the fraction of genes that need to be co-activated, Q/M, and the total number of genes that need to be inactive, $M - Q$. This suggests that it is costly to maintain genes that are never expressed, as they will be frequently spuriously expressed, arguing against unlimited accumulation of obsolete genes in organisms.

Optimal crosstalk in the regulation regime is dominated by the second term of Eq. 2.12,

and thus increases as $\sim \sqrt{S}$ and as $\frac{Q}{M}\sqrt{M-Q}$ for sufficiently small S. At the boundary between regions I and III, where regulation breaks down, we have $S(M-Q)=1$, with the crosstalk in region I being independent of S as $X^* = Q/M$, because all genes that need to be active are in a crosstalk state due to the absence of TFs, with the rest being inactive as required. To be in the regulation regime (region III), $S < 1/(M-Q)$, which sets an upper bound on the total number of genes M as $M < 1/S$, corresponding to the maximum number of genes the organism can accumulate given a certain similarity of BS sequences.

As can be seen from Fig. 2.4b, for a given value of S, an intermediate value of Q such that $0 < Q < M$ results in the largest optimal crosstalk. This arises out of the fact that the two crosstalk types, x_1 and x_2, show an opposite dependence on the number of active genes Q. While x_1 (genes that need to be active) decreases with Q, x_2 (genes that need to be silent) increases with Q, because of the change in C^* with Q. As the total crosstalk is a weighted sum of these errors with varying weights that depend on Q, X^* has a non-monotonic dependence on the number of active genes Q with a maximum at an intermediate value - see Fig. 2.5.

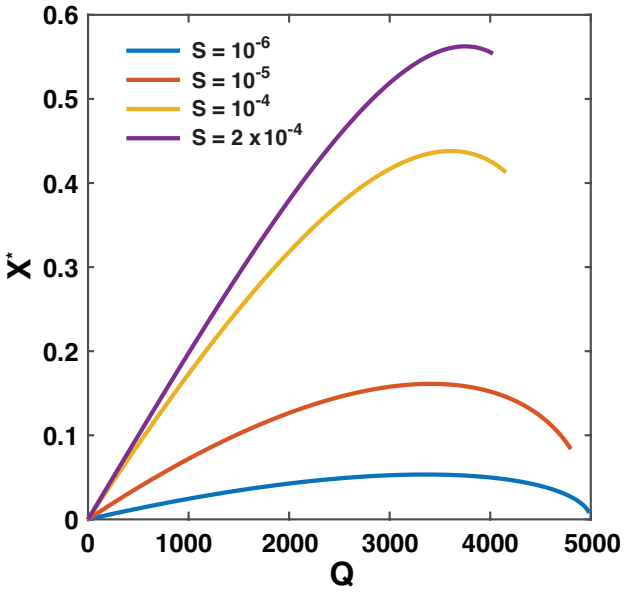


Figure 2.5: **Minimal crosstalk X^* is an increasing function of the similarity S and has a non-monotonic dependence on the number of active genes Q.** The balance between genes that need to be active (x_1 crosstalk type) and genes that need to remain inactive (x_2 crosstalk type) causes a non-monotonic dependence of the *total crosstalk* on the number of active genes Q, which has a maximum at an intermediate Q value. Curves are shown only in the regulation regime, where crosstalk is minimized by a finite TF concentration. The curves are truncated at the point of transition to regime II where TF concentration formally diverges to infinity.

Fig. 2.4b also suggests a surprising insight that crosstalk in the basic model is surprisingly high for an organism of $M = 5,000$ genes. For instance, at $Q = M/2$, when typically

about half of the genes are activated in each environment, and with TF specificity typical of metazoans ($\log(S) = -10.5$), which we call the "baseline" parameters, optimal crosstalk is $X^* \approx 0.23$. This implies that almost a quarter of the genes at any point in time are in an erroneous regulatory state. Also, as Fig. 2.6, plotted for $M = 20000$, a typical number of eukaryotic genes, suggests, a larger M results in increased crosstalk at the same value of S. This suggests that global crosstalk is a serious constraint, and that more complex regulatory mechanisms that connect TF-BS binding to transcriptional activation have evolved, at least in part, to permit reliable regulation despite non-cognate TF binding.

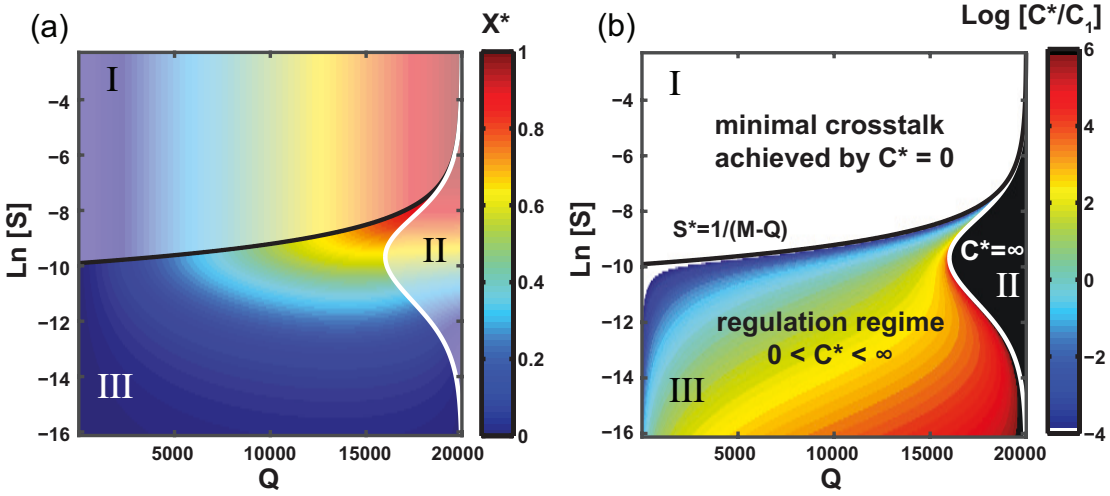


Figure 2.6: **Crosstalk in the basic model for $M = 20000$ genes that are regulated. (a)** Minimal crosstalk, X^*; **(b)** Optimal TF concentration, C^*. These results are analogous to Fig. 2.4, which is computed for $M = 5000$. The results for two different M are qualitatively similar and show 3 different regimes of regulation. We make the following observations: **(i)** for larger M, the $C^* = 0$ regime expands to include lower S values, as expected from the analytical solution for the regime boundaries; **(ii)** if the fraction of co-activated genes, Q/M, remains constant, the crosstalk *increases* with M, as it also depends on the absolute number of inactive genes $M - Q$ (see Eq. (2.12)). The discrepancies at small Q between the black solid curve separating the "no regulation" and "regulation" regimes, and the numerically computed C^* values are due to the approximation $Q - 1 \approx Q$.

We have assumed that gene regulation is achieved by using specific TF activators to drive the expression of genes that would otherwise remain inactive. An alternative formulation of the problem postulates that genes are strongly expressed without TFs bound to their regulatory sites, but need to be repressed by the binding of specific regulators to stop their expression. In this complementary model, in which all regulators are repressors instead of activators, results (Fig. 2.7) are a mirror image of the results shown in Fig. 2.4b for the activator-only basic model. They can be obtained simply by mapping $Q \to M - Q$. Since we keep the convention that Q is the number of genes that are active, the difference in regulation strategies amounts to having either Q activator types and keeping $M - Q$

binding sites unbound (activator-only) or having $M-Q$ repressor types and keeping Q binding sites unbound.

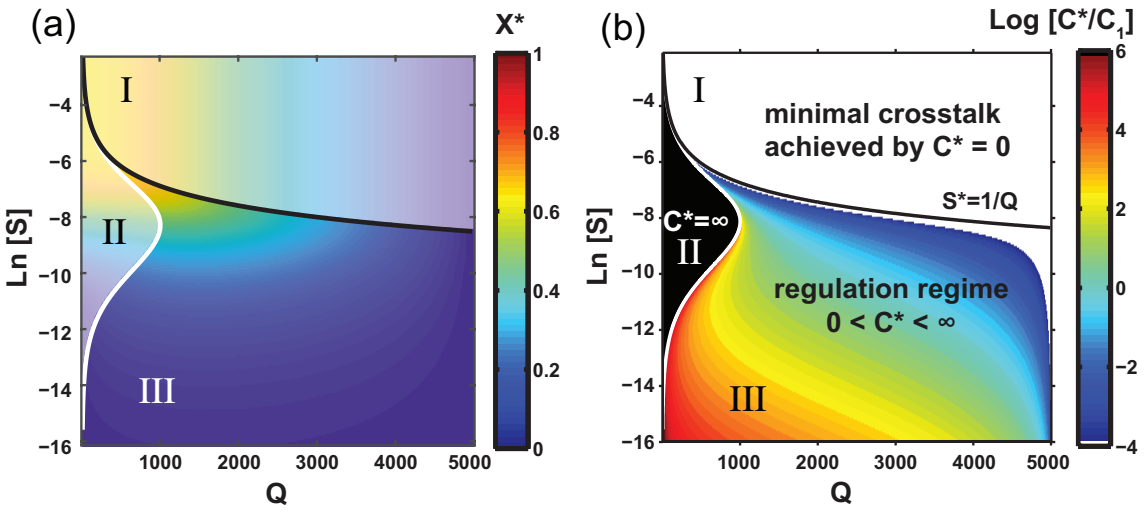


Figure 2.7: **Crosstalk in the basic model with regulation by repressors alone is a mirror image of regulation with activators only.** **(a)** Minimal crosstalk, X^*; **(b)** Optimal TF concentration, C^*. These results are analogous to Fig. 2.4, which is computed for regulation with activators only. The observed picture is an exact mirror image of Fig. 2.4, namely Q maps to $M-Q$, where we keep the convention that Q denotes the number of genes that should be active. The difference is that in the activator-model activating Q genes requires Q types of activators, whereas in the repressor model this requires $M-Q$ types of repressors.

2.4 Validity of the mean-field assumption

In computing crosstalk at a given M and Q, we have made a mean-field assumption on the similarity measure S. Given a set of M binding site sequences of length L in sequence space of size 4^L, this amounts to assuming the following about the distribution of mismatches between pairs of binding site sequences. The distribution of all mismatches (with the $M-1$ other BS sequences) corresponding to each binding site comes from the same underlying distribution, independent of the binding site considered. For a particular selection of Q genes, for each binding site i from the M binding sites, similarity S_i can be defined using d_{ij} where $j \neq i$ indexes over the binding sites of the Q selected genes.

$$S_i = \sum_{j \neq i} e^{-\epsilon d_{ij}}. \tag{2.15}$$

From this, we have for crosstalk for this particular selection of Q genes,

$$X(\{S_i\}) = \frac{1}{M}\Big[\sum_{i\in Q} x_1(S_i) + \sum_{i\in M-Q} x_2(S_i)\Big] \tag{2.16}$$

$$= \frac{1}{M}\Big[\sum_{i\in Q} \frac{e^{-E_a} + CS_i}{C/Q + e^{-E_a} + CS_i} + \sum_{i\in M-Q} \frac{CS_i}{e^{-E_a} + CS_i}\Big], \tag{2.17}$$

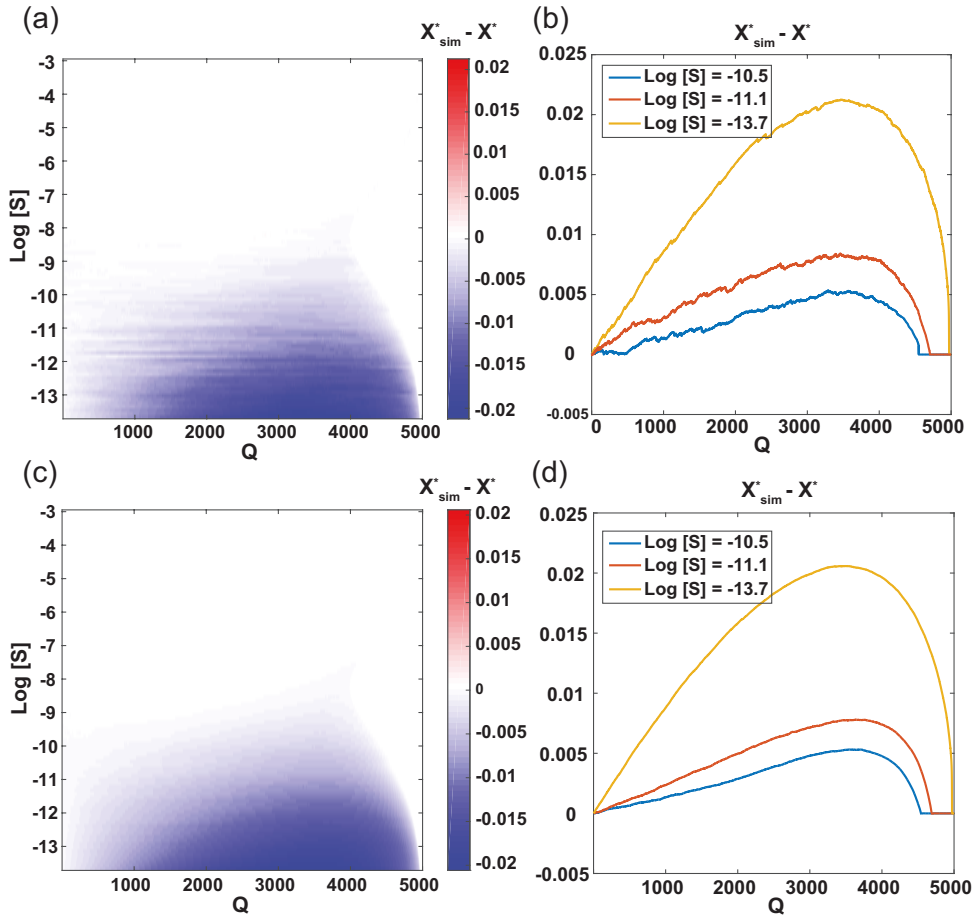


Figure 2.8: **Comparison of mean-field and simulations.** **(a)** and **(c)** We plot the difference in optimal crosstalk between simulations and the mean-field approach, $X^*_{sim} - X^*$ for different Q and S. **(b** and **(d)** We plot $X^*_{sim} - X^*$ against Q for three different S. Here, $M = 5000$, $L = 10$, and S has been varied by tuning ϵ. X^*_{sim} is a Monte Carlo estimate of the mean crosstalk, obtained over n_{sel} different selections of Q out of M genes. $n_{sel} = 1$ in the top row, and $n_{sel} = 30$ in the bottom row. The mean-field approach is in general a very good approximation of the simulations. The maximal crosstalk difference is 0.02. At smaller S, the difference is larger.

where $x_1(S_i)$ and $x_2(S_i)$ depend on S_i as shown. We're interested in the mean crosstalk $X = \langle X(\{S_i\})\rangle$ over all selections of Q out of M genes, which is equivalent to the joint distribution of all S_i. Each S_i comes from the same underlying distribution with mean S. So we have,

$$X = \langle X(\{S_i\})\rangle = \frac{1}{M}\Big[\sum_{i\in Q}\langle x_1(S_i)\rangle + \sum_{i\in M-Q}\langle x_2(S_i)\rangle\Big]. \tag{2.18}$$

In the mean-field assumption, we have $\langle x_1(S_i)\rangle \approx x_1(\langle S_i\rangle) = x_1(S)$ and $\langle x_2(S_i)\rangle \approx x_2(\langle S_i\rangle) = x_2(S)$ which gives us

$$X = \frac{Q}{M}x_1(S) + \frac{M-Q}{M}x_2(S). \tag{2.19}$$

From this, one can obtain the optimal crosstalk X^*. To check the validity of such a mean-field assumption, we performed simulations by picking binding sites from the sequence

space and computing optimal crosstalk X^*_{sim}, and compared this with the mean-field crosstalk X^*. To do this, we first picked M binding sites (genes) randomly from the sequence space and fixed them. Now, for each Q, we perform n_{sel} different selections of Q out of M genes. For each such selection, after computing the binding site mismatches and occupancies, we compute the crosstalk. To get the mean crosstalk for Q, we perform a Monte Carlo estimate of the mean crosstalk over these n_{sel} different selections of Q out of M genes. We see that the mean-field crosstalk systematically over-estimates the actual crosstalk, but is nevertheless a very good approximation to it.

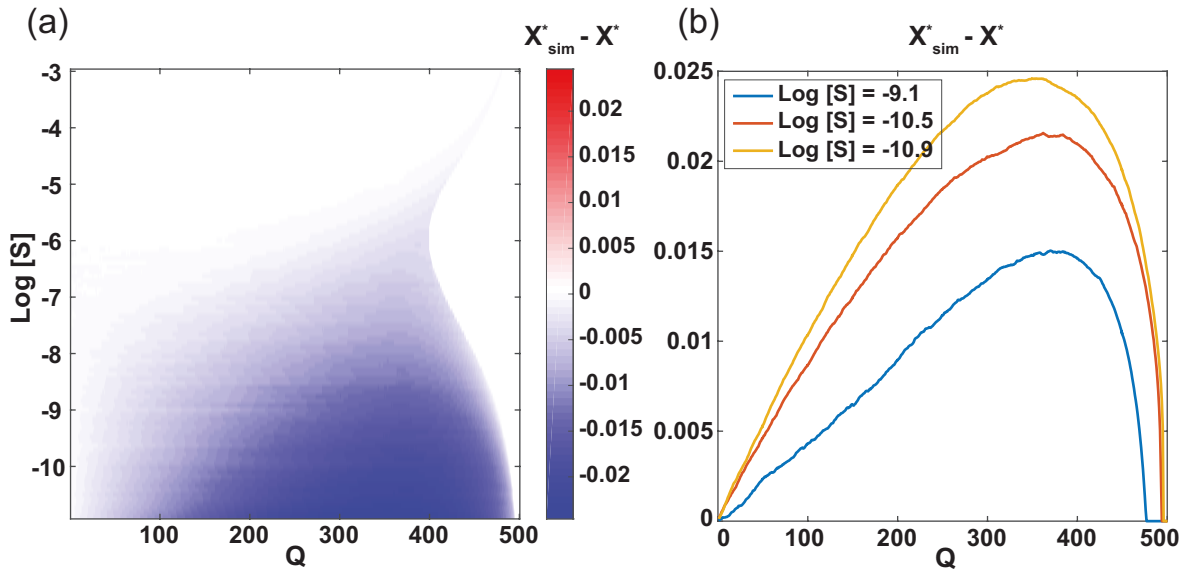


Figure 2.9: **Comparison of mean-field and simulations. (a)** We plot the difference in optimal crosstalk between simulations and the mean-field approach, $X^*_{sim} - X^*$ for different Q and S. **(b)** We plot $X^*_{sim} - X^*$ against Q for three different S. Here, $M = 500$, $L = 8$, and S has been varied by tuning ϵ. X^*_{sim} is a Monte Carlo estimate of the mean crosstalk, obtained over $n_{sel} = 100$ different selections of Q out of M genes. Again, as with $M = 5000$, the mean-field approach is a very good approximation of the simulations. The maximal crosstalk difference is only slightly larger than 0.02.

2.5 Mixed models of activators and repressors

In the baseline models, we have M genes, all of which are regulated by only activators or only repressors. Here, we consider mixed models, where some genes are regulated by activators and the other genes by repressors. Here, we assume that M_A genes are regulated by activators and M_R genes are regulated by repressors. We have $M = M_A + M_R$. In a particular environment, we assume that Q genes need to be ON. Out of these, we assume that Q_A genes are regulated by activators and Q_R genes are regulated by repressors. Hence, we have $Q = Q_A + Q_R$. For the Q_A genes (out of M_A activator-regulated genes) that need to be activated (and hence, ON), we have Q_A activators present in the cell. For the Q_R genes (out of the M_R repressor-regulated genes) that need to be not repressed

(and hence, ON), we don't have their repressors present. But for the rest $M_R - Q_R$ genes (out of the M_R repressor-regulated genes) that need to be OFF, we have $M_R - Q_R$ repressors present in the cell. So, a total of $T = Q_A + M_R - Q_R$ TFs are present in the cell. As before, S is the similarity of the binding sites and C the total concentration of TFs (activators+repressors). The concentration of a particular TF type, when present, will now be C/T. We assume that any non-cognate interaction ("activation out-of-context" or "repression out-of context") contributes towards crosstalk error.

For a gene that is activated and needs to be ON, it needs to be bound by the cognate activator. The unbound state and any non-cognate binding (non-cognate activator or repressor) are crosstalk states.

$$x_1^A = \frac{e^{-E_a} + CS}{\frac{C}{T} + e^{-E_a} + CS} \qquad (Q_A \text{ out of } M \text{ genes}). \tag{2.20}$$

For a gene that is activated and needs to be OFF, it needs to be unbound. Any non-cognate binding is a crosstalk state.

$$x_2^A = \frac{CS}{e^{-E_a} + CS} \qquad (M_A - Q_A \text{ out of } M \text{ genes}). \tag{2.21}$$

For a gene that is repressed and needs to be ON, it needs to be unbound. Any non-cognate binding is a crosstalk state.

$$x_1^R = \frac{CS}{e^{-E_a} + CS} \qquad (Q_R \text{ out of } M \text{ genes}). \tag{2.22}$$

For a gene that is repressed and needs to be OFF, it needs to be bound by the cognate repressor. The unbound state and any non-cognate binding (non-cognate repressor or activator) are crosstalk states.

$$x_2^R = \frac{e^{-E_a} + CS}{\frac{C}{T} + e^{-E_a} + CS} \qquad (M_R - Q_R \text{ out of } M \text{ genes}). \tag{2.23}$$

Notice that $x_1^A = x_2^R$ and $x_2^A = x_1^R$. Now, the overall crosstalk error reads

$$\begin{aligned} X_{mixed,full}(Q_A, Q_R, M_A, M_R) &= x_1^A \frac{Q_A}{M} + x_2^A \frac{M_A - Q_A}{M} + x_1^R \frac{Q_R}{M} + x_2^R \frac{M_R - Q_R}{M} && (2.24)\\ &= x_1^A \frac{M_R + Q_A - Q_R}{M} + x_2^A \frac{M_A + Q_R - Q_A}{M} && (2.25)\\ &= x_1^A \frac{T}{M} + x_2^A \frac{M - T}{M} && (2.26)\\ &= X(Q_{eff} = T, M_{eff} = M). && (2.27) \end{aligned}$$

Hence, given a set of (Q_A, Q_R, M_A, M_R) of the mixed model, crosstalk is same as that in an equivalent baseline activator model with $Q_{eff} = T = M_R + Q_A - Q_R$ and $M_{eff} = M = M_A + M_R$.

Given M, there are many mixed models possible. In each of the mixed models, the cell has different number of genes put under the control of activators (M_A). This can

be tuned on an evolutionary timescale. Once M_A is chosen, different selections of Q genes have different numbers of genes under the control of activators (Q_A) and repressors ($Q_R = Q - Q_A$). For each mixed model (Q_A, Q_R, M_A, M_R), there exists an optimal concentration which depends on the number of TFs, at which one can compute an optimal crosstalk.

For a given M, Q and S, we find the best possible M_A, which minimizes the crosstalk. For some M_A, we define the optimal crosstalk as the average optimal mixed crosstalk for all selections of Q genes (different Q_A),

$$X^*(M, Q, S, M_A) = \sum_{Q_A} P_{Q_A} X^*_{mixed,full}(Q_A, M, Q, S, M_A), \tag{2.28}$$

where P_{Q_A} is the fraction of Q gene selections that have Q_A activated genes.

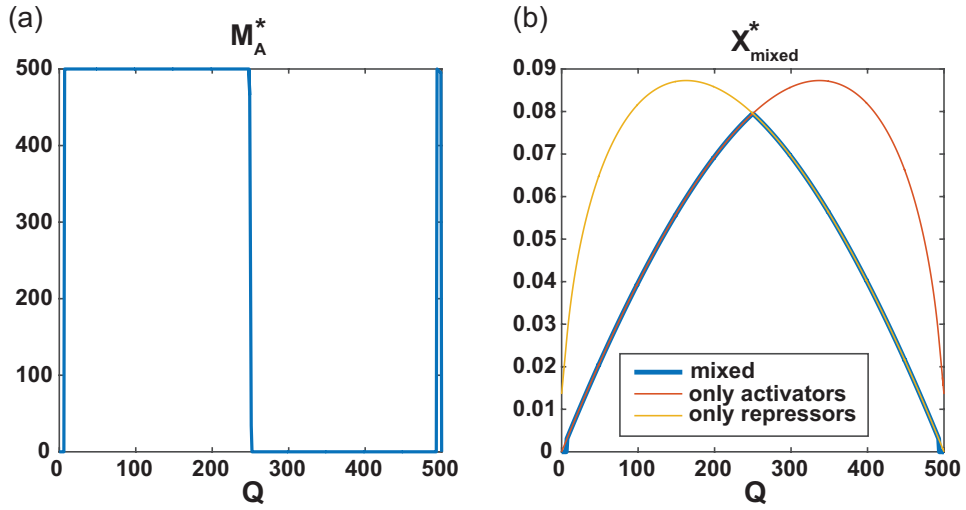


Figure 2.10: **Mixed model at best M_A. (a)** We plot the optimal number of activated genes M_A^* for different Q at $M = 500$ and $\log(S) = -10.5$. For $Q < 250$, it is best to have all under activators ($M_A^* = 500$) and for $Q \geqslant 250$, it is best to have all genes under repressors ($M_A^* = 0$). **(b)** We plot the optimal mixed crosstalk, computed at M_A^*, and averaged over different gene selections using P_{Q_A}.

We have

$$P_{Q_A} = \frac{\binom{M_A}{Q_A}\binom{M-M_A}{Q-Q_A}}{\binom{M}{Q}}, \tag{2.29}$$

$$X^*_{mixed}(M, Q, S) = \min\Big[X^*(M, Q, S, M_A)\Big], \tag{2.30}$$

$$M_A^* = \arg\min_{M_A} X^*(M, Q, S, M_A), \tag{2.31}$$

where M_A^* is the optimal M_A. As in Fig. 2.10, we see that for $Q < M/2$, the best strategy is to use all activators $M_A = M$, and for $Q >= M/2$, the best strategy is to use all repressors.

2.6 Alternative crosstalk definition

So far, we considered "activation out-of-context" — i.e., activation by the binding of a non-cognate TF when the cognate TF is present (but not bound) — to be a crosstalk state. Our reasoning was motivated by viewing transcriptional regulation as a signal transmission apparatus. In this interpretation, gene activation by a non-cognate TF amounts to generating a response (transcriptional activity) to a wrong input signal. Consequently, this should count as crosstalk, despite the fact that (by chance) the correct signal was simultaneously present in the cell. This is perhaps easiest to appreciate if one considers more realistic setups in which genes are not simply "ON" and "OFF", but can be quantitatively regulated by the level of their cognate TF. In such a model, there might be two TFs present and varying in concentration as a function of time: one cognate for the gene of interest and one not. In this case it is clear that the correct response of the gene is to track the changes in the cognate TF, and not to simply be expressed in a constant "ON" state; consequently, tracking the non-cognate TF due to crosstalk is obviously an error, even if the cognate TF is present at the same time.

One could, however, argue that "activation-out-of-context" should not be considered as an error state. If the presence or absence of TF signals is a binary variable and if the binary response is defined solely by the state of transcriptional activity (activation/inactivation of gene), then when the presence of the signal matches the response state, the regulation outcome is correct, irrespective of the molecular details on the promoter. For example, for a gene whose cognate TF is present, activation by any means (either by cognate or non-cognate binding) is the correct response. In this scenario, the "out-of-context activation" is actually what one might call beneficial crosstalk: here, non-cognate TF can be seen as helping to activate the gene when the cognate TF is also present. For a gene whose cognate TF is absent, activation is still an incorrect response, like before.

Hence, $x_2(i)$ retains the same expression, but $x_1(i)$ changes to

$$x_1(i) = \frac{e^{-E_a}}{C_i + e^{-E_a} + \sum_{j \neq i} C_j e^{-\epsilon d_{ij}}}. \tag{2.32}$$

As shown in Fig. 2.11, optimizing C results in three distinct regulatory regimes, like in the default basic setup. For small S in the regulation regime, the optimal C is given to the leading order by:

$$C^* \sim \frac{e^{-E_a}}{\sqrt{S}} \frac{Q}{\sqrt{M-Q}}. \tag{2.33}$$

The minimal crosstalk error at the optimal concentration C^* is given by

$$X^* = -SQ + 2\frac{Q}{M}\sqrt{S(M-Q)(1+SQ)}. \tag{2.34}$$

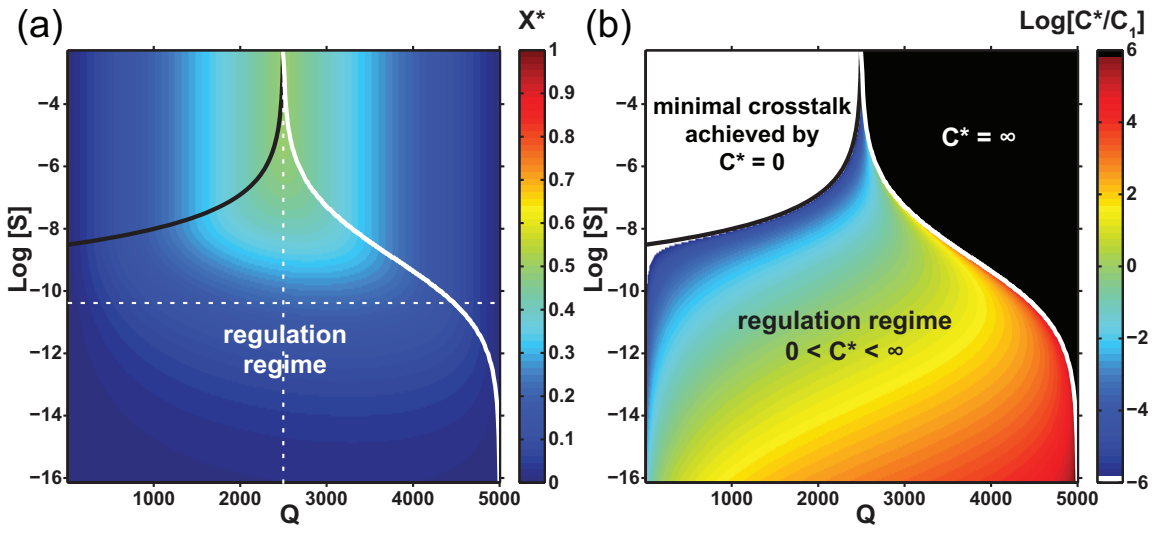


Figure 2.11: **Basic model with alternative crosstalk definition also exhibits three distinct regulation regimes.** The alternative definition does not count "activation out-of-context" as an error state. **(a)** Minimal crosstalk error, X^*, shown in color, as a function of the number of co-activated genes Q, and binding site similarity S. **(b)** Optimal TF concentration C^*, that minimizes the crosstalk, relative to C_0, the optimal concentration at the baseline parameters (see main text).

2.7 Estimating the binding site similarity, S

2.7.1 Optimal packing

In real organisms, binding site sequences for different genes could depart from a random distribution (even after taking into account the statistical structure of the genomic background). For example, to achieve high specificity of regulation, we could hypothesize that binding site sequences evolved to minimize the overlap between any pair of consensus sequences. To explore the crosstalk limit under such optimal use of sequence space and contrast it with the random choice of binding sites in our basic model, we synthetically constructed binding site sequences that are as distinct as possible. Specifically, our optimal codes are described by a parameter $d_{\min}$, which is the minimum required number of basepair differences between any pair of binding site sequences. This is the Hamming distance HD between sequences. The problem of choosing M sequences of length L such that each pair differs by at least $d_{\min}$ is not solvable in general. We construct numerical approximations to these optimal codes using the following algorithm:

1. Generate all possible sequences of length L and store them in a list called *words*. Create an empty list, called *codewords*, which will store the binding site sequences.

2. Pick the first entry, s, from the list *words*, to be a binding site sequence, and append it to the list *codewords*.

3. Erase s and all of its Hamming neighbours at distance strictly less than $d_{\min}$ from the list *words*.

4. If the list *words* is not empty, repeat from step 2. If the list *words* is empty, stop.

When the procedure terminates, the list *codewords* will contain binding site sequences that are separated by at least $d_{\min}$ mismatches. The outcome of this procedure depends on the initial ordering of the list of all possible sequences. The procedure is not guaranteed to generate the maximal set of sequences satisfying the Hamming distance criteria.

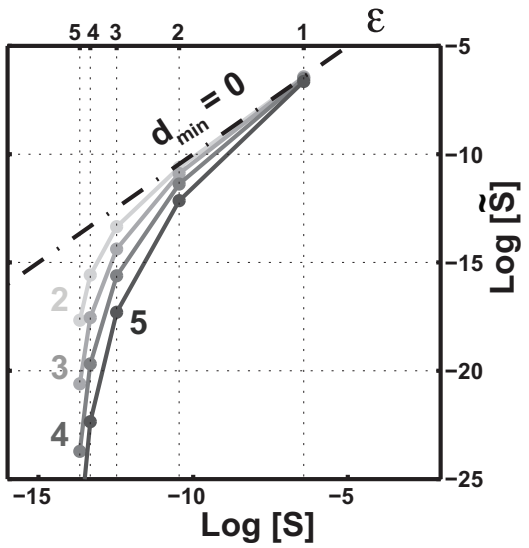


Figure 2.12: **Optimal packing of binding sites in sequence space.** This alternative model with optimal packing of binding sites in sequence space leads to values for $\tilde{S}$ (y-axis) that can be remapped to the $S(\epsilon, L)$ (x-axis) for the random code with the mismatch energy model, $E(d) = \epsilon d$ and $L = 10$ bp binding sites (corresponding scale for ϵ shown in the top axis). Dashed lines denote equality. Optimally designed binding sites effectively decrease S. Here, their sequences are at least $d_{\min}$ bp distant from each other (gray lines = different $d_{\min}$ as indicated).

From the list of generated binding site sequences, we obtain $P(d)$, the distribution of mismatch distances between all pairs of binding sites, and hence obtain the value of S as

$$\tilde{S}(d_{\min}) = \sum_{d \geqslant d_{\min}} P(d) e^{-\epsilon d}. \tag{2.35}$$

$d_{\min} = 0$ corresponds to the "random code" and results in $\tilde{S}(d_{\min} = 0) = S = (\frac{1}{4} + \frac{3}{4} e^{-\epsilon})^L$. Note that increasing $d_{\min}$ decreases the maximum possible M as sequences move further apart in sequence space whose size is fixed. A well-known upper bound on the number of sequences satisfying the Hamming distance criterion is the Singleton bound [Lin and Costello, 2004]: $M(d_{\min}, L) \leqslant 4^{L - d_{\min} + 1}$. As shown in Fig. 2.13, with $L = 8$ and $d_{\min} = 3$, we already have $M \leqslant 4096$. With $L = 10$ and $d_{\min} = 4$, we have $M \leqslant 16384$. As L becomes smaller, the possible range of M also decreases. This suggests that for

prokaryotes, optimally packed binding site sequences can be significantly better than random packing, because they typically have $L > 10$ and $M < 10^4$. On the other hand, eukaryotes have smaller L and larger M, and therefore, might not have enough sequence space for optimal packing to be significantly better than random packing.

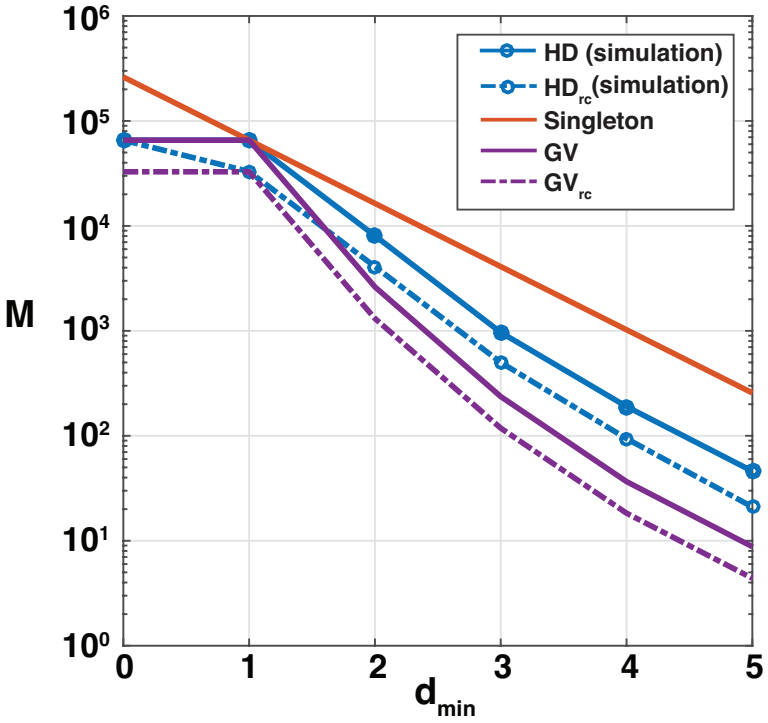


Figure 2.13: **Bounds on the maximal number of binding site sequences for different $d_{\min}$ with binding sites of length $L = 8$.** Two bounds from the coding theory (Singleton upper bound and Gilbert-Varshamov (GV) lower bound [Lin and Costello, 2004]) are shown together with the values of M obtained by our numerical approximation procedure. These are shown both for the usual definition of distance between sequences as the Hamming distance HD as well as for a definition that considers the reverse complements of the sequences HD_{rc}. For $d_{\min} = 0$ there are $M = 4^8 \approx 65000$ possible sequences where all sequence pairs are at least $d_{\min}$ distant from each other, but the number quickly decreases with increasing $d_{\min}$. From the HD to HD_{rc}, the Singleton bound doesn't change from the usual situation but the Gilbert-Varshamov (GV) bound, which takes into account the "volume of restricted ball" around each sequence goes down. Because of stronger constraints, the number of sequences that can be packed goes down from the usual situation but only by a factor of ≈ 2.

2.7.2 Reverse complemented sequences

We also consider a different definition of distance between sequences that considers the double-stranded nature of DNA into account. If a TF that binds a sequence s can also bind its reverse complemented sequence $r = RC(s)$ and thus r cannot be another BS sequence. Hence, one needs to consider the reverse complement of both the sequences in

question. If s_i and s_j are two sequences with reverse complements r_i and r_j respectively, this new definition of Hamming distance is

$$HD_{rc}(s_i, s_j) = \min\Big[HD(s_i, s_j), HD(r_i, s_j), HD(s_i, r_j), HD(r_i, r_j)\Big], \tag{2.36}$$

where $HD(s_i, s_j)$ is the usual Hamming distance as considered previously.

This restricts the sequence space much more than with the usual definition and as such, as seen in Fig. 2.13, we can pack fewer binding sites in the sequence space at a specific $d_{\min}$. In Fig. 2.14, we map S from the reverse complement code to S from a random code. See that S increases by about a factor of ≈ 2 for realistic $\epsilon \in [2, 5]$, because of these stricter constraints.

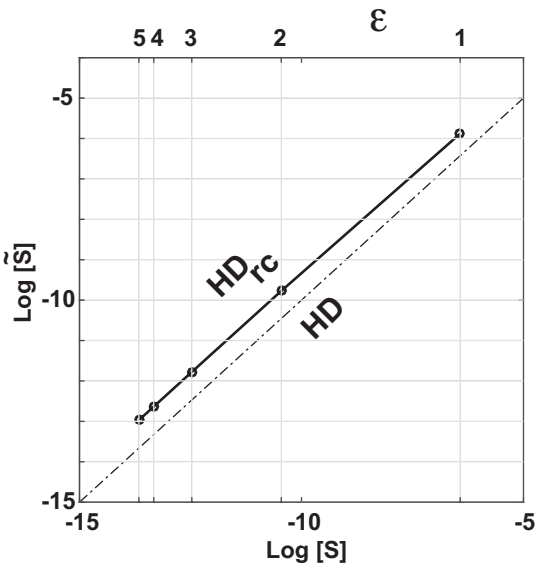


Figure 2.14: **Reverse complemented sequences.** Using an alternative definition HD_{rc} of distance between binding site sequences, which takes into account the double-stranded nature of DNA by considering the reverse complements as well of the sequences in question, leads to values for $\tilde{S}$ (y-axis) that can be remapped to the $S(\epsilon, L)$ (x-axis) for the random code with the usual Hamming distance definition HD. Here, we have considered $L = 8$ bp binding sites (corresponding scale for ϵ shown in the top axis). Dashed lines denote equality. This alternative definition increases S because more sequences are now found in the "shells" around the consensus to which the TF can bind on the reverse strand. S increases by about a factor of ≈ 2 for $\epsilon \in [2, 5]$, and by about a factor of ≈ 1.7 for $\epsilon = 1$.

2.7.3 Saturating model of TF-DNA binding energy

It has been experimentally observed that the binding energy between TF and DNA saturates to some nonspecific value after a certain number of mismatches between the TF's cognate sequence and the DNA sequence in question [Maerkl and Quake, 2007]. We consider such a saturating energy model, characterized by a parameter d_0, the number of mismatches after which binding energy saturates.

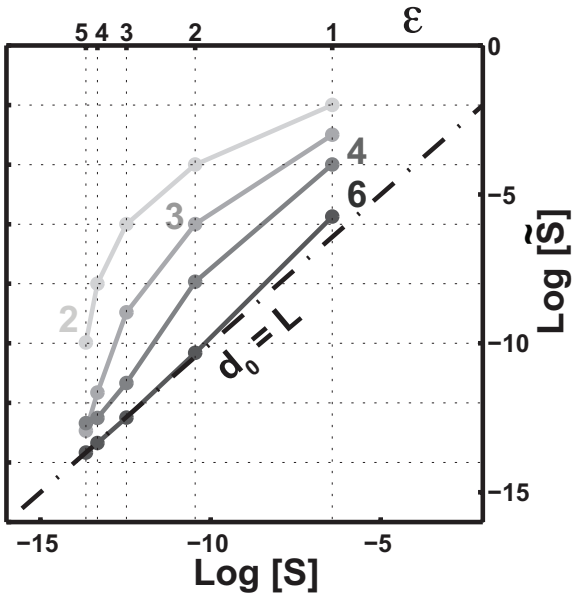


Figure 2.15: **Saturating energy model.** An improved affinity model where the mismatch energy saturates after d_0 mismatches, $E(d) = \epsilon \min(d, d_0)$ (gray lines = different d_0 as indicated), effectively increases S. $d_0 \sim 4$ has been reported experimentally [Maerkl and Quake, 2007]. This alternative model leads to values for $\tilde{S}$ (y-axis) that can be remapped to the $S(\epsilon, L)$ (x-axis) for the random code with the mismatch energy model, $E(d) = \epsilon d$ and $L = 10$ bp binding sites (corresponding scale for ϵ shown in the top axis). Dashed lines denote equality.

The binding energy is given by $E(d) = \epsilon \min(d, d_0)$. We obtain S as

$$\tilde{S}(d_0) = \sum_d P(d) e^{-E(d)}, \tag{2.37}$$

where $P(d)$ is the distribution of mismatch distances between all pairs of binding sites picked at random from the sequence space. $d_0 = L$ corresponds to a mismatch model with non-saturating energy. Decreasing d_0 limits the specificity of the TF towards binding site sequences far away from the consensus and thereby increases $\tilde{S}(d_0)$.

2.7.4 Empirical values

We obtain organism-specific estimates of S from known databases [Gama-Castro *et al.*, 2011; Mathelier *et al.*, 2013; Spivak and Stormo, 2012] of the binding site sequences of different TFs. In the main text, for a particular genome, we defined S for a collection of TFs with the same mismatch penalty ϵ and binding sites of a specific constant length L. In real organisms, different TFs have different ϵ and L, making it difficult to directly calculate S for a genome. Instead we obtain a value of S for each TF by defining it as the value of S of a hypothetical genome in which all TFs have the same binding site properties (ϵ, L) as our TF. Hence, for each organism, we obtain a set of S values.

Many databases document the binding site sequences of TFs in Position Count Matrices (PCMs). The PCM of a TF with a binding site of length L is a $4 \times L$ matrix B with b_{ij} denoting the number of known TF binding site sequences that have nucleotide i in position j. One can obtain estimates of ϵ and L from B, and use them to calculate S. There are two broad ways to estimate ϵ and L (and hence, S) of a TF: (a) Information method, (b) Pseudo-count method. In (a), we calculate the information contained in the whole binding site motif and obtain an ϵ that distributes this information uniformly among all sites in an equivalent "effective" motif that has the same length as the original, but only has 0 or ϵ mismatch energy values. In (b), we obtain ϵ for all entries of the PCM and calculate an average ϵ from these entries. To handle zeros in the PCM which lead to undefined ϵ, (b) uses an arbitrary pseudo-count. Method (a) can, in contrast, avoid the use of pseudo-counts and, additionally, reproduces by construction the information content of each known motif, which is the key statistical property of TF specificity [Wunderlich and Mirny, 2009; Schneider *et al.*, 1986]. Hence, we used (a) to infer S values. In both the methods, we used PCMs that have that have been constructed from at least 10 distinct binding site sequences.

Information method

In this method, we first obtain the binding site length L and also the total information I, contained in the binding site sequences of the TF.

$$I \quad = \sum_j I_j = \sum_j \sum_i p_{ij} \log_2 \frac{p_{ij}}{q_{ij}}, \tag{2.38}$$

where I_j is the information contained in position j, p_{ij} is the frequency of nucleotide i in position j, obtained in a straightforward way from B, and q_{ij} is the expected background frequency. To get rid of non-specific positions, we neglect all positions that contain information less than a certain threshold ($I_j > 0.2$ bits for position j to be considered part of the binding site). For a random genome, $q_{ij} = 0.25 \ \forall \ i, j$, resulting in

$$I = 2L + \sum_{i,j} p_{ij} \log_2 p_{ij}. \tag{2.39}$$

The maximum information in the motif is $2L$ bits (when $\epsilon \to \infty$) with each position contributing a maximum of 2 bits, which for finite ϵ, is reduced by an entropy term. Obtaining information per position $I_{pos} = I/L$, we infer an ϵ that uniformly distributes the information in the motif among individual positions. At a specific position j^*, without loss of generality, assume that $i = 4$ has the best binding energy ($= 0$). The probability of observing $i = 4$ at j^* is given by $p_4 = 1/Z$ while the probability of observing any of the three other possible nucleotides is given by $p_{1,2,3} = e^{-\epsilon}/Z$, with $Z = 1 + 3e^{-\epsilon}$ [Berg

and von Hippel, 1987]. Hence,

$$I_{pos} = 2 + \sum_i p_i \log_2 p_i \tag{2.40}$$

$$= 2 - \frac{1}{Z}\log_2 Z + 3\frac{1}{Z\ln 2}\epsilon e^{-\epsilon} - 3\frac{e^{-\epsilon}}{Z}\log_2 Z \tag{2.41}$$

$$= 2 - \log_2 Z + 3\frac{1}{Z\ln 2}\epsilon e^{-\epsilon}. \tag{2.42}$$

The mismatch energy ϵ can be obtained from the above expression, and from ϵ and L, we obtain $S(\epsilon, L) = (\frac{1}{4} + \frac{3}{4}e^{-\epsilon})^L$.

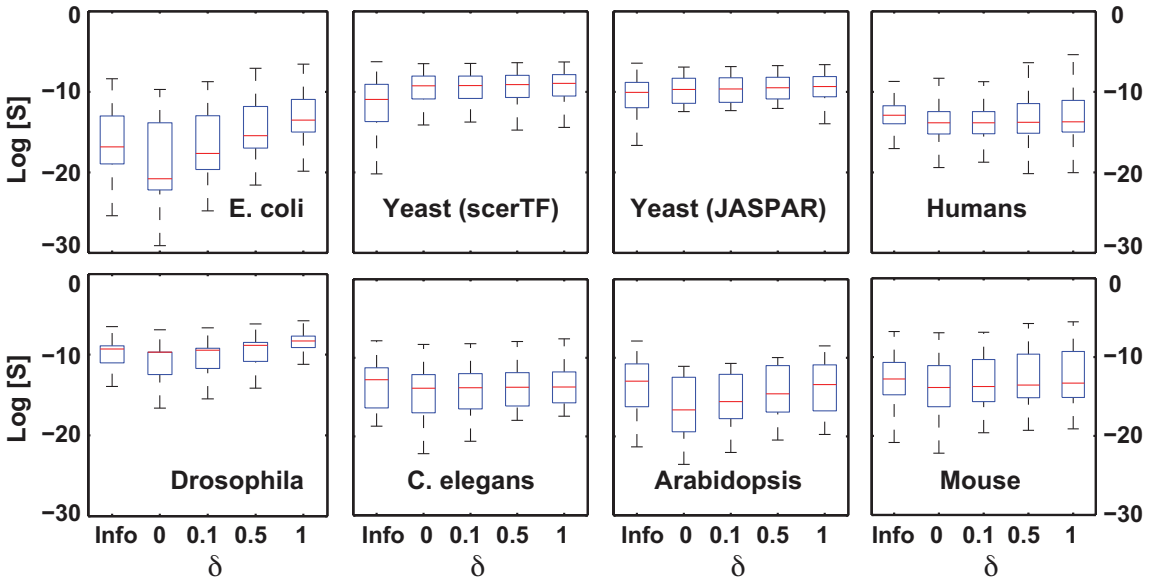


Figure 2.16: **Distributions of S for TFs from different databases.** In each panel, organism-specific (from a single database) boxplots of S are shown. The first boxplot in each panel corresponds to S values obtained from information estimates, and the remaining four correspond to S values obtained using the psuedo-count method with $\delta = 0, 0.1, 0.5, 1$ from left to right. *E. coli* TFs were obtained from RegulonDB [Gama-Castro *et al.*, 2011] and yeast (*S. cerevisiae*) from two different databases - scerTF [Spivak and Stormo, 2012] and JASPAR [Mathelier *et al.*, 2013]. All the other organism specific TFs were obtained from JASPAR. Notice that in the pseudo-count method, δ has the biggest influence on the estimates in *E. coli*. Importantly, for all other organisms, the estimates are invariant to δ and in general seem to agree with the information estimate.

Pseudo-count method

In this method, we infer ϵ for all three non-cognate nucleotides in each position, and obtain ϵ for the TF as an average of these $3L$ values. For an arbitrary position j, as before, assume that $i = 4$ has the maximum counts ($b_{4j} > b_{ij}$, $i = 1, 2, 3$). We obtain $\epsilon_{ij} = \log \frac{b_{4j}}{b_{ij}}$ and mismatch penalty for position j as $\epsilon_j = \frac{1}{3}(\epsilon_{1j} + \epsilon_{2j} + \epsilon_{3j})$. If some entry $b_{kj} = 0$, ϵ_{kj} is undefined. To take care of this, we first add a pseudocount δ to all

Parameter	Explanation
ϵ	Energetic mismatch penalty
L	Binding site/TF motif length
S	Binding site similarity
B	PCM of a TF, matrix of size $4 \times L$
I	Total information in the TF motif
p_{ij}	Frequency of nucleotide i in position j
q_{ij}	Expected frequency of nucleotide i in position j
ϵ_{ij}	Energetic contribution of nucleotide i at position j
δ	Pseudo-count added to PCM B
B_δ	Pseudo-count added PCM

Table 2.2: **Explanation of parameters involved in similarity estimation.**

entries of B and obtain a modified PCM B_δ to infer ϵ. The value of δ chosen is arbitrary and it is common practice to use $\delta = 0.5$ or $\delta = 1$. As before, to get rid of non-specific positions, we consider positions that have $\epsilon_j \geqslant 1$. This is similar to the previous exclusion criterion in the information method; requiring $\epsilon_j \geqslant 1$ is equivalent to requiring $I_j \geqslant 1.7$ bits. From the remaining, we take a mean to obtain $\epsilon = \frac{1}{L}\sum_j \epsilon_j$, and finally obtain $S(\epsilon, L) = (\frac{1}{4} + \frac{3}{4}e^{-\epsilon})^L$.

2.8 Combinatorial regulation (AND gate)

So far, we have been dealing with models in which each gene is regulated by a single type of TF, whether as a single activator or a repressor. Here, we will consider a simple model of combinatorial regulation and compute optimal crosstalk for this setup as a function of different parameters of interest.

As before, we have M genes in total, with each gene having two binding sites, corresponding to two different (cognate) TF types. For a particular gene to be ON, we need the presence of both the cognate TF types, which need to occupy both the binding sites. Transcription of the gene occurs only when both the binding sites are occupied. In the non-crosstalk setup, this corresponds to an implementation of an AND gate. We don't specify how this AND gate is implemented on the molecular level.

To bring combinatorial regulation into picture, we allow a particular TF to interact with many different TFs in regulating a set of genes. In the basic activation setup, the total number of genes M was equal to the total number of TFs. In the combinatorial regulation setup which is an extension of the basic activation setup, the total number of genes M will be equal to the total number of different TF-TF interactions that exist. This will depend on the extent of combinatorial regulation which we quantify using f, the fraction of TF-TF interactions each TF type realizes.

2.8.1 Total number of TFs T

If there are T TFs in total, each TF can potentially interact with $N_{int} = f(T-1)$ other TF types, where f is the fraction of total interactions each TF type realizes. This gives us $M = TN_{int}/2$, which gives us $T \approx \sqrt{2M/f}$ and $N_{int} \approx \sqrt{2Mf}$. But each TF should interact with at least one other TF, so we require $N_{int} \geqslant 1$. Taking both of these into account, we have, for N_{int}, the number of TFs each TF interacts with, and the number of total TFs T,

$$N_{int} = \max(1, \sqrt{2Mf}), \tag{2.43}$$

$$T = \frac{2M}{N_{int}}. \tag{2.44}$$

If each TF interacts wtih all other TFs, we have $f = 1$ and $N_{int} = T - 1$, which give us $T \approx \sqrt{2M}$. This we call "perfect combinatorial regulation" because it minimizes the number of TFs needed to express a certain number of genes. If each TF realizes only a fraction $1/2M < f < 1$ of its interactions, we have $N_{int} > 1$ interactions for each TF, which gives us $T \approx \sqrt{2M/f}$. This we call "imperfect combinatorial regulation". If $f \leqslant 1/2M$, we have $N_{int} = 1$, which gives us $T = 2M$. This we call "worst combinatorial regulation".

2.8.2 Number of TFs for Q genes to be ON

As before, we will compute the optimal crosstalk when Q genes are required to be ON. Here, we compute the "typical" number of TFs t by following a similar recipe as before. We have $Q = tn_{int}/2$, where n_{int} is the number of interactions per TF now. This will be smaller as there are fewer TFs ($t \leqslant T$). As before, we have

$$n_{int} = \max(1, \sqrt{2Qf}), \tag{2.45}$$

$$t = \frac{2Q}{n_{int}}. \tag{2.46}$$

When $f > 1/2Q$, we have $t = \sqrt{2Q/f}$ and when $f \leqslant 1/2Q$, we have $t = 2Q$.

2.8.3 Number of genes with only one TF present and none present

Unlike in the basic activation setup, in the combinatorial regulation, when Q genes are required to be ON, apart from genes that do not have any cognate TFs, we have genes that have only one out of the two cognate TFs present. In such a situation, as calculated above, we have t TFs and each TF has n_{int} interactions, while the total number of interactions it can have are N_{int}. So each TF that is present has $N_{int} - n_{int}$ missing

interactions. So the number of genes that have only one TF present can be obtained as

$$Q_1 = \frac{t(N_{int} - n_{int})}{2}. \tag{2.47}$$

The number of genes with no TFs present now is $Q_0 = M - Q - Q_1$. In Table 2.3, we have listed all possible configurations for the two binding sites of a gene, along with details of crosstalk states and statistical weights. From this, we get the per-gene crosstalk for different types of genes. For genes that have both the cognate TFs present (Q out of M), the per-gene crosstalk error is

$$x_{both} = 1 - \frac{(C/t)^2}{\begin{aligned}(C/t)^2 + 2e^{-E_a}(C/t) + 2(C/t)CS + 2e^{-E_a}CS \\ + (CS)^2 + (C/t)CS(2\epsilon, L) + e^{-2E_a}\end{aligned}}. \tag{2.48}$$

For genes that have only of the two cognate TFs present (Q_1 out of M genes), the per-gene crosstalk error is

$$x_{one} = \frac{(C/t)CS + (CS)^2 + (C/t)CS(2\epsilon, L)}{e^{-E_a}(C/t) + (C/t)CS + 2e^{-E_a}CS + (CS)^2 + (C/t)CS(2\epsilon, L) + e^{-2E_a}}. \tag{2.49}$$

For genes that do not have any of their two cognate TFs present ($M - Q - Q_1$ out of M genes), the per-gene crosstalk error is

$$x_{none} = \frac{(CS)^2 + (C/t)CS(2\epsilon, L)}{2e^{-E_a}CS + (CS)^2 + (C/t)CS(2\epsilon, L) + e^{-2E_a}}. \tag{2.50}$$

Hence the total crosstalk is

$$X = \frac{Q}{M}x_{both} + \frac{Q_1}{M}x_{one} + 1 - \frac{Q + Q_1}{M}x_{none}. \tag{2.51}$$

At a given M and f, for each (Q, S) pair, we compute the optimal concentration C^* numerically, and obtain the minimal crosstalk X^*_{comb}.

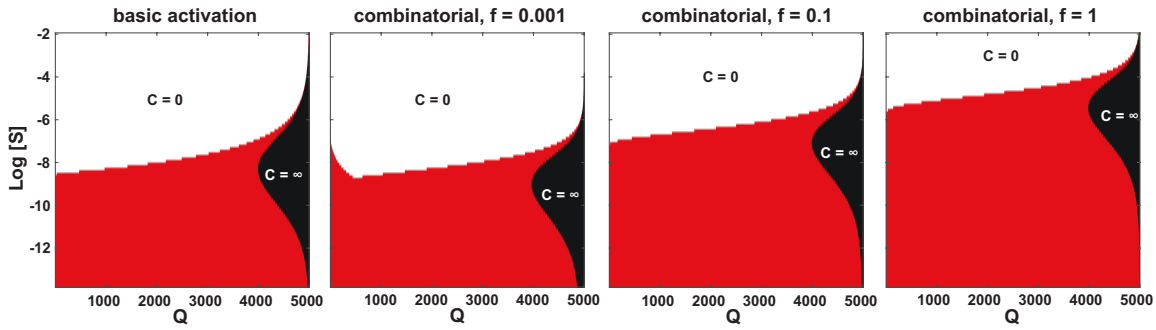


Figure 2.17: **Different regimes on the (Q, S) plane for the basic and combinatorial setup.** Here we show how the boundaries between different regulatory regimes shift between the basic activation setup and combinatorial regulation setups. In the leftmost panel, we have the regimes for the basic activation setup. In all the other panels, we have the regimes for the combinatorial setup for $f = 0.001, 0.1$ and 1 respectively from left to right. For $f = 0.001$, the "regulation regime" is slightly smaller than in the basic activation setup. As f increases, the "regulation regime" increases in size (and is bigger than in the basic activation setup) and the boundary with $C = 0$ is pushed higher towards bigger similarity.

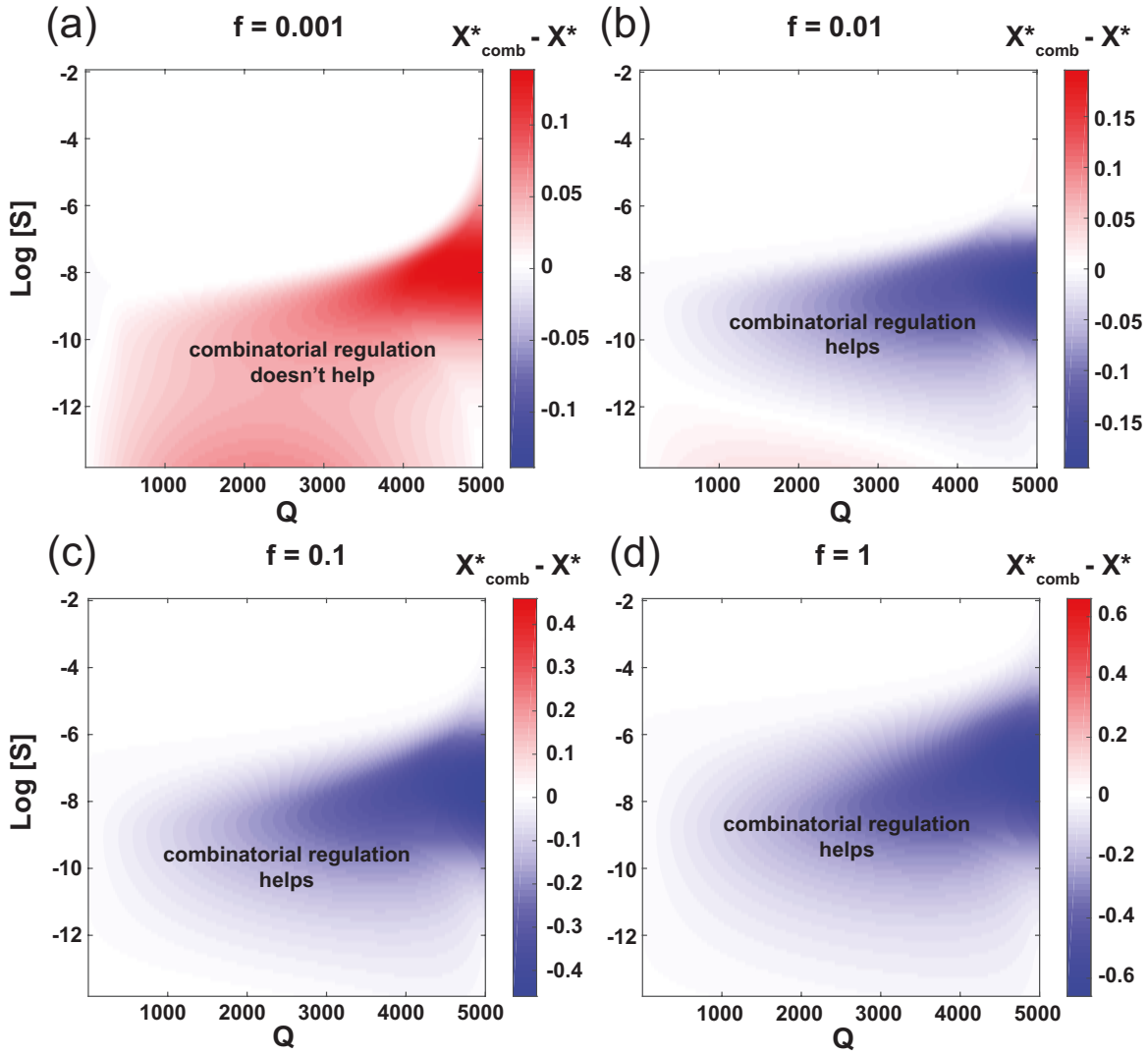


Figure 2.18: **Difference in optimal crosstalk between combinatorial setup and the basic activation setup for different f.** We show the difference in optimal crosstalk between combinatorial setup and basic activation setup for different f. **(a)** $f = 0.001$; here, combinatorial regulation performs worse in comparison to the basic activation setup. **(b,c,d)** $f = 0.01, 0.1, 1$ respectively; here, combinatorial regulation is almost always helpful and gives a significant improvement over basic activation in terms of optimal crosstalk. At the baseline parameters of $Q = 2500, M = 5000$ and $\log(S) = -10.5$, optimal crosstalk for the combinatorial setups reads $X^*_{comb} = 0.28, 0.18, 0.11$ and 0.07 for $f = 0.001, 0.01, 0.1$ and 1 respectively, compared to $X^* = 0.23$ for the basic activation setup.

As plotted in Fig. 2.17, the boundaries between different regimes shift in the combinatorial setup. In particular, while at small f, the "regulation regime" shrinks in the (Q, S) plane, as f increases, it expands. As f increases towards 1, the boundary between the "regulation regime" and "$C = 0$" regime moves towards larger S. In Fig. 2.18, we have plotted the difference in optimal crosstalk between combinatorial regulation and the basic activation setup. For $f = 0.001$, combinatorial regulation doesn't improve from

the basic activation setup in terms of optimal crosstalk. But for $f = 0.01, 0.1, 1$, combinatorial regulation gives a lower optimal crosstalk than the basic activation setup. So, there exists a threshold in f such that for combinatorial regulation below that threshold, the "regulation regime" shrinks in comparison to the basic activation setup and performs worse. Above the threshold, the "regulation regime" expands towards larger S and gives a lower optimal crosstalk than the basic activation setup. At the baseline parameters of $Q = 2500, M = 5000$ and $\log(S) = -10.5$, optimal crosstalk for the combinatorial setups reads as $X^*_{comb} = 0.28, 0.18, 0.11$ and 0.07 for $f = 0.001, 0.01, 0.1$ and 1 respectively, compared to $X^* = 0.23$ for the basic activation setup.

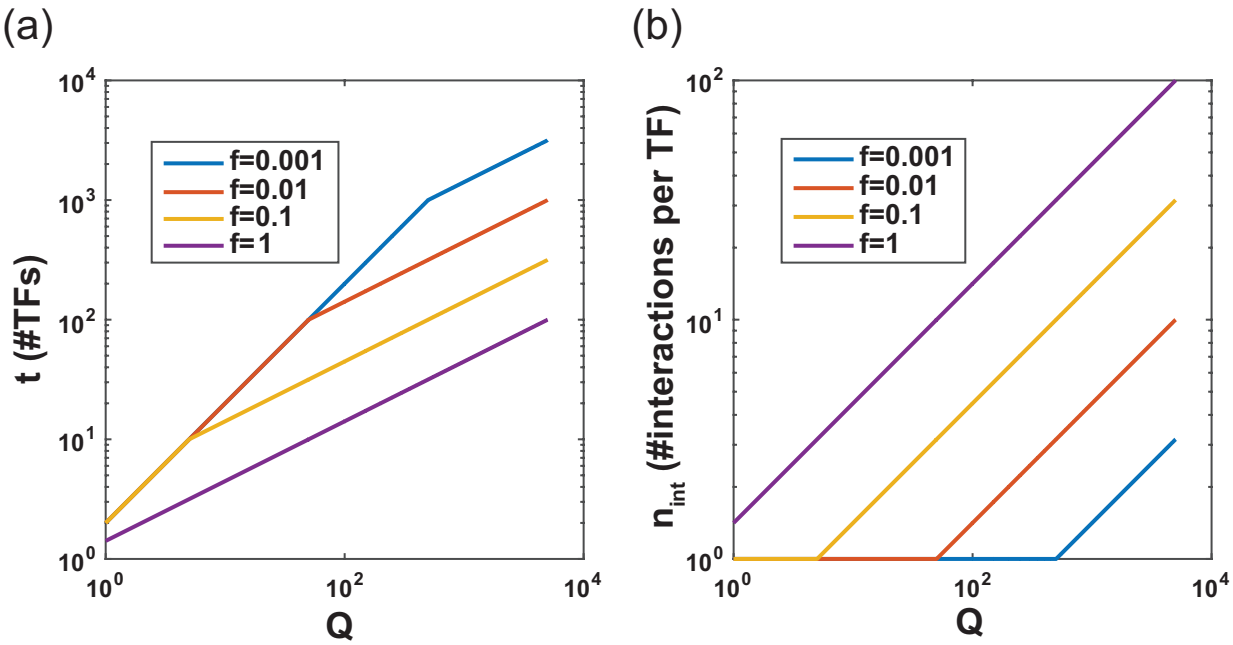


Figure 2.19: **Number of TFs present - t, and number of interactions per TF - n, against Q for different f, on log-log scale.** Here we show **(a)** how number of TFs t and **(b)** number of interactions per TF n vary with Q for different f. For each f, for Q smaller than some threshold value which depends on f, the number of TFs t varies as $Q = 2t$ and the number of interactions per TF n is constant at 1. For all Q greater than this threshold value, $\log n$ increases linearly with $\log Q$ (n changes with Q in a power-law fashion).

This decrease in crosstalk is consistent simply with the reduction in the number of regulatory components (T and t, the number of TFs, see Fig. 2.19), as discussed in the next section. In the case of perfect combinatorial regulation $f = 1$, we have about $\sqrt{2M}$ instead of M TF species in the basic activation setup, which is a significant reduction in the number of regulatory components. Hence, each TF now effectively controls $\theta = M/\sqrt{2M} = \sqrt{M/2}$ genes, and so the decrease in crosstalk is expected to be roughly $\sqrt{\theta}$ compared to the basic activation setup as argued below. The actual reduction in crosstalk is not as large because of certain differences between the combinatorial setup and θ-genes setup of the next section. One major difference is that in the θ-genes setup, the cell can only activate sets of genes of size θ, while in the combinatorial setup, the cell

has the power to activate single genes at will, albeit at the cost of partially activating genes that are not needed. Fundamentally, therefore, crosstalk decrease gains come from the decrease in the number of regulatory components in the system, which again points to the explosion in the number of possible noncognate interactions as the crucial origin of the crosstalk.

Although this combinatorial strategy allows crosstalk reduction and has been documented at specific promoters, we point out that the predicted, square-root scaling of the number of TF species with the total number of genes, M, is inconsistent with published reports [van Nimwegen, 2004; Maslov *et al.*, 2009], which is in fact quadratic (!) making it unlikely that crosstalk reduction is achieved through genome-scale combinatorial control as analysed here.

	configuration (XY)	activity	crosstalk if gene needs to be ON	OFF, C can be			Energy	Weight
				X	Y	none		
1	CC	ON	-				0	$(C/t)^2$
2	UC	OFF	+		-		E_a	$e^{-E_a}(C/t)$
3	NC	ON	+		+		ϵd	$(C/t)CS$
4	CU	OFF	+	-			E_a	$e^{-E_a}(C/t)$
5	CN	ON	+	+			ϵd	$(C/t)CS$
6	UU	OFF	+	-	-	-	$2E_a$	e^{-2E_a}
7	UN	OFF	+	-	-	-	$E_a + \epsilon d$	$e^{-E_a}CS$
8	NU	OFF	+	-	-	-	$E_a + \epsilon d$	$e^{-E_a}CS$
9	N_xN_y	ON	+	+	+	+	$\epsilon(d_1 + d_2)$	$(CS)^2$
10	N_xN_x	ON	+	+	+	+	$2\epsilon d$	$(C/t)CS(2\epsilon, L)$

Table 2.3: Caption next page.

Table 2.3: (Previous page.) **All possible binding configurations and the corresponding energies for a combinatorial regulation setup implementing an AND gate.** Each gene has two binding sites which bind two different cognate TF types. The "configuration" column lists all the configurations of the two binding sites of a gene. "C" denotes binding by cognate factor, "N" - binding by non-cognate and "U" - means that the site is unbound. We distinguish between binding of non-cognate molecules of the same type (N_xN_x) and different types (N_xN_y). The "activity" column denotes whether in the given configuration the gene is either ON or OFF. To implement the AND gate, we assume that transcription occurs (ON) only when both the binding sites are bound. The next four columns denote whether this configuration is counted as crosstalk (+) or not (-). In the leftmost column "ON", both the cognate transcription factors are present (and the gene should be ON). In the next three "OFF" columns, at least one of the cognate TFs is absent (and the gene should be OFF). In "C can be X" column, the cognate TF of only the left binding site (X) is present, in "C can be Y", the cognate TF of only the right binding site is present, and in "C can be none" column, both the cognate TFs are absent. Blank space denotes a non-existing configuration: these are the configurations including a cognate factor bound in the situation that it is absent. The column "Energy" specifies the energy of these configurations. We define the reference energetic level $E = 0$ as the state "CC" when both sites are bound by their cognate factors, such that all other energies are positive. The column "Weight" denotes the statistical weight of the configurations, taking into account the concentrations of the relevant TFs and the energy of the configurations. Note that the statistical weight of the last binding configuration N_xN_x uses $S(2\epsilon, L)$ instead of the otherwise $S(\epsilon, L)$.

2.9 Every transcription factor regulates Θ genes

When every gene has its own unique TF type, this allows for maximal flexibility in regulating each gene individually. But real gene regulatory networks inevitably have fewer TFs than the number of target genes, so that at least some transcription factors regulate several genes.

To accommodate this picture, here we consider a simple extension of the basic model, in which each TF regulates Θ genes rather than one. We assume no overlap between the sets of genes regulated by various TFs, so that the total number of TFs species is now Θ times smaller than before. If Q genes should be active, then Q/Θ TF species should be present in a given condition. Assuming that $Q/\Theta \gg 1$, we can approximate $Q/\Theta - 1 \approx Q/\Theta$ as before. The only change from the basic crosstalk formulation is in x_1,

because the concentration of cognate factors is now Θ times larger than before:

$$x_1^{\Theta} = \frac{e^{-E_a} + CS}{\frac{C}{Q/\Theta} + e^{-E_a} + CS} \tag{2.52a}$$

$$x_2^{\Theta} = \frac{CS}{e^{-E_a} + CS}. \tag{2.52b}$$

This formulation is analytically solvable, yielding

$$X_\Theta^* = \frac{Q}{\Theta M}\left(-S(M-Q) + 2\sqrt{S\Theta(M-Q)}\right) \tag{2.53a}$$

$$x_1^{\Theta *} = \frac{\sqrt{S(M-Q)}}{\sqrt{\Theta}} \tag{2.53b}$$

$$x_2^{\Theta *} = \frac{SQ}{\Theta}\left(\frac{\sqrt{\Theta}}{\sqrt{S(M-Q)}} - 1\right) \tag{2.53c}$$

$$C_\Theta^* = \frac{e^{-E_a}Q(\Theta - S(M-Q))}{S^2(M-Q)Q + S(M-2Q)\Theta + \sqrt{S(M-Q)}\Theta^{3/2}}. \tag{2.53d}$$

For small S the leading term in the optimal concentration is

$$C_\Theta^* = \frac{1}{\sqrt{\Theta}}\frac{e^{-E_a}Q}{\sqrt{S(M-Q)}} + O(1). \tag{2.54}$$

Compared to the basic model result of Eq. (2.11), the optimal TF concentration is now reduced by factor of $\sqrt{\Theta}$, as is the minimal crosstalk error of the first type, $x_1^{\Theta *}$. The dependence on Θ of the crosstalk of the second type, x_2^{Θ}, is more complicated. These gains in crosstalk have, however, been achieved by sacrificing the ability to regulate each gene individually: now, the smallest set of genes that can be co-activated is of size Θ. Typically, TFs might constitute $\gtrsim 10\%$ of the genes [van Nimwegen, 2004]; with $\Theta \sim 10$, the crosstalk could be reduced by a factor of ~ 3 at best.

2.10 Discussion

Molecular recognition events, which are essential to the functioning of the cell and the organism, are ultimately limited in their specificity by the finite specificity of their underlying monomer interactions. For instance, in the case of transcriptional regulation, the specificity limits of the hydrogen bond mediated interactions between amino acids of the TF's DBD and the nucleotide base pairs on the DNA, set the limits with which a TF molecule can bind to the correct target site (based on the latter's sequence) while avoiding spurious binding to off-target sites that might trigger unwanted cellular programs. Such consequences, which might be commonplace inside the cell because transcription takes place in a mix of cognate and a large number of non-cognate TF species, might be severe

to the cell. But studies so far have not considered this issue to their completion and quantitatively analyzed what role such off-target interactions might play in the functioning of the cell. In this chapter, we constructed a theoretical framework to study such crosstalk, accounting for all possible cross-interactions between TFs and binding sites. As crosstalk is a systemic phenomenon, we constructed our model as such, enabling us to compute a lower bound on crosstalk (with respect to TF concentrations) and thereby estimate what level of cross-interactions cannot be overcome by the cell, given a particular class of molecular mechanisms behind transcriptional regulation. This lets us not only assess the effectiveness of various regulatory strategies in decreasing crosstalk, but also enable us to derive limits that crosstalk places on gene regulatory system design.

We show that crosstalk depends primarily on the total number of genes M, the typical number of co-activated genes, Q, and the average level of similarity between pairs of binding sites, S, and that these parameters robustly define three possible regulatory regimes. An important regime is the "regulation regime", in which a non-zero and finite TF concentration that minimizes crosstalk exists with binding sites sufficiently distinguishable from each other (S not too large) and the typical number of co-activated genes not extreme (Q not too high). The other two regimes are anomalous cases where regulation is dysfunctional - either the optimal TF concentration that minimizes crosstalk is zero or infinity. A closer look at the boundaries between these regimes indicates that the average similarity between binding sites, S, puts an upper bound to the total number of genes that an organism can effectively regulate [Itzkovitz *et al.*, 2006].

Another paradigmatic example of molecular recognition is protein-protein interaction networks [Zhang *et al.*, 2008; Johnson and Hummer, 2011], studies on the evolution of which have applied a combination of positive and negative design using computer simulations, concluding that 'negative design' seriously constrains the possible architectures [Sear, 2004a; Sear, 2004b; Myers, 2008; Johnson and Hummer, 2011]. Analogous to the binding sites similarity, S, Johnson *et al.* [Johnson and Hummer, 2011] used the minimal energy gap between specific and nonspecific interactions as a quantitative measure for the likelihood of specific versus nonspecific interactions. They inferred that the energy gap follows a power-law with the total number of proteins in the network, and that it depends inversely on the size of the binding surface, L, analogous to our results. Further, they found the network designs that have hubs - proteins having multiple specific partners - have higher crosstalk compared with networks that inhabit only pairwise interactions. This is different between protein-protein interaction networks as all nodes interact with each other in them, and TF-DNA interactions, as BSs do not interact with themselves. Another study by Zhang *et al.* [Zhang *et al.*, 2008] concluded that a trade-off exists between proteome diversity and concentrations and that the empirical values are close to the limits set by crosstalk.

Much like TF concentrations in our model, protein concentrations face a trade-off that they should be high enough to form specific interactions, but not so high as to form many nonspecific ones. This explosion of the number of non-cognate configurations comes up in other molecular contexts like prebiotic metabolism [Schuster, 2000] and the immune

system [Košmrlj *et al.*, 2008], where receptors are designed by selection to recognize foreign peptides while avoiding self-binding. In the context of TF-DNA interactions Sengupta *et al.* [Sengupta *et al.*, 2002] studied how mutation and selection counteract each other to tune TF specificities. They identified a trade-off between avoiding the loss of current targets (for which a lower specificity is favoured) and avoiding the spurious recruitment of new ones (for which a higher specificity is favoured); they also report an inverse relation between the number of different targets and the TF specificity for each. In this light, an intriguing direction for future research is to explore how crosstalk might limit the complexity of regulatory networks in an evolutionary setting, a first step in which we take in Chapter 4.

In the parameters space of M, Q and S, where are real organisms placed? Prokaryotes usually have longer binding sites and fewer genes than eukaryotes, and hence crosstalk is low between 1 and 10%. Crosstalk is high in eukaryotes, which have significantly more genes and shorter binding site. Even for a short genome of $M = 5,000$ genes, such as yeast, or for longer genomes of metazoans where most of the genes have been non-transcriptionally silenced, we expect minimal crosstalk of $X \approx 0.23$, almost a quarter of genes in erroneous states at any point. But the equilibrium thermodynamics based biophysical model of transcription considered here perhaps might not be completely relevant for eukaryotic systems, and the activation of transcription might require binding events of multiple different types of TFs that involve kinetic proofreading as well.

Traditional knowledge suggests that complex regulatory schemes increase the specificity of gene regulation by cognate factors and hence completely mitigate the problem of crosstalk. In contrast, by considering mechanisms that involve combinatorial regulation by multiple TFs, we reveal a more intricate picture. We showed that [Friedlander *et al.*, 2016] cooperativity, and a combined use of activators and repressors, do not eliminate spurious interactions for the reason that by adding new regulatory components, the number of non-cognate interactions also drastically increases, therefore making crosstalk often worse. In this chapter, I showed how a simple AND-gate type combinatorial regulation that requires co-binding of TFs of various types helps overcome crosstalk by decreasing the number of regulators necessary to regulate a fixed number of genes. But further work is needed to fully elucidate crosstalk limits in more general models of combinatorial control and cooperativity, with interesting parallels to precision in biochemical sensing, in equilibrium as well as out-of-equilibrium scenarios [Govern and ten Wolde, 2014; Mora, 2015; Skoge *et al.*, 2013; Cepeda-Humerez *et al.*, 2015].

Taken together, our work on crosstalk suggests that this global constraint poses significant challenges in eukaryotic regulation that can be mitigated, but not easily removed. Although the initial conclusion was based on the simplest model of gene regulation, even an analysis of more complex regulatory strategies revealed that crosstalk remains a challenge. One reason for this is because a major determinant of crosstalk is the binding site similarity S, denoting how similar different binding site sequences are, which primarily depends on the typical mismatch energy ϵ and the length of the binding sites, L. Although crosstalk could be reduced by extending binding site length and/or augmenting the bind-

ing energy, both parameters are severely constrained by a combination of biophysical and evolutionary factors. The scale of the mismatch energy is set by the energetics of hydrogen bonds to $\sim 2-4k_BT$, whereas the length of individual binding sites in eukaryotes appears strongly constrained by evolutionary considerations to $\sim$ 10 bp [Sengupta *et al.*, 2002; Stewart *et al.*, 2012; Tuğrul *et al.*, 2015]. Furthermore, the performance of complex regulatory schemes is also limited by the explosion of possible non-cognate configurations that may lead to erroneous regulation, hence only those that significantly reduce the number of regulatory components mitigate the problem of crosstalk. These constraints apply universally: any regulatory scheme operating at equilibrium, no matter how complex, faces a fundamental limit to its achievable error, for reasons that led Hopfield [Hopfield, 1974] to propose kinetic proofreading.

The primary conclusion from this chapter is that crosstalk in gene regulation is an important constraint, and is far from being solved. Equilibrium mechanisms of gene regulation, unless they significantly reduce the number of regulatory components cannot overcome crosstalk and face fundamental limits from its emergence. In this light, it is conceivable that cells might have evolved out-of-equilibrium solutions where energy is deliberately spent to counteract the detrimental effects of crosstalk. For instance, permanent gene silencing by spending energy to compactify DNA [Allshire and Madhani, 2018; Wang *et al.*, 2016], or localization of transcriptional activity to specific cellular compartments via phase-separation like mechanisms [Hilbert *et al.*, 2018] that spend energy to switch between dynamic states, or molecular reaction schemes for gene regulation that implement variants of kinetic proofreading [Cepeda-Humerez *et al.*, 2015].

Chapter 3

General theoretical formulation of TF-BS coevolution

In the preceding chapters, I described questions concerning the biophysical setup of gene regulation that relies on transcription factor-DNA binding. While in Chapter 1, I introduced a thermodynamic equilibrium model of TF-DNA binding using the grand-canonical ensemble, and described how simple constraints based on such biophysical considerations shape the sequence space of DNA sequence involved in regulation, in Chapter 2, I showed how to quantify and investigate crosstalk, a global biophysical phenomenon that corresponds to incorrect cross-interactions between TFs and BSs, using the biophysical model. Another set of important questions concern the evolution of gene regulatory systems; as gene regulatory networks are ultimately embedded in the DNA sequence – various functional domains TF coding genes, binding site sequences on the regulatory elements – they can be completely understood only in the context of how the biophysical relationship between DNA sequence and regulatory function interacts with functional constraints that dictate fitness and other evolutionary considerations, ultimately shaping the evolution of DNA sequence that represents these GRNs.

In the chapter, I will introduce a general setup that embeds the biophysical model of gene regulation (Chapter 1) inside a population genetic framework of evolution. By combining the biophysical model with a fully specified evolutionary model comprising important evolutionary parameters like population size, fitness, mutation rates etc., we can ask evolutionary questions on the structural and functional design of gene regulatory networks with an arbitrary number of interacting components. For instance, in Chapter 4, I will develop, as a special case of the general setup described in this chapter, a model to describe the evolution of a gene regulatory network of two TFs under different signal sensing constraints. We will answer questions about the various pathways and the

evolutionary timescales involved in the specialization of duplicate TFs, something that cannot be achieved without merging the biophysical model with a population genetics framework. Also, such a generic framework will serve as a platform to construct models that can be used in future research on the evolution of gene regulatory systems.

3.1 General setup

We have Q TF types and M binding sites, each of which is associated with a particular gene. As described in Chapter 1, we describe each TF type by its consensus sequence, which is the BS sequence it prefers to bind to, and the mismatch penalty ϵ which corresponds to the increase in TF-BS binding energy with every mismatch between the TF consensus sequence and the BS sequence. Each TF i is present in concentration c_i, which in turn could depend on upstream signals that the TFs sense using signal sensing domains (see 4 for more details). The phenotype of the gene regulatory network is the set of expression levels of all M genes, which we quantify by considering the equilibrium probability of the corresponding binding sites being bound. These probabilities depend on the $Q \times M$ mismatches between the TF consensus sequences and the BS sequences, and also the concentrations of the TFs.

3.2 Environment

The set of concentrations of all TFs, $C_t = (c_1, c_2, \ldots, c_Q)$ defines an environmental state, where t indexes either time, space, or some other external condition. The environmental state can be modeled in two ways:

Set of possible states. The total set E of possible environmental states, $C_t \in E$ for all $t = 1, 2, \ldots$, can be listed out explicitly. In each state C_t, the concentrations c_i are deterministic. When t indexes time, the sequence $\{C_t\}$ might, for instance, correspond to a particular temporal sequence of TF concentrations, like in developmental programs. When t indexes space, this might correspond to fixed TF concentration patterns that arise in different tissues. Here, one can just explicitly list the sequence of environments $\{C_t\}$ or consider the distribution P_E over possible environments E, $C_t \sim P_E(.)$.

Stochastic treatment. The concentrations c_i can be treated as continous random variables. For instance, one can assume that C_t comes from a multivariate Gaussian distribution. $C_t \sim \mathcal{N}_Q(\langle C \rangle, \Sigma_c)$. Such a model can be considered to represent pattern of TF concentrations with just one "cell state". Different cell states, which are to be understood as clearly distinct environmental contexts (for example, presence of different sets of nutrients) with different means $\langle C \rangle$, can be modeled by considering a multivariate Gaussian mixture distribution, $C_t \sim \sum_i \beta_i \mathcal{N}_Q(\langle C \rangle_i, \Sigma_i)$ with $\sum_i \beta_i = 1$, where i indexes over the possible cell states and β_i represents the frequency of that cell state. When t indexes time, we sample TF concentration patterns in the same cell over different times.

When t indexes space, we are sampling cells from different tissues (each tissue has a different cell state) with β_i reflecting the "size" of different tissues. Note that listing out possible states is a special case of the stochastic treatment in which all the variances are zero, E is the set of all possible cell states $\{\langle C\rangle_i\}$ and $\beta_i = P_E(\langle C\rangle_i)$.

3.3 Phenotype

In a given environmental state C_t, the expression levels of each gene j depend on the equilibrium probability of its binding site being occupied (occupancy probability) by some TF molecule, $p_b(j,t) = \sum_{i=1}^{Q} p_b^{(i)}(j,t)$, where $p_b^{(i)}(j,t)$ is the probability that a TF molecule of type i is bound. These occupancy probabilities comprise the phenotype of the GRN. For each TF i and binding site j pair, k_{ij} denotes the number of mismatches between the consensus sequence of TF i and the sequence at binding site j. For each binding site j and TF x, $p_b^{(x)}(j,t)$ depends, apart from the mismatches k_{ij} for all TFs $i \in \{1,2,\ldots,Q\}$ and the mismatch penalties $\{\epsilon_i\}$, on the chemical potentials $\{\mu_i(t)\}$, which depend on concentrations C_t and other sequestration factors (see Chapter 1), as

$$p_b^{(x)}(j,t) = \frac{e^{\mu_x(t)} e^{-\epsilon_x k_{xj}}}{1 + \sum_{i=1}^{Q} e^{\mu_i(t)} e^{-\epsilon_i k_{ij}}}. \tag{3.1}$$

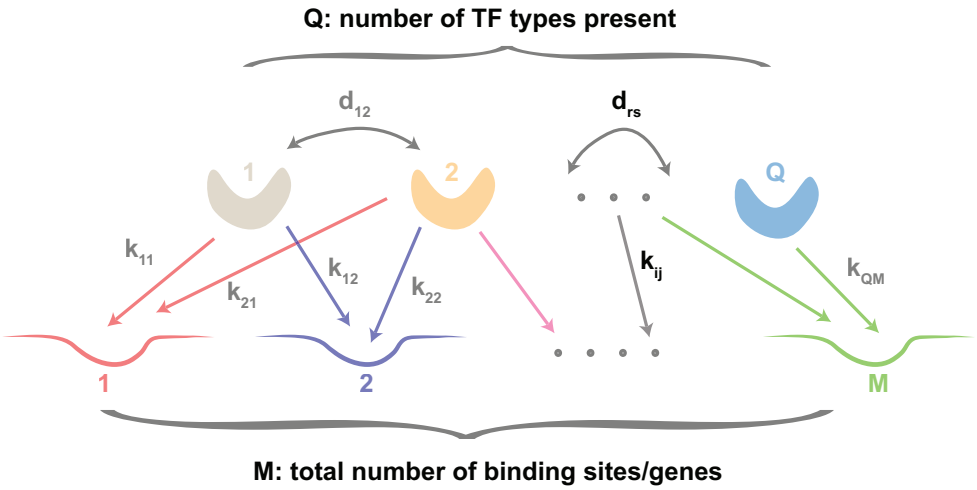


Figure 3.1: **Schematic of the biophysical model.** We have Q TF types, each of which can bind to M binding sites with different affinities. These affinities depend on the genotype which is specified by the mismatches $\{k_{ij}\}$ between the i^{th} TF's consensus sequence and j^{th} binding site sequence, and the mismatches $\{d_{rs}\}$ between the consensus sequences of r^{th} and s^{th} TFs.

3.4 Genotype

The primary actors in the GRN model are the TFs, which can sense upstream signals and bind to binding sites, resulting in regulated gene expression patterns. In principle, the genotype is the set of coding sequences of the TFs and the BS sequences corresponding to the binding sites of the various target genes. However, as we are interested in TF-DNA binding as the phenotype, and as empirical data characterizing the binding preferences of TFs is readily available, we consider the genotype to be comprised of (a) the binding site sequences on the regulatory regions of the target genes, and (b) the position weight matrices (PWMs) of various TFs, characterizing their binding preferences to different BS sequences, and possibly. The PWMs contain information about the TF consensus sequences, which are BS sequences that the TFs prefer to bind, and the relative penalty on binding energy as BS sequences deviate from the consensus sequence. Any other domains of interest in the TF, like the signal sensing domain, can also be included in this genotypic description.

This genotypic space characterized by sequences is very large, bringing practical limitations of computation into consideration. However, as the occupancy probabilities of binding sites, which is our phenotype of interest, depend only on the mismatches between the TF consensus sequences and the BS sequences (constant mismatch penalty model described in Chapter 1 and above), we will instead work in the "mismatch" space. This space of mismatches, which we call the "reduced genotype space" is defined by $(\{k_{ij}\}, \{d_{rs}\}, \{\epsilon_i\}, \{L_i\})$ for $i, r, s = 1, 2, \ldots, Q$ and $j = 1, 2, \ldots, M$, where L_i is the length of sequences bound by TF i, ϵ_i is the mismatch penalty of TF i, k_{ij} is the number of mismatches between i^{th} TF's consensus sequence the BS sequence j, and d_{rs} is the number of mismatches between the consensus sequences of TFs r and s. While this reduced space of genotypes is more convenient computationally as it drastically reduces the number of dimensions, it introduces feasibility constraints on the set of possible mismatches. Not all combinations of $\{k_{ij}\}$ and $\{d_{rs}\}$ are feasible, and some combinations occurring more often than others; depending on the values of mismatches, the number of underlying genotypic sequences that result in them can vary from 0 to very large.

3.5 Fitness

Given a sequence of environmental states $\{C_t\}$ for $t = 0, 1, 2, \ldots$, at each environmental state, we define optimal occupancy probabilities $p_b^0(j, t)$ for every binding site j. For instance, $p_b^0(j, t)$ could be defined to be 0 for genes which need to OFF, and 1 for genes which need to be ON in different environments, a template we use in Chapter 4. Hence, for a given genotype, we obtain the phenotype of occupancy probabilities using the biophysical model of TF-DNA binding, and by comparing how close these are with optimal occupancy probabilities defined for various environments, we obtain the genotype's fitness.

Directional selection: Optimal occupancy probabilities $p_j^0(t) \in \{0, 1\}$, meaning binding sites are required to be either empty or saturated. Let $S = \{1, 2, \ldots, M\}$ be the set of all binding sites and $S_1(t)$ be the subset of binding sites which are expected to be saturated. So, $p_j^0(t) = 1$ if $j \in S_1(t)$ and $p_j^0(t) = 0$ if $j \notin S_1(t)$. In this setup, we define the fitness of a genotype $x = (\{k_{ij}\}, \{d_{rs}\})$ for C_t as

$$F_t(x) = \prod\nolimits_j (1 + s_j(t)p_j(t)) \tag{3.2}$$

$$= 1 + \sum_j s_j(t)p_j(t) + \mathcal{O}(s^2), \tag{3.3}$$

where $s_j(t)$ is the selection coefficient for binding site j. Note that $s_j(t) > 0$ if $j \in S_1(t)$ and $s_j(t) < 0$ if $j \notin S_1(t)$. In the limit of weak selection, the higher order terms can be neglected and fitness can be reduced to a linear function,

$$F_t(x) = 1 + \sum_j s_j(t)p_j(t). \tag{3.4}$$

Stabilizing selection: Optimal occupancy probabilities $p_j^0(t) \in [0, 1]$, meaning they can take intermediate values. In this setup, we define fitness of a genotype $x = (\{k_{ij}\}, \{d_{rs}\})$ for C_t as

$$F_t(x) = \prod\nolimits_j \Big(1 + s_j(t)(p_j(t) - p_j^0(t))^2\Big) \tag{3.5}$$

$$= 1 - \sum_{j \in S} s_j(t)(p_j(t) - p_j^0(t))^2 + \mathcal{O}(s^2), \tag{3.6}$$

where $s_j(t)$ is strength of selection for binding site j in environmental state C_t.

Now, given a set of environmental states C_t, there are two ways to define overall fitness. **Multiplicative fitness**: This is suitable for the case of precise sequential (either temporal or spacial) developmental processes. Multiplicative fitness gives us

$$F(x) \;=\; \Big(\prod_{t=1}^{T} F_t(x)\Big)^{1/T}, \tag{3.7}$$

where T is either the total number of time steps in the developmental program, or the number of tissues that involve in the regulation of the M binding sites. **Average (over environmental states) fitness**: This is suitable when the environmental state follows some statistics and we are interested in just the average behaviour over space or time. When C_t comes from a multivariate Gaussian mixture distribution, we have

$$F(x) = \sum_j \beta_j \idotsint_{c_1, c_2, \ldots, c_Q} f_C(\{c_i\}; \langle C \rangle_j, \Sigma_j) F_{\{c_i\}}(x), \tag{3.8}$$

where $f_C(\{c_i\}; \langle C \rangle, \Sigma)$ is the probability density function of the $Q-$dim multivariate normal distribution with mean $\langle C \rangle$ and covariance Σ. When the environmental states can be listed out as a set E, fitness can be reduced to

$$F(x) \;=\; \sum_{m \in E} \beta_m F_m(x), \tag{3.9}$$

where β_m is the probability of experiencing an environmental state m.

3.6 Evolutionary dynamics

We model the evolution of the population as a Markov chain, with the state space defined as the set X of all possible reduced genotypes $x = (\{k_{ij}\}, \{d_{rs}\})$. Here, we consider ϵ_i and L_i to be constant for all TFs and do not consider their evolution, but an extension of the model to include them is not hard. We think of each mismatch, either k_{ij} or d_{rs} as a dimension, which can take $L + 1$ possible values, L being the length of a binding site. We also define Y as the set of all mismatches $y = \{k_{ij}\}$ between TF consensus sequences and binding site sequences and Z as the set of mismatches $\{d_{rs}\}$ between TF consensus sequences. We have $X = Y \times Z$ and the number of dimensions are $MQ+Q(Q-1)/2$. The total size of the state space is $(L+1)^{MQ}(L+1)^{Q(Q-1)/2}$. Such a Markov chain treatment amounts to assuming the monomorphic fixed state limit in which the population is almost always composed of some single genotype x. An occasional mutation x' either leads to fixation, in which case the new state is x' or gets lost, in which case the population stays in x state. Fixation or loss of x' occurs before another mutation occurs. A mutation and its subsequent fixation move the population to a new state, and the transition rates R in the Markov chain are defined in terms of the mutation rate and fixation probability as

$$R_{x,x'} = NU_{x,x'}p_{fix}(x \to x'), \tag{3.10}$$

where N is the population size, $U_{x,x'}$ is the mutation rate from x to x', $p_{fix}(x \to x')$ is the fixation probability of a x' mutant in a population of x. We can ask questions related to the stationary distribution of the Markov chain, or the dynamics (using the transition rates) on the state space. Under certain conditions [Wright, 1931; Sella and Hirsh, 2005; Barton and Coe, 2009], the stationary distribution over states x is given by

$$P(x) \sim \Omega(x)e^{2N_eF(x)}, \tag{3.11}$$

where $\Omega(x)$ is proportional to the distribution in the neutral case, and N_e is the effective population size. There are two contributions to the stationary distribution $P(x)$ over states X - (a) the entropy term $\Omega(x)$, which specifies the number of microstates (sequences) that give that particular set of mismatches $x = \{k_{ij}\}$, and (b) the drift-selection term $e^{2N_eF(x)}$, which specifies the join action of drift and selection. After normalizing, the exact stationary distribution is given by

$$P(x) = \frac{\Omega(x)e^{2N_eF(x)}}{\sum_{x'}\Omega(x')e^{2N_eF(x')}}. \tag{3.12}$$

If $F(x) = g(x) + h$, where the constant h is independent of x, $P(x)$ doesn't depend on h.

3.7 Timescale separation

As a first approximation to the general problem of TF-binding site coevolution, we assume a timescale separation. Specifically, we assume that binding sites evolve on a faster timescale while TFs evolve on a slower timescale. The rate of evolution is affected by two factors - mutation rate and fixation probability. While mutation rate decides how much raw material is produced that is capable of evolutionary change, fixation probability tells us how capable the new raw material is of producing evolutionary change. That TFs evolve much slower than binding sites can thus be argued in two independent ways.

3.7.1 Lower TF mutation rate

TFs could have a lower effective mutation rate than binding sites. DNA binding domains of TFs have about $4-7$ amino acids that interact with binding sites, which corresponds to about $8-14$ nucleotides (in the gene coding for the TF) that are involved in the TF-DNA binding. This is comparable to the number of nucleotides in a single binding site but is much smaller compared to the size of a CRE (like enhancer) in which binding sites can evolve in various positions. So, while in a TF-single BS setup, the mutation rates are comparable, in a CRE-TF setup, the mutation rate for TFs is much lower than for binding sites.

3.7.2 Effect of fixation probability

Each mutation, either in the TF or the binding site, has a certain fixation probability that depends on its selective effect and population size. If most mutations in the nucleotides coding the DBD of TFs are deleterious, they have negligible rates of fixation and hence only a few mutations in TFs are expected to get fixed. It is not clear if this is the case for a generic TF (due to the biochemical properties of amino acids in the context of TF-DNA interactions) and it would be worth investigating the mutational spectrum of TFs [Maerkl and Quake, 2007; Maerkl and Quake, 2009]. More importantly, this can be the result of the pleiotropic nature of TFs. Because a TF regulating many genes (by binding to many binding sites) is constrained highly, a major fraction of mutations can be expected to be deleterious. Hence, a major chunk of mutations in such TFs would be quickly lost from the population. The number of mutations that are capable (have non-negligible fixation probability) of getting fixed, either due to selection or drift are low, which shows as a lower rate of TF evolution. Further research into the mutational effects of TF mutations would help us understand the relative rates of the molecular evolution for TFs and their BSs.

3.8 Fast timescale: fixed $\mathbf{d_{rs}}$, evolution of only binding sites

In the monomorphic fixed state limit we assume that the population resides in some specific state $x = (y, z)$ where $y = \{k_{ij}\}$ is the set of binding site mismatches and $z = \{d_{rs}\}$ is the set of TF mismatches. When a mutation occurs, it gets either fixed or lost quickly before another mutation occurs. On the fast timescale, mutations occur only in the binding sites giving rise to a new potential state $x' = (y', z)$. So we are essentially fixing the TF mismatches z and considering the evolution of binding sites alone. The state space X can be separated into $(L+1)^{Q(Q-1)/2}$ groups of $(L+1)^{MQ}$ states each, where each group corresponds to the states that have a specific z. On the fast timescale, the population is constrained to one of these groups that is specified by z. So, we describe binding site evolution as a Markov chain on Y. We have

$$R_{y,y';z} = NU^{BS}_{y,y'} p_{fix}(y \to y'; z), \tag{3.13}$$

where $U^{BS}_{y,y'}$ is the mutation rate from y to y', and $p_{fix}(y \to y'; z)$ is the fixation probability of y' in y for z.

For a fixed z, the entropy term $\Omega_z(y)$ factorizes into M terms, each with Q dimensions,

$$\Omega_z(y) = \prod_j \Omega^{(j)}(k_{1j}, k_{2j}, \ldots, k_{Qj}), \tag{3.14}$$

where $\Omega^{(j)}(k_{1j}, k_{2j}, \ldots, k_{Qj})$ specifies the number of microstates associated with the set of mismatches of j^{th} binding site. When $F(y)$ is also linear in the contribution from each binding site, it also factorizes into terms for each binding site. Hence, $P_z(y)$ also factorizes into M terms of Q-dim each, with each term describing the distribution of mismatches $\{k_{ij}\}$ associated with the j^{th} binding site.

$$P_z(y) = \prod_j P^{(j)}(k_{1j}, k_{2j}, \ldots, k_{Qj}), \tag{3.15}$$

where $P^{(j)}(k_{1j}, k_{2j}, \ldots, k_{Qj})$ is the distribution of mismatches associated with binding site j. Each binding site evolves independently of the others. Even if selection is not weak or the appropriate definition of fitness is multiplicative, one can obtain a steady state distribution $P_z(y)$, only that it will not factorize into terms for each binding site.

3.9 Slow timescale: evolution of TFs

On the slower timescale, in some state $x = (y, z)$, we have mutations in TFs which then get quickly fixed or are lost from the population. A mutation that changes the consensus sequence of TF r can potentially change the $Q-1$ mismatches $\{d_{rs}\}$ to the consensus

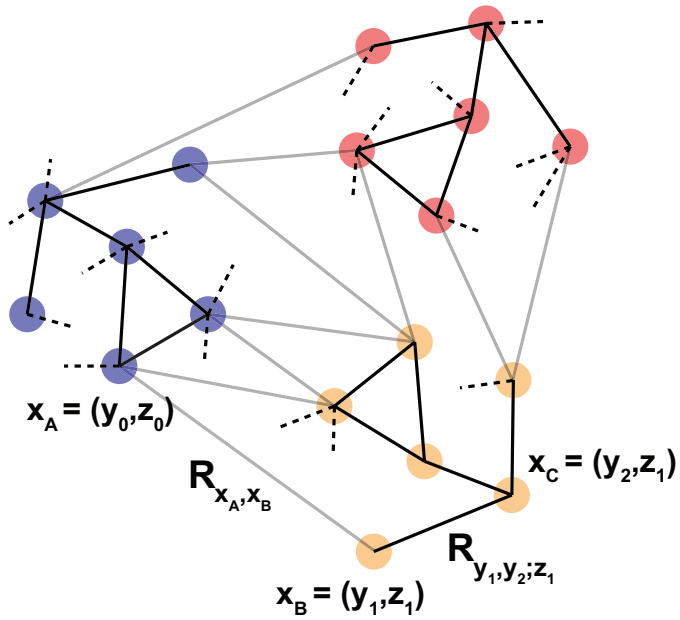


Figure 3.2: **Markov chain for the co-evolution of TFs and binding sites: general considerations and fast-timescale dynamics.** Shown is a Markov chain modeling the coevolution of TFs and binding sites. Each node is a state $x = (y, z)$ in the Markov chain, defined by the set of mismatches $y = \{k_{ij} : i = 1, 2, \ldots, Q, j = 1, 2, \ldots, M\}$ between TF consensus sequences and binding site sequences and the set of TF consensus sequence mismatches $z = \{d_{rs} : r, s = 1, 2, \ldots, Q, r \neq s\}$. Each link defines a transition on this Markov chain with R defining the transition rate. On the fast timescale, only transitions on darker edges occur, and the population is constrained to states with a single color, which shows states with the same TF mismatches z. States x_B and x_C are two such states with the same TF mismatches z_1 (yellow color) but with binding site mismatches y_1 and y_2 respectively. $R_{y_1,y_2;z_1}$ denotes the transition rate between these states. On the slower timescale, transitions on lighter edges occur, changing both y and z, and the population moves to a new colored state. It moves on the dark edges (fast timescale) among the states in the new color till another such light edge transition occurs on the slower timescale. x_A and x_B show a slow timescale transition.

sequences of the other TFs, and also the M mismatches $\{k_{rj}\}$ to all the binding sites. Hence the new mutation $x' = (y', z')$ changes both y and z. This new mutation gets fixed or lost before another mutation, either in the TFs or the binding sites, occurs. The probability of its fixation depends on the fitness difference $\Delta F_{x',x} = F(x') - F(x)$. For TF mismatch state z, the binding site mismatches y would be in their steady state distribution $P_z(y)$ on the fast timescale. So the genetic background y for z on which the new TF mutation z' arises follows a distribution $P_z(y)$.

This gives rise to a picture of TF evolution as a Markov chain with states described by the set of TF mismatches $z = \{d_{rs}\}$, where each state is composed of many internal binding site states y that follow a distribution $P_z(y)$. One can define the transition rates $R_{z,z'}$ between TF states z and z' as an average of all possible binding site mismatch transitions

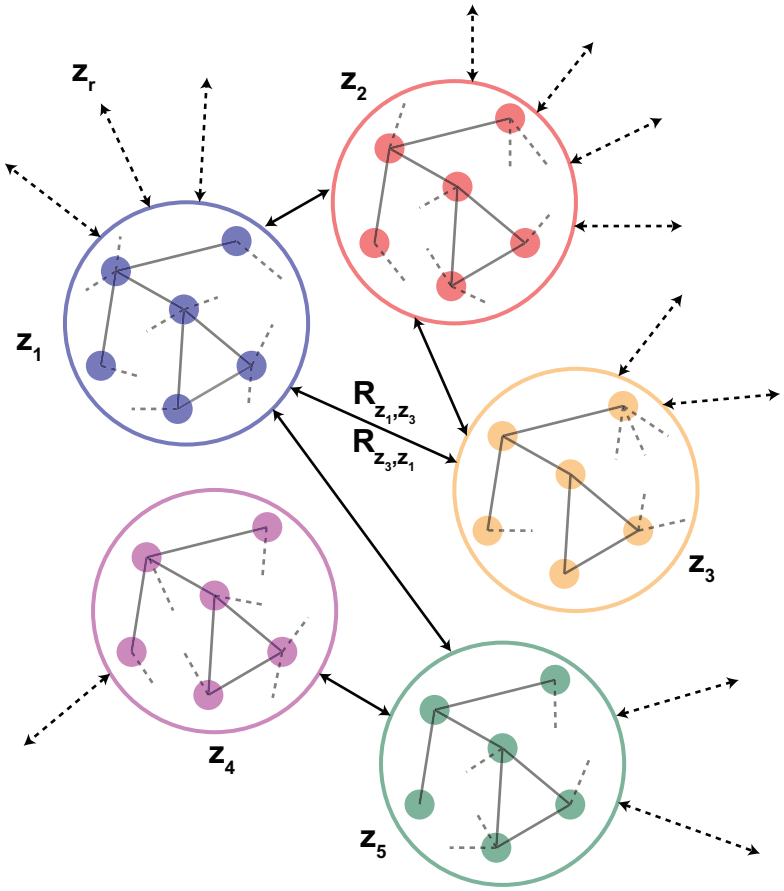


Figure 3.3: **Markov chain for the co-evolution of TFs and binding sites: slow-timescale dynamics.** On the slower timescale, we have a Markov chain describing the jumps between states with different TF mismatches. Here, each state is given by a colored circle, for example, the blue circle denotes the state with TF mismatches z_1 while the green circle denotes the state with TF mismatches z_5. Inside each TF state is embedded a fast-timescale Markov chain (with its states as filled small circles of the same color) whose dynamics happen as described before. For example, the blue state z_1 has a fast-timescale Markov chain with blue small circles inside it, etc. The transition rate from z_1 to z_3 is given by R_{z_1,z_3}.

$y \to y'$.

$$R_{z,z'} = NU^{TF}_{z,z'} \sum_{y} \sum_{y'} \Big[P_z(y) V(y'; y, z \to y', z') p_{fix}(y, z \to y', z') \Big], \tag{3.16}$$

where $U^{TF}_{z,z'}$ is the TF mutation rate from z to z', $P_z(y)$ is the steady state distribution of y for TF state z, $p_{fix}(y, z \to y', z')$ is the fixation probability of (y', z') in (y, z) and $V(y'; y, z \to y', z')$ is the entropic term. Calculating U and V is challenging, and here we will go through an example calculation to get some insight.

3.10 Calculating U and V

Imagine that $x = (y, z)$ where $y = \bigcup_i \{k_{ij} : j = 1, 2, \ldots, M\}$ and $z = \bigcup_r \{d_{rs} : s \neq r\}$. Now suppose, w.l.g., that there is a mutation in the first TF which changes its consensus sequence. This changes its mismatches to the consensus sequence of all other TFs. Suppose that the new TF state is given by

$$z' = \{d'_{1s} : s \neq 1\} \cup \Big[\bigcup_{r \neq 1} \{d_{rs} : s \neq r\} \Big]. \tag{3.17}$$

For a particular z', $U^{TF}_{z,z'}$ can be obtained by the number of ways (sequences) z' can be obtained from z. This mutation also changes the first TF's mismatches to all M binding sites so that the new binding site state is of the form

$$y' = \{k'_{1j}\} \cup \Big[\bigcup_{i \neq 1} \{k_{ij}\} \Big]. \tag{3.18}$$

Suppose that Y is the set of all such y' that can result from this transition. To get V, first we count the number $\Omega_{z'}(y_{new})$ of sequences with y_{new} in z'. Now, for each $y' \in Y$, we obtain V as

$$V(y'; y, z \to y', z') = \frac{\Omega_{z'}(y')}{\sum_{y'' \in Y} \Omega_{z'}(y'')}. \tag{3.19}$$

As a first step in making progress with this calculation, we need to find an expression for $\Omega_z(y)$, the entropic term for binding site mismatches $y = \{k_{ij}\}$ for a given $z = \{d_{rs}\}$. At a given z, this decomposes into terms for each binding site $\Omega^{(j)}_z(y^{(j)})$, where $y^{(j)}$ is the mismatch set for binding site j. While this can be done for 2 TFs (see Eq. 4.18 from Chapter 4), the general solution is hard to obtain.

3.11 Treatment in energy space

Because of these computational hurdles, instead of considering the sequence space with mismatch classes, we directly treat evolution in binding energy space. Genotype is specified by MQ binding energies $\{E_{ij}\}$ between TFs and binding sites, and a covariance matrix Σ of size $Q \times Q$. While the non-diagonal elements of Σ specify how similar the consensus sequences of different TF pairs are, the diagonal elements specify the variance of each binding energy E_{ij}. For a given Σ (which means fixing the TF consensus sequences), in the neutral evolution case, the binding energies $E_{\{i\}j}$ for each binding site j are follow a multivariate Gaussian distribution with covariance matrix Σ. This is equivalent to the entropic contribution Ω in the mismatch setup. While a mutation in binding site j changes $E_{\{i\}j}$ (in a state-dependent way), mutations in TFs change Σ.

In contrast to the sequence-based framework, here we instead consider on the fast timescale and in the neutral case, the steady state distribution for the binding energies $E_i :=$

$\{E_{ij}\}_{j \in TFs}$ each binding site i to be $P(E_i|\sigma) \sim \mathcal{N}(\bar{E}, \Sigma)$, a Gaussian with mean $\bar{E} = 3\epsilon L/4$ and covariance matrix Σ which depends on ϵ, L and TF mismatches d_{rs} as $\Sigma_{rs} = \frac{\epsilon^2}{4}(3L/4 - d_{rs})$. In the selection case, we have

$$P(E|\sigma) = \frac{\prod_i P(E_i|\sigma) \exp(2N_e F(\sigma, E))}{\langle \exp(2N_e F(\sigma, E)) \rangle_{\prod P_i}}, \tag{3.20}$$

where $E = \{E_i\}_i$ are the binding energies of all sites together, $F(\sigma, E)$ is the fitness of E, σ, $\prod_i P(E_i|\sigma)$ is the total entropic contribution (joint distribution of all binding energies E), $\langle X \rangle_{\prod P_i}$ is the mean of X on this distribution. Now, on the slow timescale, we have for $P(\sigma)$

$$\begin{aligned} \frac{dP(\sigma)}{dt} &= \sum_{\sigma'} \mu_{\sigma' \to \sigma} \langle p_{fix}(\sigma', E' \to \sigma, E) \rangle_{E,E'} P(\sigma') \\ &\quad - \sum_{\sigma'} \mu_{\sigma \to \sigma'} \langle p_{fix}(\sigma, E \to \sigma', E') \rangle_{E,E'} P(\sigma), \end{aligned} \tag{3.21}$$

where $\mu_{\sigma \to \sigma'}$ is the mutation rate from σ to σ' and viceversa, $p_{fix}(\sigma, E \to \sigma', E')$ is the fixation probability of σ', E' in σ, E, and $\langle X \rangle_{E,E'}$ is the mean of X over the joint distribution of E and E'. We have

$$\langle p_{fix}(\sigma, E \to \sigma', E') \rangle_{E,E'} = \iint dE dE' p_{fix}(\sigma, E \to \sigma', E') P(E, E'|\sigma, \sigma') \tag{3.22}$$

$$= \iint dE dE' p_{fix}(\sigma, E \to \sigma', E') P(E'|E, \sigma, \sigma') P(E|\sigma, \sigma') \tag{3.23}$$

$$= \iint dE dE' p_{fix}(\sigma, E \to \sigma', E') P(E'|E, \sigma, \sigma') P(E|\sigma), \tag{3.24}$$

where $P(E, E'|\sigma, \sigma')$ is the joint distribution of E and E', given σ and σ', which can be written as the product of $P(E'|E, \sigma, \sigma')$ and $P(E|\sigma)$. $P(E'|E, \sigma, \sigma')$ is the distribution of new binding energies E', given that you jump from energies E and TF sequences σ to σ'. We expect this "jump probability distribution" to satisfy two broad constraints: E' should be close to E with a width of size ϵ and E' should satisfy the constraints due to σ' via $P(E'|\sigma')$ or the like.

As an example, consider the mutation in TF r such that σ changes to σ'. This changes E to E'. The binding energies $\{E_{rj}\}_j$ of each binding site j change to $\{E'_{rj}\}_j$ but all other energies remain same: $\{E'_{sj}\}_{s \neq r, j} = \{E_{sj}\}_{s \neq r, j}$. Let us consider one particular binding site i, whose binding energies $E_i := \{E_{si}\}_s$ intially come from $P(E_i|\sigma)$. In the weak selection limit, we can consider binding sites separately. In the new state E', for binding site i, only E_{ri} changes to E'_{ri} while for all other $s \neq r$, $E'_{si} = E_{si}$. We need to find $P(E'_{ri}|E'_{si} = E_{si}\ \forall s \neq r, E_{ri}, \sigma, \sigma')$. We have

$$P(E'_{ri}|E'_{si} = E_{si}\ \forall s \neq r, E_{ri}, \sigma, \sigma') \ = \frac{P(E'_{ri}, E_{si}\ \forall s \neq r|E_{ri}, \sigma, \sigma')}{P(E_{si}\ \forall s \neq r|E_{ri}, \sigma, \sigma')}, \tag{3.25}$$

where $P(E'_{ri}, E_{si}\ \forall s \neq r|E_{ri}, \sigma, \sigma')$ is the joint distribution of E'_{ri} and $E_{si}\ \forall s \neq r$, given everything else. E'_{ri} will have a mean E_{ri}, variance ϵ^2, and covariance between E'_{ri} and

$E_{si}\ \forall s \neq r$ will be according to Σ' corresponding to σ'. Because they do not change, the means, variances and covariances of all other energies $E_{si}\ \forall s \neq r$ follow from $P(E_{si}\ \forall s \neq r|\sigma)$ which is a marginal of the initial distribution $P(E_i|\sigma)$. We also have

$$P(E_{si}\ \forall s \neq r|E_{ri}, \sigma, \sigma') \ = \frac{P(E_i|\sigma)}{\int dE_{si} P(E_i|\sigma)}. \tag{3.26}$$

In the neutral case, $P(E'_{ri}, E_{si} \forall s \neq r|E_{ri}, \sigma, \sigma')$ is Gaussian with the means and covariance matrix changed as mentioned before, and $P(E_{si}\ \forall s \neq r|E_{ri}, \sigma, \sigma')$ comes from $P(E_i|\sigma)$ which is also a Gaussian. Once we obtain all the terms of the Master equation for $P(\sigma)$, we can numerically solve it to understand evolution in energy space.

We have assumed neutrality in the above approach, and leave the case of selection in the energy space to future research as full genotype treatment for a generic GRN is a hard and complicated problem.

3.12 Summary

In this chapter, I introduced a generic theoretical framework to investigate questions about the coevolution of many transcription factors and their binding sites involved in a gene regulatory network. At the core, this model is based in sequence space – TF consensus sequences, BS sequences, and sequences of other domains of the TF that are important towards the functioning of the GRN. By building on existing knowledge of the biophysics of TF-DNA binding (Chapter 1), we go from genotype to phenotype, which is the set of binding probabilities of various TFs to various BSs. Then by invoking a "function" for the regulatory network, I described various ways to define the fitness of the regulatory programs. Such a generic model is very rich and flexible, and allows for the modeling of temporal processes like developmental programs, and also processes that require, from a fitness point of view, a certain average behaviour from the cell, for instance, nutrient uptake and metabolism. The technical details become involved, and the problem turns hard, when one considers the full calculation of evolutionary steady states and evolutionary dynamics in the sequence space. I briefly explore an alternative way this problem by considering evolution in energy space rather than sequence space. While a more thorough exploration of this is left to future research, there are, however, tractable and biologically relevant cases if we restrict ourselves to evolutionary dynamics in sequence space for 2 TFs only, a problem we investigate in the next chapter.

Chapter 4

Coevolution of duplicated TFs with their binding sites

The work presented in this chapter was performed in collaboration with Tamar Friedlander and has been published in Nature Communications (see [Friedlander *et al.*, 2017]). Tamar Friendlander came up with the calculation in Eq. 4.18, and also did some calculations in Section 4.9.

4.1 Introduction

There is a large body of evidence showing that phenotypic evolution takes place primarily through changes in gene regulation [King and Wilson, 1975; Gilad *et al.*, 2006; Wray, 2007; Carroll, 2005], and that such evolution may be flexible and rapid [Yona *et al.*, 2015; Madan Babu *et al.*, 2006]. But the relative contribution of TF evolution, via coding-sequence changes in the genes that code for these proteins, as compared with regulatory sequence evolution that modify binding site sequences, is one that is not clearly understood yet. TFs play an important role in signal transmission, hence it is important to consider TFs together with the upstream signals they sense and the downstream target genes they regulate. Mutations in the TFs alter the affinity and specificity of TF proteins towards their upstream signals and their downstream binding sites. Together with changes in the binding sites themselves on the DNA, the organism uses TF mutations to "rewire" gene regulatory networks - weaken or remove extant interactions and add new ones, either functional or spurious - to explore the space of possible regulatory networks. However such evolutionary exploration is constrained by

the functionality of the intermediate gene regulatory network states, and hence affects the emergence of novel functions.

A primary mechanism by which regulatory networks evolve and diversify in functionality is via gene duplication [Ohno, 1970; Magadum *et al.*, 2013; Yona *et al.*, 2012] of genes that code for transcription factors. TFs exist in multiple paralogous families, the members of which have probably originated by such gene duplication events, as can be inferred from structure and sequence alignment of TF protein sequences. Different organisms have different number of TFs in these paralogous families, and some families are restricted to only a subset of species, indicating varying selection pressures across organisms. Such gene duplication of TFs gives rise to gene regulatory divergence, resulting in a variety of new phenotypes. For instance, a typical evolutionary outcome is that a regulatory function previously accomplished by a single (or several) TF(s) is now carried out by a larger number of TFs, allowing for additional fine-tuning and precision, or, alternatively, for an expansion of the regulatory scope [Kacser and Beeby, 1984; Simionato *et al.*, 2007; Larroux *et al.*, 2008; Hobert *et al.*, 2010; Achim and Arendt, 2014; McKeown *et al.*, 2014; Baker *et al.*, 2011; Sayou *et al.*, 2014; Pougach *et al.*, 2014; Nadimpalli *et al.*, 2015; Arendt, 2008]. As the two copies are degenerate in the organisms, initial preservation and fixation of duplicates takes place via either a rapid weakening of expression of the duplicates [Lan and Pritchard, 2016] or alternatively a selection to increase expression levels [Conant *et al.*, 2014; Loehlin and Carroll, 2016]. Following this, the preserved extra copies of the TFs thus provide the "raw material" required for further modifications leading to evolutionary diversification with an additional TF. The post-duplication specialization of TFs often involves divergence in both their inputs (e.g., ligands) and outputs (regulated genes) [Wray, 2007; Wittkopp and Kalay, 2012]. Examples range from repressors involved in bacterial carbon metabolism that arose from the same ancestor via a series of duplication-divergence events [Nguyen and Saier, 1995], and ancestral TF Lys14 in the metabolism of *S. cerevisiae*, which diverged into 3 different TFs regulating different subsets of genes in *C. albicans* [Pérez *et al.*, 2014], and the MalR paralogs involved in the metabolism of maltose-like and palatinose-like sugars in *S. cerevisiae* [Pougach *et al.*, 2014], to many variants of Lim and Pou-homeobox genes involved in neural development across different organisms [Hobert and Westphal, 2000] and many more.

After a TF duplication event, because the two copies of the TF genes are identical in sequence, molecular recognition between TFs, their input signals, and their binding sites is specific but undifferentiated between the TF copies. Under selection pressure for the TFs to specialize into different functions, recognition sequences and ligand preferences of the two TFs can diverge by subsequent mutations. But this can happen only under the constraint that some degree of matching between TFs and their binding sites (and upstream signals) is continually maintained to ensure network function. This results in a coevolutionary constraint between the TFs and the binding sites, ultimately affecting the likelihood of potential evolutionary trajectories, and the relevant timescales involved in traversing them. However, very little is known about the resulting limits to evolutionary outcomes; specifically, it is unclear how these likelihoods and timescales depend on important parameters, such as the number of regulated genes, the length and specificity

of the binding sites, the correlations between the input signals, etc.

On the other hand, TF duplication has been studied very little from a theoretical perspective. One class of models consider the framework of gene duplication-differentiation and largely focus on the sub-functionalization of proteins like enzymes that do not have any regulatory role [Innan and Kondrashov, 2010]. A few studies have included cis-regulatory mutations [Force *et al.*, 1999; Lynch and Force, 2000; Force *et al.*, 2005; Proulx, 2012], but only in a simplified fashion, e.g., by considering a small number of discrete alleles that represent TF binding sites appearing and disappearing at fixed rates [Force *et al.*, 2005; Proulx, 2012]. Such an approach ignores the essentials of molecular recognition, and hence cannot model co-evolution between TFs and their binding sites.

Another set of models deal with regulatory sequences explicitly and consider the evolution of gene regulation using a biophysical description [Shea and Ackers, 1985; Kinney *et al.*, 2010; Sherman and Cohen, 2012; He *et al.*, 2010] based in the sequence space. Such an approach captures the essential details of the TF-DNA molecular recognition and accounts for the fact that TFs can bind a variety of DNA sequence with varying affinities [Maerkl and Quake, 2007; Wunderlich and Mirny, 2009; Payne and Wagner, 2014]. However, most studies have focused only on the evolution of binding sites while keeping the TF constant [Payne and Wagner, 2014; Berg *et al.*, 2004; Lässig, 2007; Lynch and Hagner, 2015; Tuğrul *et al.*, 2015], hence largely leaving out TF duplication and their subsequent evolution (but see [Poelwijk *et al.*, 2006; Burda *et al.*, 2010]).

In this chapter, I will describe how we combine these two complementary frameworks of a biophysical description of gene regulation and evolutionary modeling of TF-BS interactions, to define a modeling framework that will let us understand the role of TF duplication and TF-BS coevolution in biophysically realistic fitness landscape on the sequence space. This is a departure from previously considered simpler models that have been "artificially" constructed by manually adding in features, and suggest that realistic landscapes emerge out of simple biophysical and functional constraints and exert a major influence on the evolutionary dynamics and outcomes.

I will first introduce the basic model with two activating TFs and two target genes. I will describe the steady state distribution of evolutionary outcomes, showing that a few functional phenotypes describe the entire evolution over the huge genotypic space. I will show how the statistics of upstream environmental signals affects these steady state outcomes and also consider an alternative model in which TFs act as repressors. Then I will describe the possible evolutionary trajectories and timescales, showing that two major pathways exist, whose likelihoods and speeds depend on various biophysical and evolutionary parameters. I will investigate the role of crosstalk interactions on steady states and evolutionary pathways, and also compare our sequence space models with a simpler biallelic type model. Next, I will extend the basic model to consider the case of each TF regulating multiple downstream target genes, and show that the resulting timescales becomes long as the fitness landscapes become increasingly rugged. Finally, I will study the effect of "promiscuity-promoting mutations", a new type of mutation that has been observed empirically [Sayou *et al.*, 2014; Pougach *et al.*, 2014] to decrease the

binding specificity of TFs in specific sites, on the timescales of specialization.

4.2 Model description and parameters

4.2.1 Biophysical model

In our model, n_{TF} transcription factors regulate n_G genes by binding to sites of length L base pairs; for simplicity, we consider each gene to have one such binding site. The specificity of a TF for any sequence is determined by the TF's preferred (consensus) sequence; sequences matching consensus are assigned lowest energy, $E = 0$, which corresponds to tightest binding, and every mismatch between the consensus and the binding site increases the energy by ϵ; this additive "mismatch" model has a long history in the literature [Maerkl and Quake, 2007; Lässig, 2007; Von Hippel and Berg, 1986; Gerland *et al.*, 2002].

The equilibrium probability that the binding site of gene j ($j = 1, \ldots, n_G$) is bound by active TFs of any type i ($i = 1, \ldots, n_{\mathrm{TF}}$) is a proxy for the gene expression level and is given by the thermodynamic model of gene regulation [Shea and Ackers, 1985; Bintu *et al.*, 2005]:

$$p_{jm}(\{k_{ij}\}, \{C_i(m)\}) = \frac{\sum_i C_i(m) e^{-\epsilon k_{ij}}}{1 + \sum_i C_i(m) e^{-\epsilon k_{ij}}}, \tag{4.1}$$

where $C_i(m)$ is dimensionless concentration of active TFs of type i in condition m, k_{ij} is the number of mismatches between the consensus sequence of the i-th TF species and the binding site of the j-th gene, and ϵ is the energy per mismatch in units of $k_B T$. Concentration $C_i(m)$ of active TFs depends on condition m, which can represent either time or space (e.g., during developmental gene expression programs) or a discrete external environment (e.g., the presence/absence of particular chemical signals).

The simplest case considered here assumes the existence of two such signals that can be either present or absent, in any combination, for a total number of 4 possible environments ($m = 00, 01, 10, 11$), occurring with probabilities α_m; an important parameter will be the correlation, $-1 \leqslant \rho \leqslant 1$, between the two signals. The presence ('1') or absence ('0') of these two signals defines the different environments $m \in \{00, 01, 10, 11\}$ that are possible, with α_m denoting the frequency of environment m. These probabilities can be expressed in terms of three important parameters - f_1, f_2, the frequencies of each signal, and ρ, the correlation between the signals. We have

$$\alpha_{11} = f_1 f_2 + \rho\delta, \tag{4.2}$$
$$\alpha_{10} = f_1(1 - f_2) - \rho\delta, \tag{4.3}$$
$$\alpha_{01} = f_2(1 - f_1) - \rho\delta, \tag{4.4}$$
$$\alpha_{00} = 1 - \alpha_{11} - \alpha_{10} - \alpha_{01}, \tag{4.5}$$

where $\delta = \sqrt{f_1 f_2 (1-f_1)(1-f_2)}$. The frequency of each signal can be obtained as $f_1 = \alpha_{10} + \alpha_{11}$ and $f_2 = \alpha_{01} + \alpha_{11}$. We assume that both the signals are present at equal frequencies, and that each signal is present (or absent) half the time, resulting in $f_1 = f_2 = 0.5$. Hence, we have

$$\alpha_{00} = \alpha_{11} = \frac{1}{4}(1+\rho), \tag{4.6}$$

$$\alpha_{10} = \alpha_{01} = \frac{1}{4}(1-\rho). \tag{4.7}$$

Thus when the signals are uncorrelated ($\rho = 0$), we have $\alpha_{00} = \alpha_{10} = \alpha_{01} = \alpha_{11} = 1/4$. When the signals are fully correlated ($\rho = 1$) we obtain $\alpha_{00} = \alpha_{11} = 0.5$ and $\alpha_{10} = \alpha_{01} = 0$ and vice versa for anti-correlation ($\rho = -1$). We explore asymmetric environments in Sec. 4.4.

Fig. 4.1a,b illustrate this setup for this simple case $n_{\mathrm{TF}} = n_G = 2$, assuming that the two copies of the TF emerged through an initial gene duplication event and are fixed in the population. Transcription factors are equipped with an evolvable signal sensing domain (captured by $\sigma_i \in [00, 01, 10, 11]$). The original TF regulates two downstream genes by binding to their binding sites. It is sensitive to both external signals, which can be present with a varying degree of correlation (Fig. 4.1a). Each of the downstream genes is suitable to respond to only one of the two signals. Before duplication the genes are constrained to follow the only TF available which responds to both signals. The extra TF formed in the duplication event offers an additional degree of freedom in regulating these genes, if the TFs specialize such that each of them senses only one of the two signals and regulates only a subset of the genes. If the TF i is responsive to a signal and that signal is present in environment m, then its active concentration $C_i(m) = C_0$; otherwise, $C_i(m) = 0$. Given constants C_0, ϵ, and the genotype $\mathcal{D}$—comprising TF consensus and binding site sequences as well as TF sensitivity alleles σ_i—the thermodynamic model of Eq. (4.1) fully specifies expression levels for all genes in all environments.

This framework is applicable to more general pathway architecture than a TF that implements both signal sensing and gene regulation in the same molecule. Often these two functions are split between different components of the same pathway; for example, a separate upstream component senses the signal(s) and consequently activates the TF (e.g. by phosphorylation or another modification). Additionally, TF production is also regulated. One can also think of the evolution of the regulatory sequences of the gene coding for the TF in terms of our model. Since our model is defined in very general terms, it can capture such situations as well.

Gene birth can occur via different biological mechanisms, some of them allowing for the emergence of slightly modified copies of original genes or allowing for different regulation of the same coding sequence. One such mechanism is called 'retroposition': creation of duplicate gene copies in new genomic positions through the reverse transcription of mRNAs from source genes (also known as RNA-based duplication or retroduplication) [Kaessmann *et al.*, 2009]. These newly formed genes often lack regulatory elements of the parental gene and may also be slightly modified due to transcription errors (that are significantly more common than DNA-duplication errors). It was shown that transcription

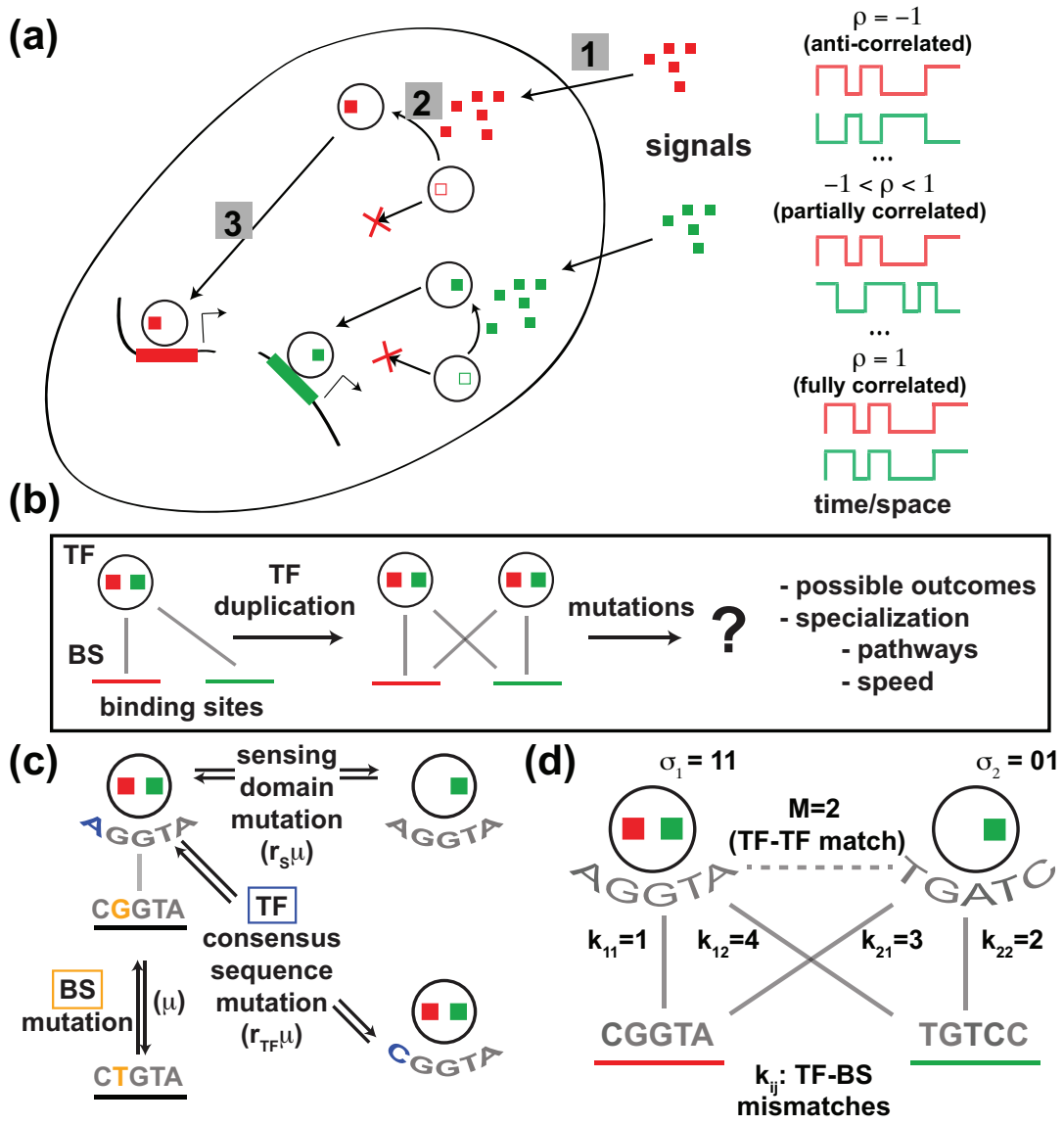


Figure 4.1: **Schematic of the TF duplication model.** **(a)** Simplified physiology of signal transduction: external signaling molecules (red and green squares) are sensed by the cell (1), activate transcription factors inside the cell (2), which in turn activate the corresponding downstream genes (3). The temporal / spatial appearance of the two external signals can be correlated to different extent, as measured by correlation coefficient, ρ. These signals can correspond to different time periods in development, spatial regions in the organism or tissue, or external conditions / ligands. **(b)** TF, initially responsive to two external signals (red and green "slots") and regulating two genes, duplicates and the additional copy fixes in the population. Immediately after duplication, the two copies are undifferentiated. (Continued in the next page.)

Figure 4.1: (Continued from the previous page.) **(c)** Various mutation types that can occur post-duplication with their associated rates. **(d)** After accumulating several mutations, the pattern of mismatches between TF consensus sequences and the binding sites is reflected in new values of $\{k_{ij}\}$, which determine the activation levels of the two genes according to Eq. (4.1). M, the number of matches between the consensus sequences of the two TFs (with a value between 0 and L), keeps track of the overall divergence of the TF specificities. For a list of model parameters and baseline values see Table 4.1.

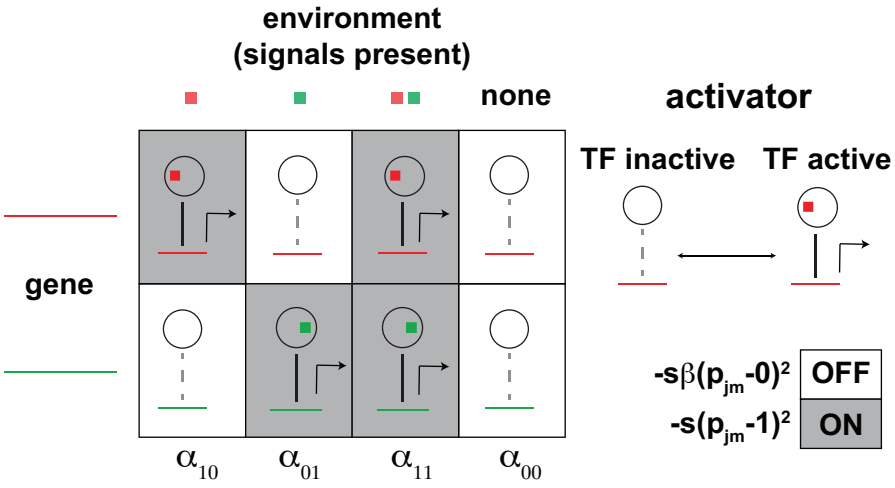


Figure 4.2: **Optimal expression patterns and fitness contributions in different environments.** In the basic model, TFs are considered to be activators, which in the presence of signals, can bind to BSs and activate the corresponding genes. Shown are the optimal expression patterns of the two genes in the four different environments, and the mechanistic components of the genotype that can achieve these optimal patterns.

of these so-called 'retrogenes' is very common and often relies on regulatory elements of neighboring genes [Vinckenbosch *et al.*, 2006].

4.2.2 Evolutionary model

After duplication, three types of mutation can occur, as shown in Fig. 4.1c: point mutations in the binding sites (rate μ), mutations in the TF coding sequence that change TF's preferred (consensus) specificity (rate $r_{\mathrm{TF}}\mu$) and mutations in the two signal-sensing alleles (rate $r_{\mathrm{S}}\mu$), which can give each TF specificity to both signals, to one of them, or to neither. An example in Fig. 4.1d shows the state of the system after several mutations have affected the degree of (mis)match between the TFs and the binding sites, k_{ij}; an especially important quantity that tracks the overall divergence of the TF specificity is denoted as M, the match between the two TF consensus sequences.

We define fitness such that the specialized genotypes have higher fitness compared to the initial non-specialized genotypes. As shown in Fig. 4.2, we define the ideal expression level of gene j in environment m, p^*_{jm}, such that $p^*_{jm} = 1$ if signal j is present in environment

m and $p^*_{jm} = 0$ if signal j is absent in environment m. The fitness of a genotype equals the squared deviation of the actual expression p_{jm} from the ideal one p^*_{jm}, summed over all genes j and averaged over all environments m:

$$F = -s \sum_j \sum_m \alpha_m \beta_{jm} (p_{jm} - p^*_{jm})^2, \tag{4.8}$$

where s denotes the selection intensity and α_m is the frequency of the m-th environment. We define environments by the presence or absence of the signals, which result in different active TF concentrations depending on their signal responsiveness. β_{jm} is the penalty for each type of deviation from the ideal expression level, allowing for diverse penalties for different genes or at different environments. For example, a gene which is not expressed when needed can incur a higher penalty than the expression of a gene that is not necessary in a given environment. To capture these latter interactions, which we call crosstalk interactions, we exploited β_{jm} to tune the fitness penalty in Section 4.7. Expression levels p_{jm} for a genotype are calculated using Eq. (4.1) by obtaining the dimensionless concentrations of the TFs, $C_i(m)$, from their signal sensing alleles σ_i, and the mismatches, k_{ij}, from the TF consensus sequences and the BS sequences.

4.2.3 Putting the pieces together

With the fitness of genotypes and the mutations between them defined, we consider an evolutionary framework to study the evolutionary dynamics of this regulatory system. We assume mutation rates to be low enough such that a beneficial mutation fixes before an additional mutation (beneficial or not) arises. The condition under which this assumption is valid was recently rediscovered by Desai and Fisher [Desai and Fisher, 2007] and reads $\frac{\log(4N\Delta F)}{\Delta F} \ll \frac{1}{4N\mu_b \Delta F}$, and is based on Gillespie's Strong-Selection-Weak-Mutation (SSWM) framework [Gillespie, 1983; Gillespie, 1984; Gillespie, 1994]. ΔF is the fitness advantage of the beneficial mutant, N is the population size and μ_b is the rate of beneficial mutations.

Under this condition the population is almost always fixed (monomorphic), and its evolutionary trajectory is captured by a series of discrete transitions between different genotypes. Consequently, when a new mutation emerges, it competes with only one other genotype. The fixation probability of a new mutation that alters the genotype from y to x equals

$$\Phi_{y \to x} = \frac{1 - \exp(-(F(x) - F(y)))}{1 - \exp(-2N(F(x) - F(y)))}, \tag{4.9}$$

where the fitness F is defined by Eq. (4.8) given the frequencies of the various environments α_m and the desired expression pattern of the genes p^*_{jm} at each. Eq. (4.9) applies to a diploid population in which the mutant x appears in a single copy over a uniform background of the other genotype y. For diploids, the fitness difference $\Delta F = F(x) - F(y)$ refers to the fitness difference between the two homozygotes or to twice the selective advantage of the heterozygote (one copy of the mutant) over the prevailing homozygote

Parameter	Explanation	Baseline value
L	BS length, length of sequences that TFs bind	5
ϵ	Energy contribution per bp towards TF-BS binding	$3\,k_BT$
C_0	Active TF free concentration in the presence of signal, set such that $p_{jm} = 0.5$ at $k = 1.5$	4.5
f_1	Frequency of signal 1	0.5
f_2	Frequency of signal 2	0.5
ρ	Correlation between input signals	0
β_X	β_{jm} when $p^*_{jm} = 0$ for gene j in environment m: penalty in fitness on activating a gene when it is not needed (crosstalk interaction)	0.5
other β	β_{jm} when $p^*_{jm} = 1$ for gene j in environment m: penalty in fitness on not activating a gene when it is needed (functional interaction)	1
Ns	Selection strength	25
r_S	Relative mutation rate of the signal sensing domain compared to the binding site mutation rate per bp	1
r_{TF}	Relative mutation rate of the TF consensus sequence per bp compared to the binding site mutation rate per bp	1

Table 4.1: **Model parameters and their baseline values.**

genotype [Gillespie, 2004]. The overall rate of substitution from genotype y to x is given by [Lässig, 2007]:

$$r_{xy} = 2N\mu_{xy}\Phi_{y\to x}, \tag{4.10}$$

where μ_{xy} denotes the mutation rate from genotype y to x. Note that the fixation probability in Eq. (4.9) below, depends, via the fitness, and in turn via the binding probabilities, directly on the TFs' signal sensing alleles σ_i, and the mismatches k_{ij} of the BS sequences with the TF consensus sequences, but not on M, the match between the TF consensus sequences. But, as shown in Fig. 2A of the main text, the set of possible k_{ij}'s is constrained by M, and hence, there is implicit selection on M. Also, importantly, selection does not directly depend on the TFs and BSs, but only via their biophysical interaction to result in appropriate gene regulation, thereby requiring concerted evolution of TFs and BSs.

In Table 4.1 we list the model parameters and their baseline values used in calculations (unless stated otherwise).

4.2.4 Space of reduced genotypes

The size of the genotype space is huge, $|\mathcal{D}| = 4^{4L+2} \approx 10^{13.25}$ for $L = 5$, which makes it hard to analytically track the evolutionary model. Since the fitnesses of genotypes

depend only on the mismatches k_{ij} and the signal sensing alleles σ_i, and the mutations only alter k_{ij}, σ_i and the TF consensus sequences' match M, we consider the space of "reduced-genotypes", $\mathcal{G} = \{M, k_{ij}, \sigma_i\}$, keeping track of only these reduced features of the genotype. The size of the reduced-genotype space is $|\mathcal{G}| < 16(L+1)^5 \approx 10^{5.09}$ for $L = 5$, which is tractable. Hence, for analytical calculations, we treat the regulatory network in the reduced-genotype space $\mathcal{G}$, and for simulations, we treat the regulatory network in the full genotypic space $\mathcal{D}$. Note that the reduced genotype representation in our model framework is *not* an approximation, but is an exact solution of the full genotype model, with the tractability gained due to clever bookkeeping of states in the sequence space.

Fig. 4.3 shows the interplay of biophysical constraints that give rise to a realistic fitness landscape for our problem. Given a match, M, between two TF consensus sequences, only certain combinations of mismatches, (k_{1j}, k_{2j}), of the TFs with each of the two binding sites are possible. A particular allowed combination can be realized by different numbers of genotypes, as shown in Fig. 4.3a, providing a detailed account of the entropy of the neutral distribution. For each of the four environments, Eq. (4.1) predicts gene expression at every pair of mismatch values (Fig. 4.3b); together with the probabilities of different environments occurring, the gene expression pattern determines the genotypes's fitness, F. TF specialization then unfolds on this landscape by different types of mutations (e.g., Fig. 4.3c). Although the landscape is complex and high-dimensional, it is highly structured and ultimately fully specified by only a handful of biophysical parameters. Furthermore, because of the sigmoidal shape of binding probability as a function of mismatch k [Eq. (4.1)], it is possible to assign phenotypes of "strong" and "weak" binding to every TF-BS interaction, allowing us to depict network interactions graphically, as shown in Fig. 4.3d, and to classify the possible macroscopic evolutionary outcomes, as we will show next.

4.2.5 Classification of genotypes into "macrostates"

Since our interest is in the biological function implemented by the network, we further coarse-grain the space of reduced-genotypes $\mathcal{G}$, and classify these reduced-genotypes into six possible macro-states, $\mathcal{M} = \{$`No Regulation`, `Initial`, `One TF Lost`, `Specialize Both`, `Specialize Binding`, `Partial`$\}$, by distinguishing only between "strong" and "weak" interactions. We set a threshold k_T and consider an interaction as weak, $k_{ij} \in \mathcal{W}$, if $k_{ij} > k_T$, and strong, $k_{ij} \in \mathcal{S}$, if $k_{ij} \leqslant k_T$. In the basic version of the model where both TFs have same biophysical properties (in particular same L) k_T is the same for all TF-BS interactions (but see the extension in Section 4.10). The threshold k_T for each TF-BS pair ij is set such that for mismatches $k < k_T$, $p_{jm_i} \geqslant 0.5$ and for $k > k_T$, $p_{jm_i} < 0.5$ when only TF i is present and other TF(s) are absent, $C_i(m_i) = C_0$.

The full genotypic space $\mathcal{D}$ is a union of sequences belonging to different macrostates z:

$$\mathcal{D} = \bigcup_{z \in \mathcal{M}} S_z, \tag{4.11}$$

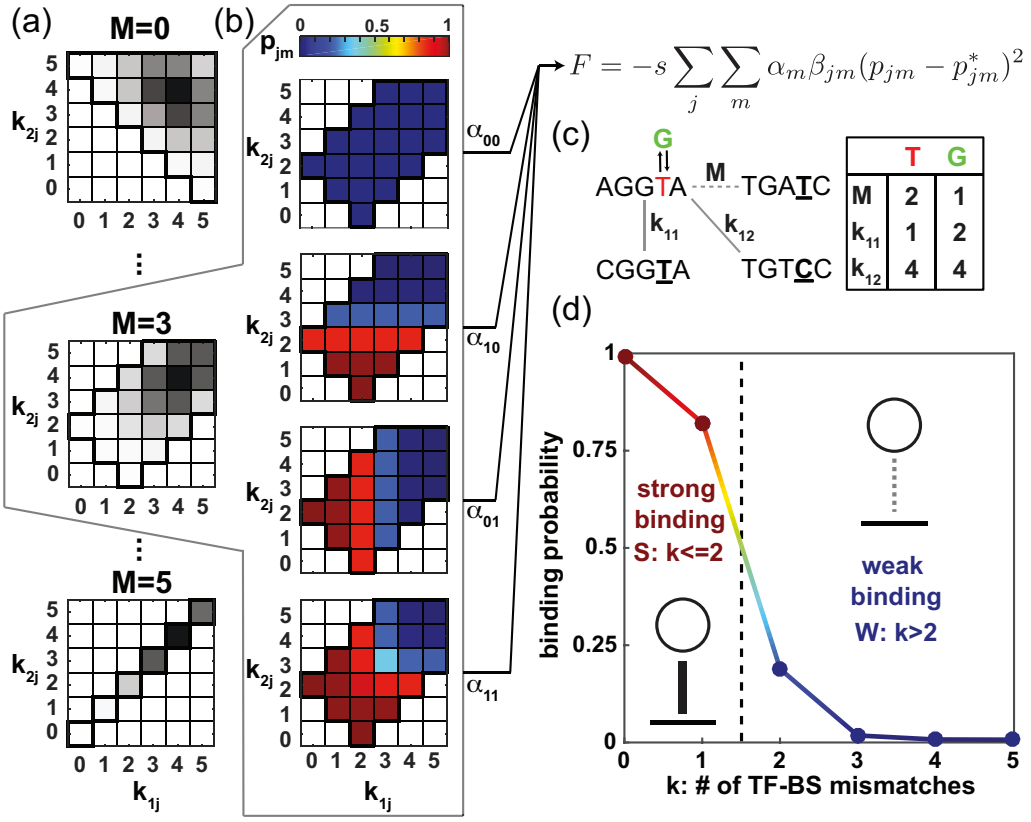


Figure 4.3: **Biophysical and evolutionary constraints shape the genotype-phenotype-fitness map after TF duplication. (a)** Match, M, between transcription factor consensus sequences (here, of length $L = 5$), constrains the possible mismatch values, k_{1j}, k_{2j}, between the gene's binding site and either TF. For example, when the two TFs are identical ($M = L = 5$, bottom left), they must have equal mismatches with all genes ($k_{1j} = k_{2j}$). Some combinations of mismatches are impossible given M (white), while others are realized by different numbers of genotypes (grayscale). **(b)** Expression level (color) for a regulated gene given all mismatch combinations, k_{1j}, k_{2j}, at $M = 3$. Impossible mismatch combinations are colored white. Each of the four panels shows expression levels in four possible environments, $m = 00, 10, 01, 11$. Fitness F depends on the structure of mismatches (a), the biophysics of binding (b), and the frequencies of different environments, α_m. Here we choose α so that the marginal probability of each input signal $f_{1,2}$ is always $f_1 = f_2 = \frac{1}{2}$ but the correlation can be varied, and assign weight $\beta_{jm} = 1$ whenever the gene should be induced but is not, and $\beta_{jm} = \frac{1}{2}$ when it is induced when it should not. The general case when $f_1 \neq f_2 \neq 0.5$ is analyzed in Section 4.4. (Continued in the next page.)

Figure 4.3: (Continued from the previous page.) **(c)** A single point mutation, e.g. a change in one TF's binding specificity from `T` to `G`, can simultaneously affect the match, M, and either increase, decrease, or leave intact the mismatches, k_{11} and k_{12}, that determine fitness. **(d)** TF-BS interactions with mismatch k that is low enough to ensure a high binding probability ($p > 1/2$) are assigned to a "strong binding" phenotype (solid link); conversely, $p < 1/2$ is a "weak binding" phenotype (dotted link).

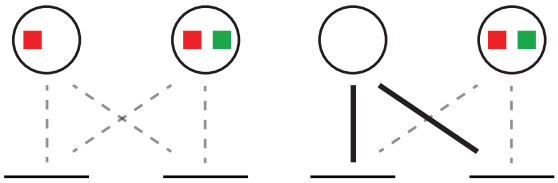


Figure 4.4: **Typical genotypes in `No Regulation` macrostate.** In the left genotype, even though both TFs sense some signals, they do not bind well to either of the binding sites, hence preventing any information transmission. In the right genotype one TF binds both the binding sites but does not sense any signal and the second TF does not bind any binding site even though it senses both signals. This way or the other no information is transmitted between the signals and the genes.

where S_z is the set of all genotypes that belong to macrostate z. We apply the following classification rules.

`No Regulation`

The `No Regulation` macrostate consists of all genotypes in which there is no regulation of any form (no information transmitted from the signals to genes). This can happen if both the TFs either do not sense any signal or do not bind well to any binding sites.

$$x \in S_{\texttt{No Regulation}} \text{ if } \forall i \left((\forall j \; k_{ij} \in \mathcal{W}) \text{ OR } (\sigma_i = 00) \right). \tag{4.12}$$

`Initial`

The `Initial` macrostate consists of all genotypes in which there is complete regulation with no form of specificity: both the TFs sense both signals and bind both binding sites. This is the typical initial state right after duplication.

$$x \in S_{\texttt{Initial}} \text{ if } \forall i \left((\forall j \; k_{ij} \in \mathcal{S}) \text{ AND } (\sigma_i = 11) \right). \tag{4.13}$$

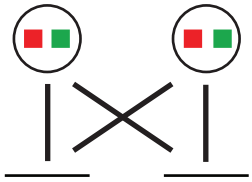


Figure 4.5: `Initial` **macrostate genotypes.** In these genotypes, both TFs sense both signals and bind both binding sites.

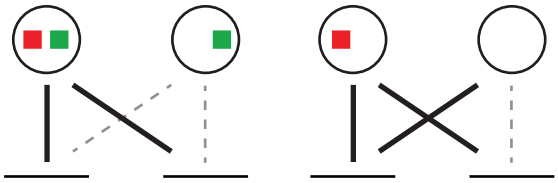


Figure 4.6: **Typical genotypes in `One TF Lost` macrostate.** In the left genotype, only the first TF is involved in regulation as it senses both signals and binds to both binding sites. The second TF senses the green signal but does not bind any of the binding sites, hence it is not involved in regulation and is "lost". In the right genotype, again only the first TF is involved in regulation as it senses the red signal and binds both binding sites. The second TF not involved in any regulation because it does not sense any signal, although it binds the first binding site.

`One TF Lost`

The `One TF Lost` macrostate consists of all genotypes in which one of the TFs is not involved in any regulation while the other is involved in some regulatory activity (namely, one TF does not sense any signal or does not bind well to any of the binding sites). This is equivalent to the genotypes before duplication, except that there is a "lost TF".

$$x \in S_{\texttt{One TF Lost}} \text{ if } \left| i : \Big((\forall j \; k_{ij} \in \mathcal{W}) \text{ OR } (\sigma_i = 00) \Big) \right| = 1. \tag{4.14}$$

`Specialize Both`

The `Specialize Both` macrostate consists of all genotypes in which there is correct specialization of TFs with respect to both signal sensing and binding sites specificity. In these genotypes, one TF senses only the first signal and binds only to the first binding site, while the other TF senses only the second signal and binds only to the second binding site.

$$\begin{aligned} & x \in S_{\texttt{Specialize Both}} \text{ if} \\ & (k_{11}, k_{22} \in \mathcal{S} \text{ AND } k_{12}, k_{21} \in \mathcal{W} \text{ AND } \sigma_1 = 10 \text{ AND } \sigma_2 = 01) \\ \text{OR } & (k_{12}, k_{21} \in \mathcal{S} \text{ AND } k_{11}, k_{22} \in \mathcal{W} \text{ AND } \sigma_1 = 01 \text{ AND } \sigma_2 = 10). \end{aligned} \tag{4.15}$$

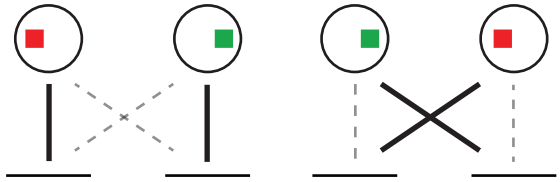


Figure 4.7: **Genotypes in `Specialize Both` macrostate.** Both genotypes have specific paths from the signals to the genes. In the left genotype, while the first TF senses the red signal and binds the first (correct) binding site, the second TF senses the green signal and binds the second (correct) binding site. Hence, the first TF mediates the red signal to first gene pathway while the second TF mediates the green signal to second gene pathway. In the right genotype, the TFs exchange roles. The first TF mediates the green signal to second gene pathway while the second TF mediates the red signal to first gene pathway.

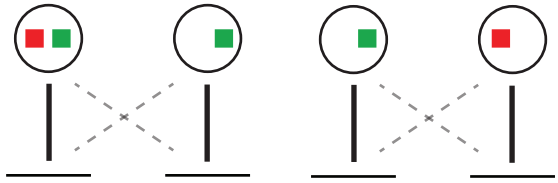


Figure 4.8: **Typical genotypes in `Specialize Binding` macrostate.** In both genotypes, the first TF binds the first binding site and the second TF binds the second binding site, but they have not correctly specialized in their signal sensing domains. In the left genotype, while the second TF has specialized correctly to sense only the green signal, the first TF still senses both the signals. Hence, while the red signal pathway is established properly, the green signal pathway is not - both genes are activated in the presence of green signal. In the right genotype, the TFs have specialized in signal sensitivities, but opposite to the desired response pattern.

`Specialize Binding`

In contrast, the `Specialize Binding` macrostate consists of all genotypes in which there is specialization of TFs with respect to binding site specificities, but not with respect to the signal sensing domains.

$$\begin{aligned} &x \in S_{\texttt{Specialize Binding}} \text{ if } (\forall i\ \sigma_i \neq 00) \text{ AND} \\ &\Big(\big((k_{11}, k_{22} \in \mathcal{S} \text{ AND } k_{12}, k_{21} \in \mathcal{W}) \text{ AND } \neg(\sigma_1 = 10 \text{ AND } \sigma_2 = 01)\big) \\ &\text{OR } \big((k_{12},\ k_{21} \in \mathcal{S} \text{ AND } k_{11}, k_{22} \in \mathcal{W}) \text{ AND } \neg(\sigma_1 = 01 \text{ AND } \sigma_2 = 10)\big)\Big). \end{aligned} \tag{4.16}$$

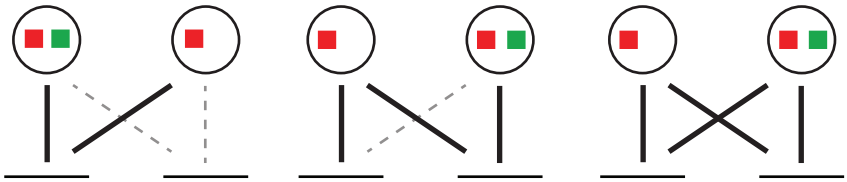


Figure 4.9: **Typical genotypes in `Partial` macrostate.** In the left genotype, both TFs regulate only the first gene while the second gene is unregulated. In the middle genotype, the first TF regulates both genes while the second TF regulates only the second gene. In the right genotype, both TFs regulate both genes but, unlike the `Initial` macrostate, here the first TF does not mediate any information from the green signal.

`Partial`

The `Partial` macrostate consists of all genotypes which do not belong in any of the other macrostates mentioned above. It contains a mixture of different regulatory architectures: both TFs regulate only one gene with the other gene unregulated, one TF regulates both genes while the other TF regulates only one gene or both TFs bind both binding sites but at least one TF has specialized in signal sensing.

Role of L in macrostate classification

Keeping ϵ and C_0 constant while changing L keeps the threshold mismatch k_T constant. Hence, the number of mismatches $|\mathcal{S}|$ in the strong binding class remains the same while the number of mismatches $|\mathcal{W}|$ in the weak binding class increases. Hence, as L increases, the number of genotypes in all macrostates except `Initial` increase. The volume of macrostates with a larger number of weak mismatches increase more than the volume of macrostates with a smaller number of weak mismatches. For instance, `No Regulation` increases more than `Specialize Binding`. As `One TF Lost` and `Specialize Binding` have the same number of weak mismatches, the ratio of the number of genotypes in them stays the same for different L.

4.3 Steady state

We consider mutation rates to be low enough that a beneficial mutation fixes before another beneficial mutation arises [Desai and Fisher, 2007], allowing us to assume that the population is almost always captured by a single genotype. The probability that the population occupies a particular genotypic state, $P(\mathcal{D}, t)$, evolves according to a continuous-time discrete-space Markov chain. Transition rates between states are a product between the mutation rates between different genotypes and the fixation probabilities that depend on the fitness advantage a mutant has over the ancestral genotypes [Lässig, 2007; Kimura, 1962], $r_{xy} = 2N\mu_{xy}\Phi_{y\to x}$, where N is the population size, μ_{xy} is the mutation

rate from genotype y to x, and $\Phi_{y\to x}$ is the probability of fixation of a single copy of x in a population of y. Our model only requires us to keep track of mismatches and not full sequences (i.e. the reduced-genotypes, $\mathcal{G} = \{M, k_{ij}, \sigma_i\}$), which significantly reduces the genotype space dimensionality. This framework allows for calculation of the steady state distribution of genotypes, or reduced-genotypes Eq. Eq. (4.17) and classification of genotypes into relevant macrostates as we describe later.

Evolutionary outcomes in steady state are determined by a balance between selection and drift. The steady state distribution over reduced-genotypes is [Gillespie, 2004]

$$P_{\rm SS}(\mathcal{G}) = P(\mathcal{G}, t\to\infty) = P_0(\mathcal{G})\exp(2NF(\mathcal{G})), \tag{4.17}$$

where P_0 is the neutral distribution of genotypes and N is the population size. Eq. (4.17) is similar to the energy/entropy balance of statistical physics [Berg *et al.*, 2004; Sella and Hirsh, 2005], with fitness F playing the role of energy and $\log P_0$ the role of entropy; in our model, both of these quantities are explicitly computable, as is the resulting steady state distribution. $P_{SS}(\mathcal{G})$ is the non-trivial solution of $\mathbf{R}P_{SS}(\mathcal{G}) = 0$. It is also possible to obtain $P_{SS}(\mathcal{G})$ by invoking the set of detailed balance conditions, $r_{xy}P_{SS}(y) = r_{yx}P_{SS}(x)$, $\forall x, y$.

To calculate the neutral distribution P_0 of the reduced-genotypes (distribution in the absence of selection), we enumerate the number of possible BS sequences j that have mismatch values (k_{1j},k_{2j}) with respect to two TFs that match each other at M out of L consensus positions:

$$\begin{aligned}
N_{\rm seq}(k_1,k_2|M) &= \sum_{j_0=j_0^{\min}}^{j_0^{\max}} \binom{M}{j_0} 3^{M-j_0} \binom{L-M}{L-j_0-k_1} \binom{j_0+k_1-M}{L-j_0-k_2} 2^{k_1+k_2+2j_0-L-M},\\
j_0^{\min} &= \max(\max(0, M-\min(k_1,k_2)), \lceil \frac{L+M-k_1-k_2}{2} \rceil),\\
j_0^{\max} &= \min(M, L-\max(k_1,k_2)).
\end{aligned} \tag{4.18}$$

The neutral distribution (up to proportionality constant) equals

$$P_0(x) \sim N_{\rm seq}(k_{11},k_{21}|M) N_{\rm seq}(k_{12},k_{22}|M) \binom{L}{M} 3^{L-M}. \tag{4.19}$$

Understanding the high dimensional distribution over genotypes is difficult, but classification of individual TF-BS interactions into "strong" and "weak" ones, as described above, allows us to systematically and uniquely assign every genotype to one of a few possible macroscopic outcomes, or "macrostates," graphically depicted in Fig. 4.10a and defined precisely in Section. 4.2.5. Thus, in the `No Regulation` state, input signals are not transduced to the target genes, either because TF-BS mismatches are high and there is no binding or because TFs themselves lose responsiveness to the input signals; in the `One TF Lost` state, a single TF regulates both genes (as before duplication), while the other TF is lost, i.e., its specificity has diverged so far that it does not bind any of the sites;

the `Specialize Binding` state corresponds to each TF regulating its own gene without cross-regulating the other but the signal sensing domains are not yet signal specific, as they are in the `Specialize Both`, the state which we have defined to have the highest fitness. Finally, the `Partial` macrostate predominantly features configurations where each of the TFs binds at least one binding site, but one of the TFs still binds both sites or retains responsiveness for both input signals; functionally, these configurations lead to large "crosstalk," where input signals are non-selectively transmitted to both target genes.

Ultimately, these macrostates are the functional network phenotypes that we care about. The number of genotypes in each macrostate, however, can vary by orders of magnitude; for example, the `No Regulation` state is larger by $\sim 10^4$ relative to the high-fitness `Specialize Both` state, for our baseline choice of parameters ($L = 5, \epsilon = 3$). Selection can act against this strong entropic bias, and the distribution of fitness values across genotypes within each macrostate is shown in Fig. 4.10b. Clearly, the mean or median fitness within each macrostate is a poor substitute for the detailed structure of fitness levels that depend nonlinearly on TF-BS mismatches and the degeneracy of the sequence space. Unlike the entropic term in Fig. 4.10b, fitness also depends on the statistics of the environment, α_m, and in particular, the correlation ρ between the two signals. For example, when the signals are strongly correlated, the `Initial` state right after duplication or the `One TF Lost` state can achieve quite high fitnesses, since responding to the wrong signal or having a high degree of crosstalk will still ensure largely appropriate gene expression pattern in all likely environments. In contrast, at strong negative correlation, many genotypes in `Specialize Binding` and `Initial` states will suffer a large fitness penalty because their sensing domains are not specialized for the correct signals, while the `Specialize Both` state will have high fitness regardless of the environmental signal correlation.

How do fitness and entropy combine to determine macroscopic evolutionary outcomes? Fig. 4.11a shows the most probable macrostate as a function of selection strength and signal correlation. From Eq. (4.17) we obtain the steady state distribution over the macrostate space. For every macrostate $z \in \mathcal{M}$ the probability to be in this macrostate at steady state equals the sum of probabilities of being in all reduced-genotypes x that are assigned to that macrostate

$$Q_{SS}(z) = \sum_{x \in S_z} P_{SS}(x). \tag{4.20}$$

We denote the the most probable macrostate at steady state by

$$z^*_{SS} := \arg\max_{z \in \mathcal{M}} Q_{SS}(z). \tag{4.21}$$

At weak selection, specific TF-BS interactions cannot be maintained against mutational entropy and the system settles into the most numerous, `No Regulation` state. Higher selection strengths can maintain a limited number of TF-BS interactions in `Partial` states. Beyond a threshold value for Ns, the evolutionary outcome depends on the signal correlation: when signals are anti-correlated or weakly correlated, the TFs reach the fully

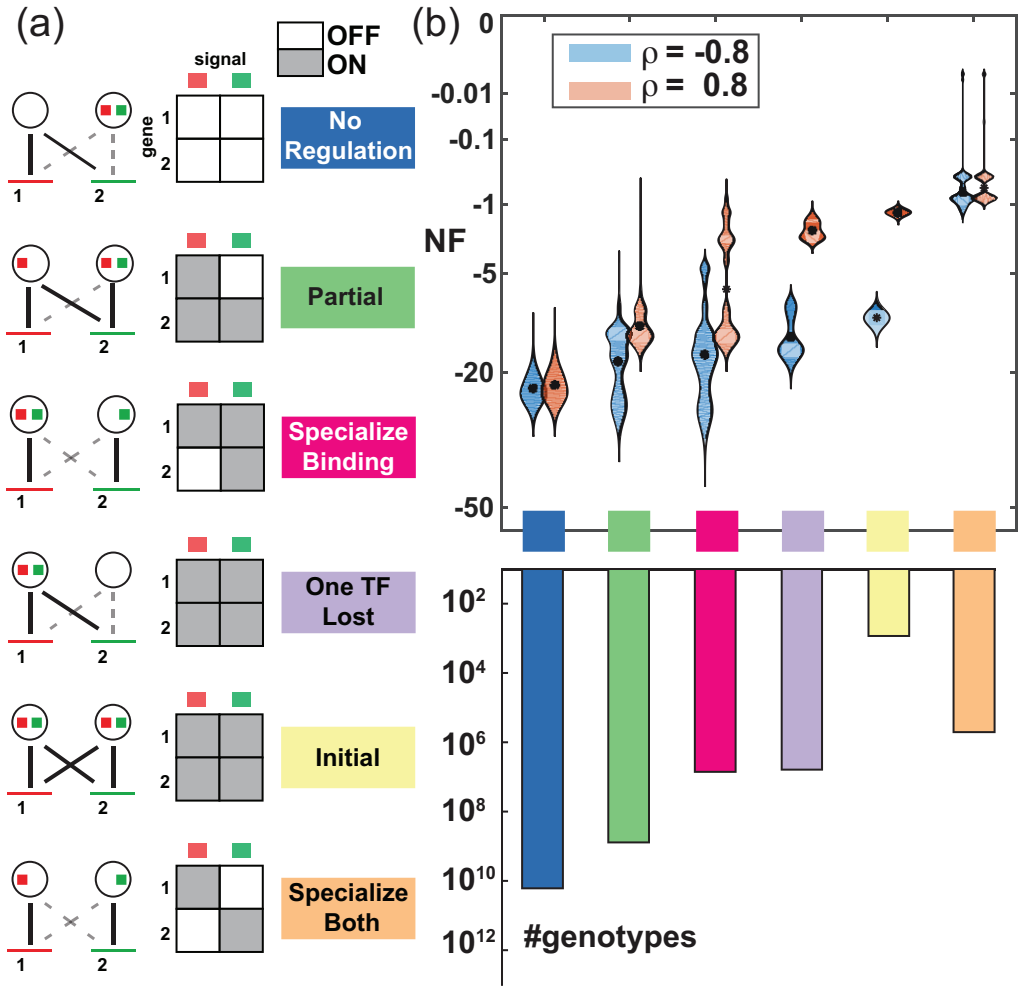


Figure 4.10: **Functional macrostates that are relevant evolutionary outcomes. (a)** Left: evolutionary macrostates (see text) depicted graphically as network phenotypes with solid (dashed) lines indicating strong (weak) TF-BS interactions. Red and green squares in the TFs represent the corresponding signal sensing domains. Right: input-output table, where columns represent the presence of either (red or green) external signal and rows represent the resulting gene activation for each phenotype. **(b)** (Top) Distribution of fitness values, NF, across genotypes in each macrostate (color-coded as in (a)), shown as violin plots, for two values of signal correlation, ρ. Black dots = median fitness in the macrostate. For each macrostate, we show the distribution of NF values for the set of underlying genotypes corresponding to that macrostate. NF is the fitness multiplied with population size – differences in NF are relevant to the strength of selection across genotypes. (Bottom) The number of genotypes in each macrostate (logarithmic scale).

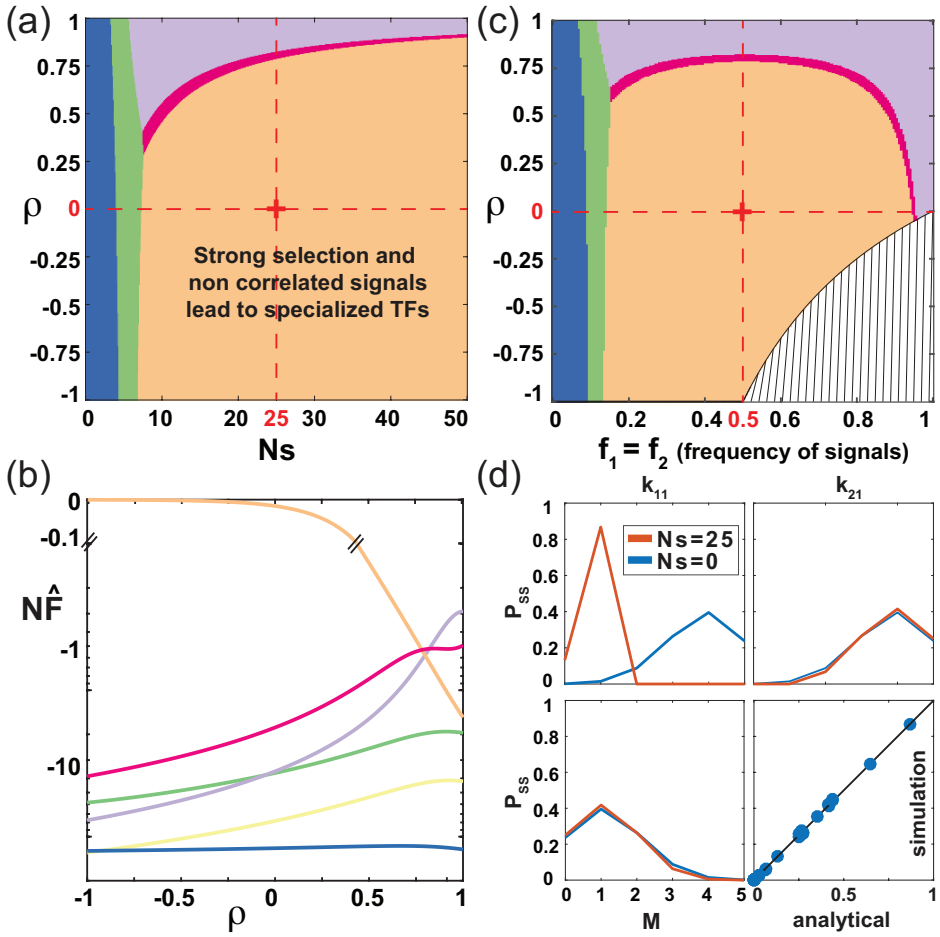


Figure 4.11: **Steady state evolutionary outcomes of TF duplication.** **(a)** Most probable outcome of gene duplication in steady state (color-coded as in (a)), as a function of selection strength, Ns, and the correlation between two external signals, ρ. **(b)** Free fitness $\hat{F}$ (at $Ns = 25$) for different macrostates as a function of correlation between signals, ρ: for most macrostates, free fitness increases with signal correlation, except for `No regulation`, which is naturally unaffected by it, and `Specialize Both`, which dominates for low correlation values. **(c)** The dominant macrostate (as in (a)), as a function of the signal frequencies, f_1, f_2, and the signal correlation, ρ, at fixed $Ns = 25$. For simplicity we plot only cases where $f_1 = f_2$. Signals in the hashed region are mathematically impossible. **(d)** Steady state distributions for mismatches ($P_{\mathrm{SS}}(k_{ij}|\sigma_1 = 10, \sigma_2 = 01)$, upper row) and the match between the two TF consensus sequences ($P_{\mathrm{SS}}(M|\sigma_1 = 10, \sigma_2 = 01)$, lower left), under strong selection (red; at baseline parameters denoted by the red cross in (a)) and neutrality (blue; Bernoulli distributions). Comparison between analytical calculation and 400 replicates of the stochastic simulation (lower right). Here and in subsequent figures, baseline parameter values are $L = 5$, $\epsilon = 3$, $r_{\mathrm{S}} = r_{\mathrm{TF}} = 1$.

specialized state, whereas high positive correlation favors losing one TF and having the remaining TF regulate both genes and respond to both signals. As signal correlation increases, so does the selection strength required to support full specialization. Detailed insight at a fixed value of Ns is provided by plotting the free fitness $\hat{F}$, as in Fig. 4.11b, which combines the fitness and the entropy of the neutral distribution from Fig. 4.10b into a single quantity that determines the likelihood of each macrostate given ρ; the macrostate with highest free fitness is shown as the most probable outcome in Fig. 4.11a for $Ns = 25$, but free fitness also allows us to see, quantitatively, how much more likely the dominant macrostate is relative to other outcomes. Fig. 4.11c examines the case where not only the correlation, ρ, but also the frequencies, f_1, f_2, of encountering both signals are varied: for low frequencies, even selection strength of $Ns = 25$ is insufficient to maintain TF specificity against drift, while for high frequencies and positive correlation one TF is lost while the remaining TF regulates both genes.

The map of evolutionary outcomes is very robust to parameter variations. The energy scale of TF-DNA interactions is that of hydrogen bonds: $\epsilon \sim 3$ (in k_BT units), consistent with direct measurements. The scale of C_0 is set to ensure that consensus sites are occupied at saturation while fully mismatching sites are essentially empty. The only remaining important biophysical parameter is L, the length of the binding sites. As expected, increasing L expands the regions of `No Regulation` and `Partial` at low Ns, due to entropic effects. Surprisingly, however, one can demonstrate that the important boundary between the `Specialize` and `One TF Lost` states is independent of L; furthermore, the map in Fig. 4.11a is exactly robust to the overall rescaling of the mutation rate, μ, and even to separate rescaling of individual rates $r_{\mathrm{S}}, r_{\mathrm{TF}}$.

We compare the steady-state marginal distributions of TF-BS mismatches and the match, M, between the two TFs, under strong selection to specialize ($Ns = 25$) vs neutral evolution ($Ns = 0$). Mismatch distributions for k_{11} and k_{21} in Fig. 4.11d display a clear difference in the two regimes: strong selection favors a small mismatch of the BS with the cognate TF, sufficient to ensure strong binding but nonzero due to entropy, and a large mismatch with the noncognate TF, to reduce crosstalk. Surprisingly, however, the distribution of matches M between two TF consensus sequences shows only a tiny signature of selection, with both distributions peaking around 1 match. As a consequence, inferring selection to specialize from measured binding preferences of real TFs might not be feasible with realistic amounts of data.

4.4 Asymmetric environments

At the baseline parameters, we assume symmetry between the occurrences of the two signals, namely their frequencies $f_1 = f_2 = 0.5$, where $f_1 = \alpha_{10}+\alpha_{11}$ is the frequency of the first signal, and $f_2 = \alpha_{01} + \alpha_{11}$ is the frequency of the second. In Fig. 4.11c, we explored the role of signal frequency f_i, together with signal correlation ρ, while maintaining symmetry ($f_1 = f_2$). Here we explore the effect of asymmetry in signal occurrence

($f_1 \neq f_2$) on the final evolutionary outcomes and in particular on the probability to fully specialize.

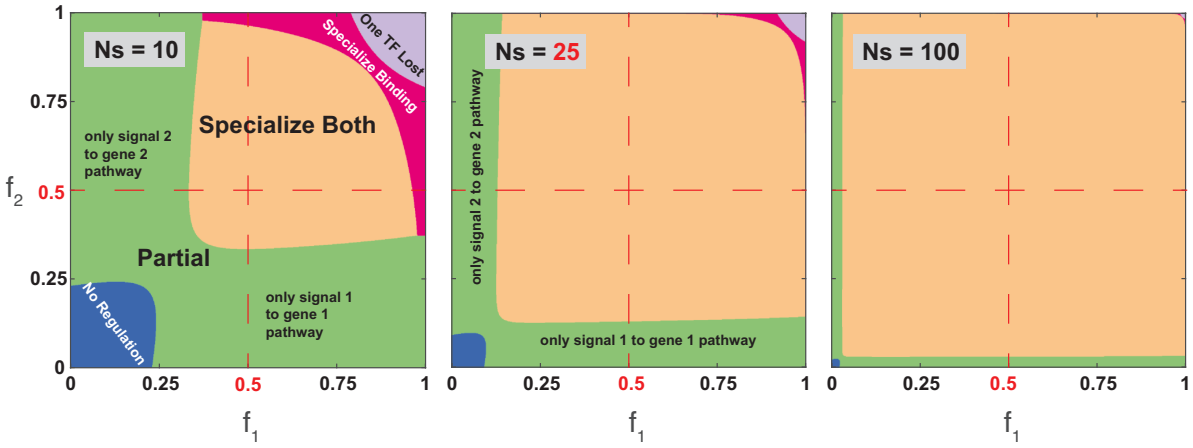


Figure 4.12: **Under medium to strong selection, specialization occurs under a broad range of signal frequencies. Under weak selection specialization occurs only if signal frequencies are sufficiently high.** Phase plots of the most probable macrostate at steady state as a function of signal frequencies f_1 and f_2, at three different selection strengths Ns = 10, 25, 100. The intersection between the red dashed lines, $f_1 = f_2 = 0.5$, denotes the baseline parameters used anywhere else in this work.

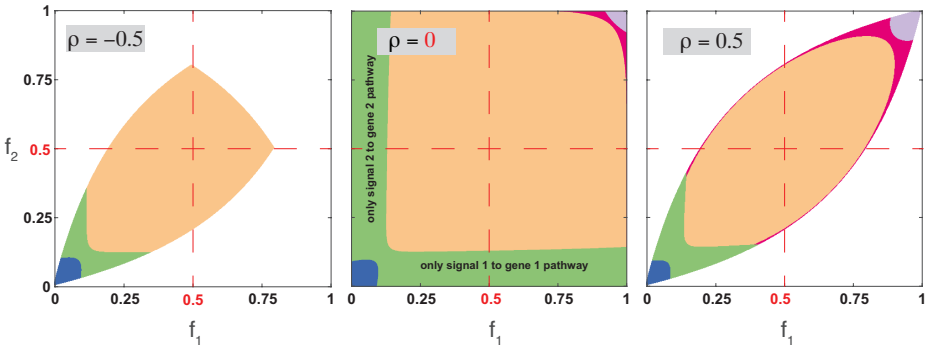


Figure 4.13: **For different ρ, f_1 and f_2 are constrained, but the phase plots in the accessible region are similar.** Phase plots of the most probable macrostate at steady state (at $Ns = 25$ and baseline parameters) as a function of signal frequencies f_1 and f_2, at three different signal correlations, ρ = -0.5, 0, 0.5. The white region of the plots denotes the forbidden areas; ρ, f_1 and f_2 are constrained and hence, not all (f_1, f_2) pairs are possible for different ρ.

In Fig. 4.12 we plot the most probable macrostate as a function of the signal frequencies f_1, f_2 for different values of selection intensities Ns when the signals are uncorrelated ($\rho = 0$); Fig. 4.13 shows that at different ρ, f_1 and f_2 are constrained but the qualitative features of the plots are retained. When both signals are rare, $f_1, f_2 \ll 1$, `No Regulation` macrostate dominates, as selection on both pathways is weak. When one of the signals is frequent while the other is rare, $f_1 \gg f_2$, only the frequently used pathway is maintained, and the

dominant macrostate is `Partial`. Only when both signals are frequent and selection is not too weak, specialization occurs. Hence, a signal-gene pathway is maintained only if it is required often enough, and the threshold for this (boundary between `Partial` and `Specialize Both`) depends on selection strength Ns. As selection strength Ns increases, this threshold moves to lower f_1 and f_2. As the frequencies of both signals increase, the dominant macrostate `Specialize Both` is replaced by `Specialize Binding`, where sensing one signal is a good proxy for the other signal as well, and later by `One TF Lost` when one TF is sufficient to transduce both signals.

4.5 TFs as repressors

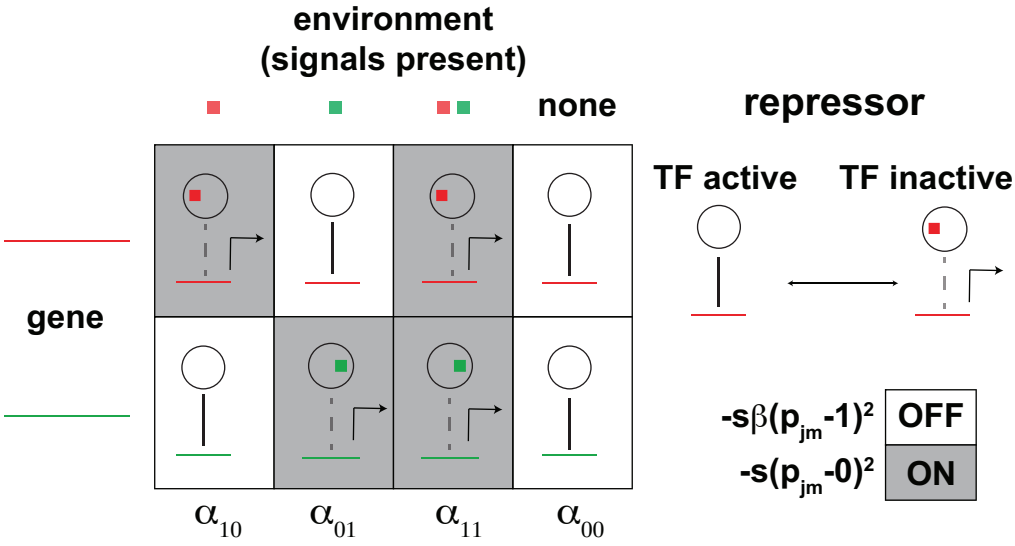


Figure 4.14: **Optimal expression patterns and fitness contributions in different environments with repressor TFs.** When TFs act as repressors, the scheme is different from when they act as activators. In the absence of a signal, they are in their active state, meaning they bind to their binding sites and thereby repress the corresponding genes. In the presence of a signal, the TFs become inactive, meaning they do not bind to their binding sites and thereby do not repress the corresponding genes. Shown are the optimal expression patterns of the two genes in the four different environments, and the mechanistic components of the genotype that can achieve these optimal patterns. Note that the environments in which repressor-BS binding is required are the environments in which the corresponding genes are required to be OFF. These are the terms in the fitness that correspond to crosstalk and hence have β_X, decreasing the effective selection to $\beta_X s$.

Here, we explore the scenario where TFs act as repressors. As described in Fig. 4.14, the primary difference is that repressor TFs, in the absence of a signal, are in their active state, meaning they bind to their binding sites and thereby, repress the corresponding genes. In the presence of a signal, the TFs become inactive, meaning they don't bind to

their binding sites and thereby, don't repress the corresponding genes. From the figure, notice that the effective selection pressure on repressor-BS binding is reduced to $\beta_X s$ because the environments in which repressor-BS binding is required are those in which the corresponding genes are required to be OFF.

In Fig. 4.15, we plot the dominant macrostate at steady state as a function of Ns and ρ at baseline parameters (with $\beta_X = 0.5$) when the TFs act as repressors. Notice that this is mostly similar to Fig. 4.11a, where TFs act as activators. One distinguishable feature is that the `No Regulation` to `Partial` to `Specialize Both` transition occurs at larger Ns values. In fact, with $\beta_X = 1$, when the selection pressures for repressor-BS binding are not diluted, these transitions occur at very similar Ns values as in the activators case.

In Fig. 4.16, we explore the role of signal frequency, f_i, on the dominant macrostate in the case of TFs acting as repressors. Note that this is a reflection, on the f_i axis, of the plot in the activators case (Fig. 4.11c). At low signal frequencies, $f_i \approx 0$, the genes are required to be OFF together most of the time, and hence, one repressor TF can regulate both the genes by always binding to their binding sites. This results in a dominance of the `One TF Lost` state. At high signal frequencies, $f_i \approx 1$, both genes are required to be ON together most of the time, and hence, repressor-BS binding occurs very rarely, thereby experiencing negligible selection pressure to maintain repressor-BS binding. Hence, the dominant state is that of `No Regulation`, where the repressor TFs don't bind to their binding sites, and hence, the genes are always ON.

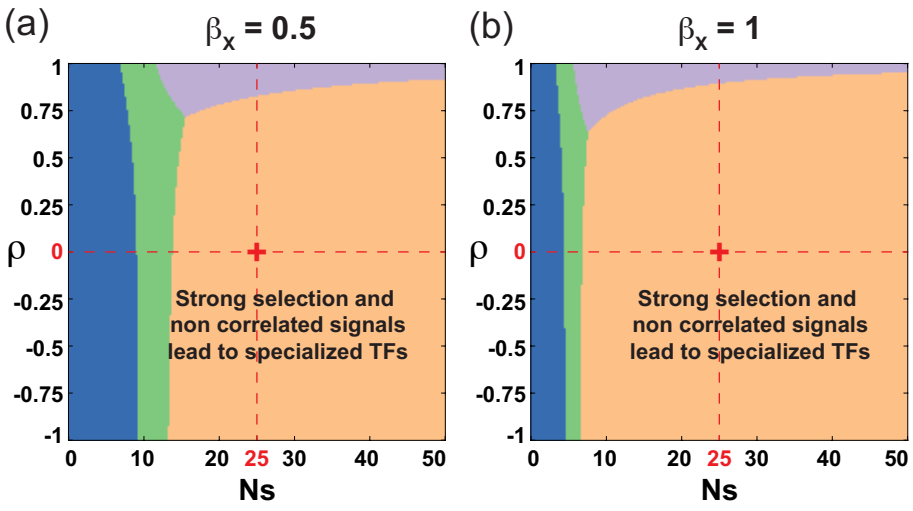


Figure 4.15: **Dominant macrostate phase plots vs Ns and ρ when TFs act as repressors.** Phase plots of the dominant macrostate against the selection strength, Ns, and the signal correlation, ρ, for **(a)** $\beta_X = 0.5$ and **(b)** $\beta_X = 1$.

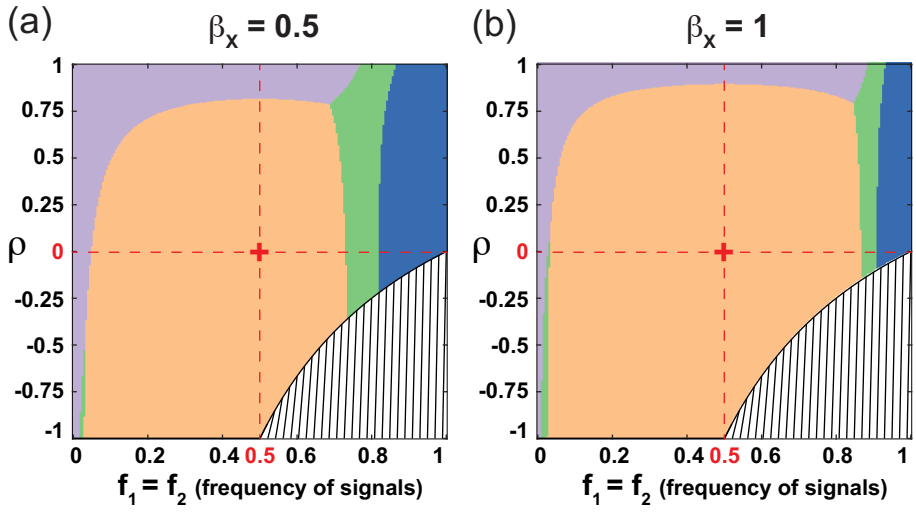


Figure 4.16: **Dominant macrostate phase plots vs ρ and $f_1 = f_2$ when TFs act as repressors.** Phase plots of the dominant macrostate against the signal correlation, ρ, and the signal frequency, $f_1 = f_2$ for **(a)** $\beta_X = 0.5$ and **(b)** $\beta_X = 1$.

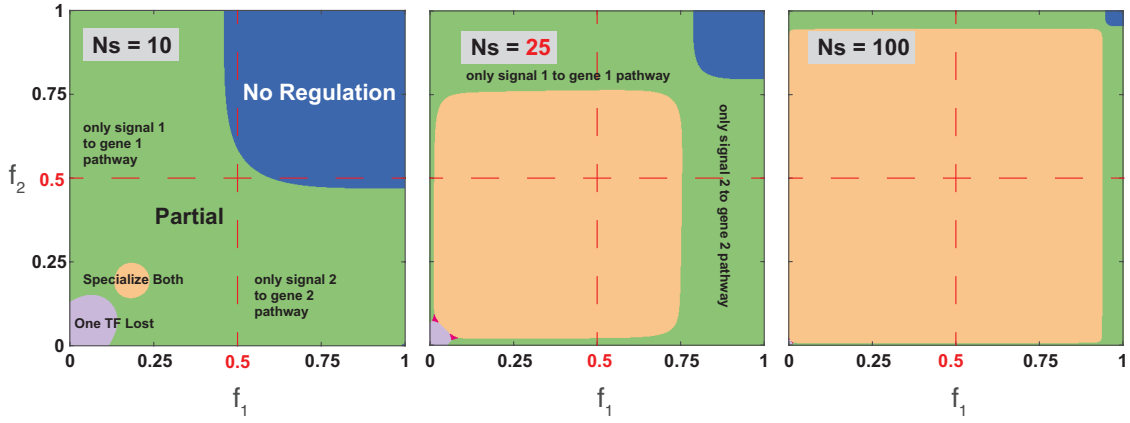


Figure 4.17: **Under medium to strong selection, specialization occurs under a broad range of signal frequencies. For repressor TFs, under weak selection specialization occurs only if signal frequencies are low.** Phase plots of the most probable macrostate at steady state (for $\rho = 0$) as a function of signal frequencies f_1 and f_2, at three different selection strengths $Ns = 10, 25, 100$ when TFs act as repressors. The intersection between the red dashed lines, $f_1 = f_2 = 0.5$, denotes the baseline parameters used anywhere else in this work.

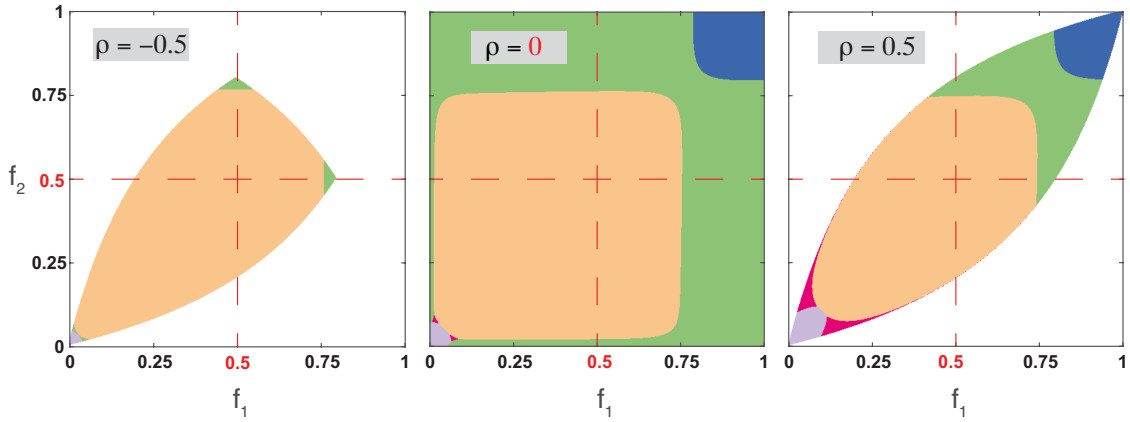


Figure 4.18: **For different ρ, f_1 and f_2 are constrained, but the phase plots in the accessible region are similar.** Phase plots of the most probable macrostate at steady state (at $Ns = 25$ and baseline parameters) as a function of signal frequencies f_1 and f_2, at three different signal correlations, ρ = -0.5, 0, 0.5. The white region of the plots denotes the forbidden areas; ρ, f_1 and f_2 are constrained and hence, not all (f_1, f_2) pairs are possible for different ρ.

In Fig. 4.17, we explore the effect of asymmetry in signal occurrence ($f_1 \neq f_2$) on the final evolutionary outcomes and in particular on the probability to fully specialize. We plot the most probable macrostate as a function of the signal frequencies f_1, f_2 for different values of selection intensities Ns when the signals are uncorrelated ($\rho = 0$); Fig. 4.18 shows that at different ρ, f_1 and f_2 are constrained but the qualitative features of the plots are retained. The principal difference from the activators case is that specialization now occurs at lower signal frequencies, with the `One TF Lost` state dominating at very low f_i, and `No Regulation` state domination at very high f_i.

4.6 Evolutionary dynamics

Next, we focus on evolutionary trajectories and the timescales to reach the fully specialized state after gene duplication.

To determine evolutionary dynamics we numerically integrate $P(\mathcal{G}, t)$ in time-steps corresponding to one generation t_g :

$$P(\mathcal{G}, t + t_g) = P(\mathcal{G}, t) + \mathbf{R} t_g P(\mathcal{G}, t), \tag{4.22}$$

where R is the Markov chain transition matrix. From $P(\mathcal{G}, t)$, we obtain the macrostate dynamics, $Q(\mathcal{M}, t)$. For every $z \in \mathcal{M}$,

$$Q(z, t) = \sum_{x \in S_z} P(x, t). \tag{4.23}$$

Again, at every time-point we determine the most probable macrostate, as illustrated in Fig. 4.21 as

$$z^*(t) := \arg\max_{z \in \mathcal{M}} Q(z, t). \tag{4.24}$$

To follow different pathways to specialization and the timescale to reach each, we calculate mean first hitting time $T_{S\leftarrow x}$ from any reduced-genotype x, to a subset of reduced-genotypes S, by using the recursive equation

$$T_{S\leftarrow x} = t_g + \sum_y a_{yx} T_{S\leftarrow y}, \tag{4.25}$$

where a_{yx} are the elements of the transition probability matrix $\mathbf{A} = \mathbf{I} + \mathbf{R}t_g$. In particular, we consider subsets S_z of genotypes that belong to a particular macrostate z, and compute the mean first hitting times, $T_{S_z\leftarrow x}$, to this macrostate. Time to specialization, τ, is the time to reach `Specialize Both` macrostate. Using a similar procedure, for every macrostate z, we also compute the dwell time, $t^{dwell}(z)$, which is the mean time to "escape" from that macrostate into any other macrostate z'. For every genotype x in S_z, the mean time to escape from S_z is by definition $T_{S'_z\leftarrow x}$, the mean time taken to hit $S'_z = \mathcal{G} - S_z$, the complementary set of S_z. We define the dwell time in macrostate z as

$$t^{dwell}(z) := \langle T_{S'_z\leftarrow x} \rangle_{x\in S_z}. \tag{4.26}$$

We supplement these analytical solutions by stochastic simulations. We use the Gillespie Stochastic Simulation Algorithm [Gillespie, 1976] to track the evolutionary trajectories of the system. Since we employ the fixed-state assumption, the time to fixation of each mutation is small compared to the waiting time between mutations and we neglect it in the calculations. At each simulation run we obtain a temporal series, $s_0, s_1, s_2, \ldots$, of genotypes (DNA sequences of TF consensus sequence and binding sites, along with signal sensing alleles), and a corresponding sequence of times, $t_0 = 0, t_1, t_2, \ldots$, at which substitutions between consecutive genotypes occurred. Here, s_0 is the initial DNA sequence with which we start the simulation. We construct s_0 by sampling a genotype from the steady state before duplication (with only 1 TF). For every i, from t_i to t_{i+1}, the DNA sequence of the system is s_i, from which there is a substitution event to s_{i+1} at t_{i+1}. We obtain s_{i+1} by appropriately sampling substitutions available from s_i, which can occur via TF consensus sequence mutations, or TF sensing domain mutations, or BS sequence mutations. We also draw $t_{i+1} - t_i$ (the waiting time) from the appropriate exponential distribution in the Gillespie framework. For each DNA sequence s_i, one can obtain the reduced-representation (M, k_{ij}, σ_i). From this, we obtain, for each simulation run r, the time trajectories of reduced-genotypes, $x_r(t)$, starting from $x_r(t = 0) = x_{r0}$. By running multiple times and computing the fractions of runs with each reduced-genotype x at each t, we obtain the dynamical trajectory of the probability distribution of reduced-genotypes, $P^{sim}(\mathcal{G}, t)$, and the steady state distribution, $P^{sim}_{SS}(\mathcal{G})$. Grouping the reduced-genotypes into macrostates, we also obtain the dynamical trajectory of the probability distribution of macrostates, $Q^{sim}(\mathcal{M}, t)$ and steady state distribution of macrostates, $Q^{sim}_{SS}(\mathcal{M})$.

The simulations enable us to compute non-trivial *path-dependent* quantities relating to an ensemble of trajectories $\{x_r(t)\}$, as well as to provide full distributions of quantities of interest. One such example is the mean hitting time to some macrostate z, conditioned on not hitting some other particular macrostate on the way. While it is possible in principle to compute such a path-dependent quantity exactly, in practice this requires too much numerical effort and Gillespie simulation becomes the method of choice.

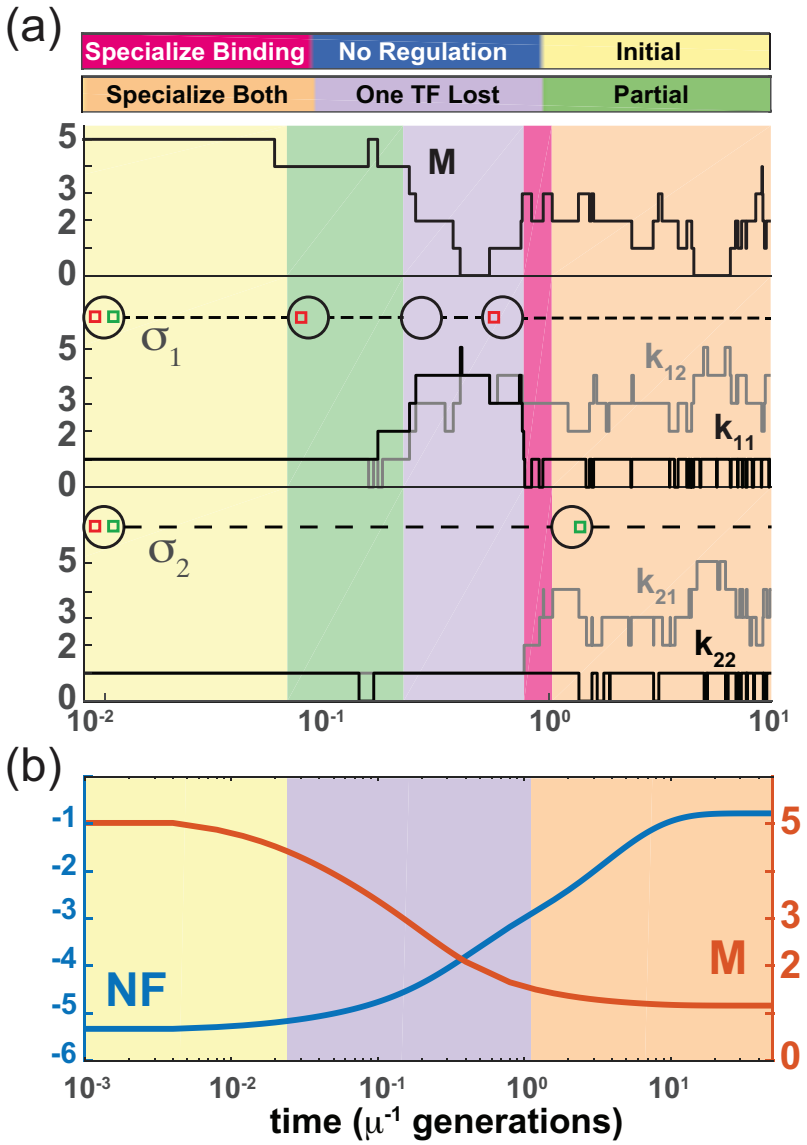


Figure 4.19: **Example trajectory.** **(a)** Temporal traces of TF-TF match M (top), and TF-BS mismatches k_{ij} (middle: TF1, bottom: TF2) with the corresponding signal specificity mutations denoted on dashed lines, for one example evolutionary trajectory at baseline parameters. Macrostates are color-coded as in the top legend and Fig. 4.11a. **(b)** Average dynamics of fitness NF (blue, left scale) and TF-TF match M (red, right scale). For every timepoint, the dominant macrostate is denoted in color.

An example trajectory is shown in Fig. 4.19a: the two TFs start off identical (with maximal match, $M = L = 5$) until, as a result of the loss of specificity for both signals, TF1 starts to drift, diverging from TF2 (sharply decreasing M in `One TF Lost` state) and losing interactions with both binding sites. Subsequently TF1 reacquires preference to the red signal, which drives the reestablishment of TF1 specificity for one binding site during a short `Specialize Binding` epoch, followed quickly by the specialization of TF2 for the green signal at the start of `Specialize Both` epoch of maximal fitness.

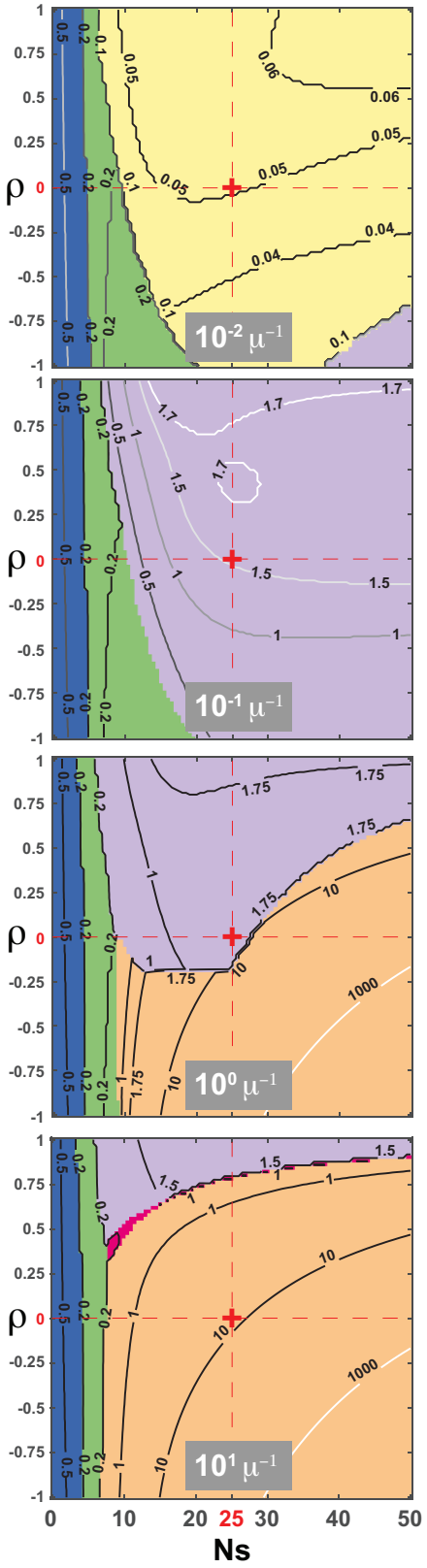


Figure 4.20: **Specialization is faster through the `Partial` state.** Snapshots of dominant macrostates (at increasing time post-duplication as indicated in the panels), shown for different combinations of selection strength Ns and signal correlation ρ as in Fig. 4.11. Contours mark dwell times in the dominant macrostates (in units of μ^{-1}). Red cross = baseline parameters.

Dynamics of the TF-TF match, M, and the scaled fitness, NF, become smooth and gradual when discrete transitions and the consequent large jumps in fitness are averaged over individual realizations, as in Fig. 4.19b. Importantly, we learn that the sequence of dominant macrostates leading towards the final (and steady) state, `Specialize Both`,

involves a long intermediate epoch when the system is in the `One TF Lost` state.

We examine this sequence of most likely macrostates in detail in Fig. 4.20, and visualize it analogously to the map of evolutionary outcomes in steady state shown in Fig. 4.11a. High Ns and correlation (ρ) values favor trajectories passing through the `One TF Lost` state, while intermediate Ns ($5 \lesssim Ns \lesssim 20$) and low correlation values enable transitions through `Partial` macrostate; along the latter trajectory, the binding of neither TF is completely abolished. Typical dwell times in dominant states, indicated as contours in Fig. 4.20, suggest that specialization via the `One TF Lost` state should be slower than through the `Partial` state, which is best seen at $t = 1/\mu$, where specialization has already occurred at intermediate Ns and low, but not high, ρ values.

It is easy to understand why pathways towards specialization via the `One TF Lost` state are slow. As the example in Fig. 4.19a illustrates, so long as one TF maintains binding to both sites and thus network function (especially when signals are strongly correlated), the other TF's specificity will be unconstrained to neutrally drift and lose binding to both sites, an outcome which is entropically highly favored. After the TF's sensory domain specializes, however, the binding has to re-evolve essentially from scratch in a process that is known to be slow [Tuğrul *et al.*, 2015] unless selection strength is very high. In contrast to this "Slow" pathway, the "Fast" pathway via the `Partial` state relies on sequential loss of "crosstalk" TF-BS interactions, with the divergence of TF consensus sequences followed in lock-step by mutations in cognate binding sites. Specifically, the likely intermediary of the fast pathway is a `Partial` configuration in which the first TF responds to both signals but only regulates one gene, whereas the second TF is already specialized for one signal, but still regulates both genes.

To calculate statistics over pathways, in each simulation run r, we calculate the time to specialization, and also record the dominant transient state. By running many simulations, we have a set of times to specialization that go predominantly via the fast pathway of `Partial` $\{\tau_{fast}\}$, and those via the slow pathway of `One TF Lost` $\{\tau_{slow}\}$. Using these, we obtain their means ($\bar{\tau}_{slow} = \langle \tau_{slow} \rangle$ and $\bar{\tau}_{fast} = \langle \tau_{fast} \rangle$); we also record the fraction of pathways proceeding via the slow and fast alternatives.

The fast and the slow pathways are summarized in Fig. 4.21a. A detailed analysis reveals how different biophysical and evolutionary parameters change the relative probability and the average duration of both pathways. In Fig. 4.22, we plot the average time to specialization via slow and fast pathways for various values of L, r_{TF} and r_S. The ratios of these times are plotted in Fig. 4.21b. Increasing the length, L, of the binding sites favours the slow pathway as well as drastically increases its duration, leading to very slow evolutionary dynamics. This is because of an increase in size of the neutral landscape; strikingly, increasing L does not lengthen the fast pathway through `Partial` states. Increasing the rate of TF-specificity-affecting mutations, r_{TF}, has a qualitatively similar effect, while increasing the mutation rate affecting the sensory domain, r_{S}, favors the fast pathway.

Indeed, in the limit when r_{S} is much larger than the other two mutation rates, the sensing

domain specializes almost instantaneously, making the complete loss of binding by either TF very deleterious and thus avoiding the `One TF Lost` state; the adaptation dynamics is initially rapid, with binding sites responding to diverging TF consensus sequences, and subsequently slow, when TF consensus sequences further minimize their match, M, in a nearly neutral process.

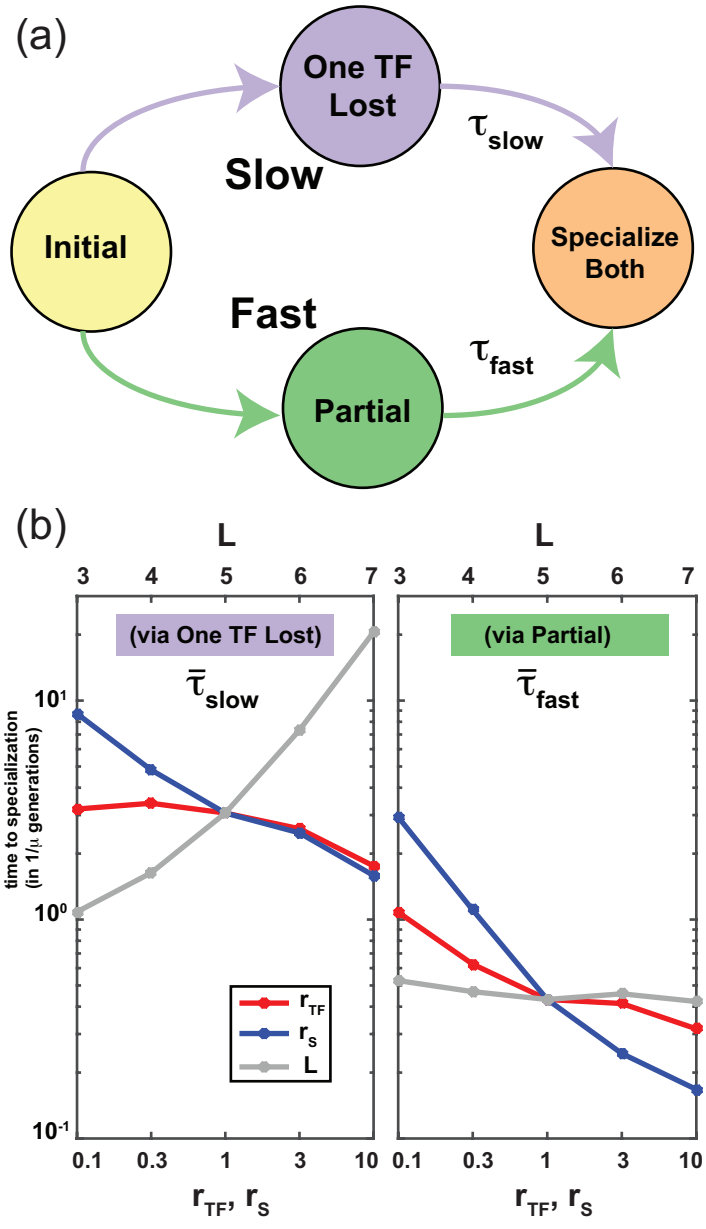


Figure 4.21: **Slow and fast pathways to TF specialization. (a)** Schematic of the two alternative pathways to specialization. τ_{slow} and τ_{fast} are the total times to specialization for the "slow" and the "fast" pathway, respectively. **(b)** We plot the mean times to specialization, $\bar{\tau}_{slow}$ and $\bar{\tau}_{fast}$, via the slow (left panel) and the fast (right panel) pathways, while varying L (grey curve, top axis), r_{TF} (red, bottom axis) and r_S (blue, bottom axis) separately. Other parameters remain at their baseline values. We find opposite dependence of the time to specialize on the binding site length L in the distinct pathways. While for pathways going via `One TF Lost` (left panel) time increases with L due to increase in the sequence space, it mildly decreases with L for pathways going via `Partial`. For all pathways specialization time decreases if mutation rates increase.

In Fig. 4.23 we detail the different pathways to specialization. The pathways proceeding via `One TF Lost` are slow compared to the pathways proceeding via `Partial` which are faster. The mutation initiating the process in all pathways is neutral and hence the ratio between r_S (signal sensing domain mutations rate) and r_{TF} (TF mutation rate) determines which pathway is more likely to occur - see Fig. 4.24.

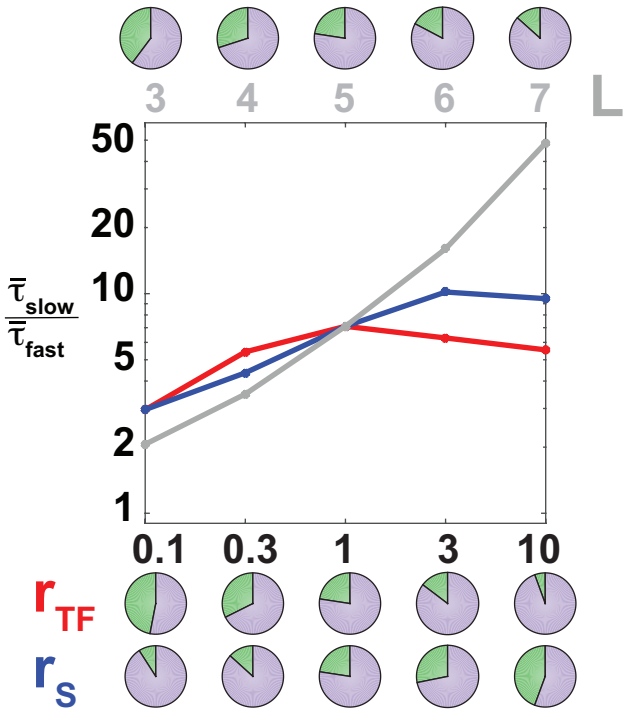


Figure 4.22: **Relative speed of specialization for different parameters.** Relative duration of the two pathways - slow via `One TF Lost` and fast via `Partial`, as a function of binding site length L (gray line, top axis), TF consensus sequence mutation rate r_{TF} (red), and signal domain mutation rate r_{S} (blue, bottom axis). Pie charts indicate the fraction of slow (pink) and fast (green) pathways at each parameter value.

Along the slow `One TF Lost` pathway, typically, first a TF consensus sequence mutation occurs that weakens the binding of one TF to both binding sites. Once binding is lost, further mutations cause the TF consensus sequence to neutrally drift away. Meanwhile, the lost TF gains a sensing mutation such that it senses only one of the two signals. Next, a BS mutation in one of the binding sites flips its TF preference such that the system moves into `Specialize Binding` macrostate. This is a beneficial mutation as one of the signal-BS pathways becomes specific. This involves evolving a TF-BS link essentially from scratch; the lost TF consensus sequence is a random number of mismatches away from the binding site sequence, and the beneficial BS mutation can occur only when the TF consensus sequence, by chance, becomes close enough to the BS sequence. From `Specialize Binding`, another beneficial sensing mutation leads the system to full specialization (BS and signal).

There are multiple routes in the `Partial` pathway. In one of the routes, first a neutral TF consensus sequence mutation occurs such that the TF loses binding to only one of the two binding sites resulting in `Partial` macrostate. This is different from the first mutation in `One TF Lost` pathway where the TF loses binding to both binding sites. From here, a sensing domain mutation specializes one of the signal-BS pathways, making this mutation beneficial. Further, a neutral BS mutation brings the system to `Specialize Binding`, from where a beneficial sensing domain mutation leads the system to specialization.

In the second and third routes via the `Partial` macrostate, first a neutral sensing domain mutation occurs. Next, either a beneficial TF consensus sequence mutation can bring the system onto the previous route or if the sensing domain mutation rate is high, another neutral sensing domain might occur first. From here, a beneficial TF consensus sequence mutation and a beneficial BS mutation again lead to full specialization.

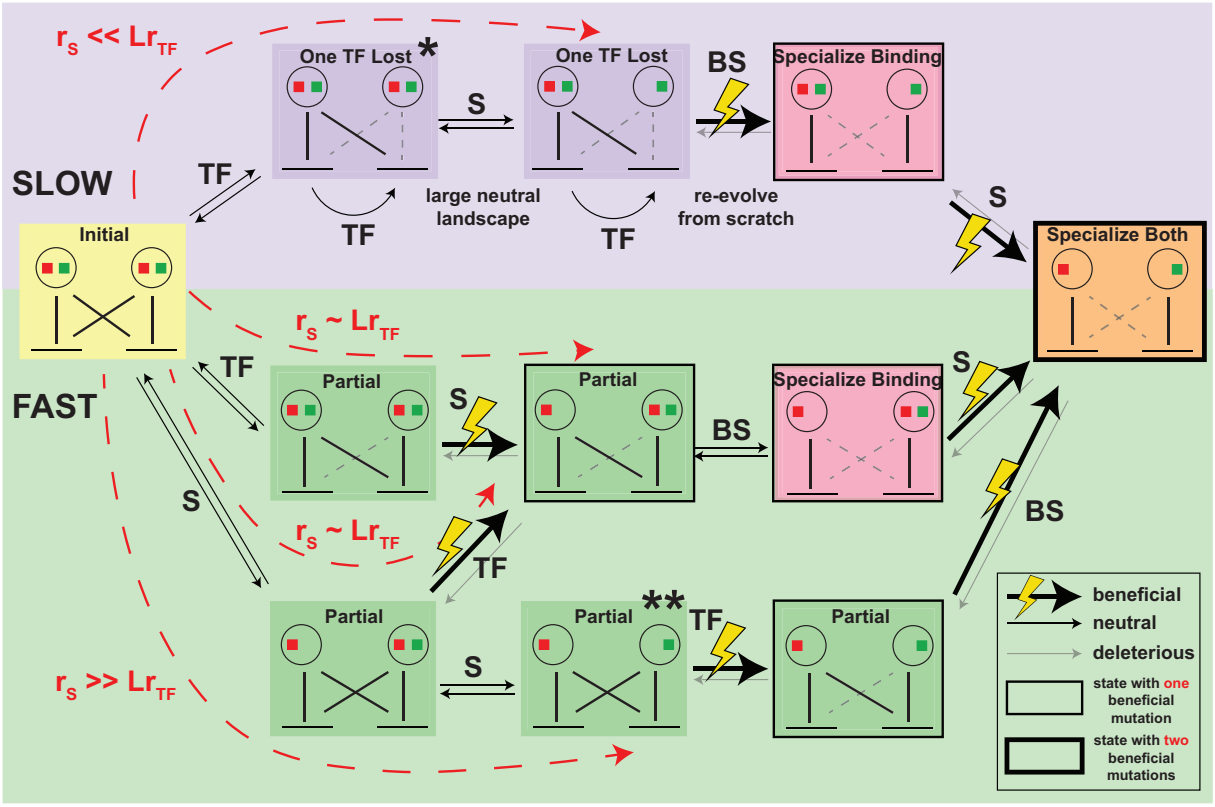


Figure 4.23: **Pathways to specialization differ in the order and nature of mutations.** Here we detail the various mutations occurring along the different pathways to specialization. For each mutation, we show the type of mutation (in text on the arrows): TF consensus sequence mutation (TF), binding site sequence mutation (BS), TF signal sensing domain mutation (S) and whether it is beneficial, (nearly) neutral or deleterious (style of the arrows). We also illustrate the macrostates along each pathway using the same color code in the background as in the main text. The number of beneficial mutations in each macrostate relative to the `Initial` macrostate is depicted by box style (see legend). Text in red indicates the conditions on mutation rates that favor the different pathways. Note that from the `One TF Lost` state marked with a star, the "lost" TF can actually take up new functions (by sensing and binding to signals and binding sites other than those considered in our model), leading to "neo-functionalization". Also, the `Partial` state marked with two stars acts as the initial condition in the alternative model variant, with the TFs already specialize in signal sensing immediately post-duplication.

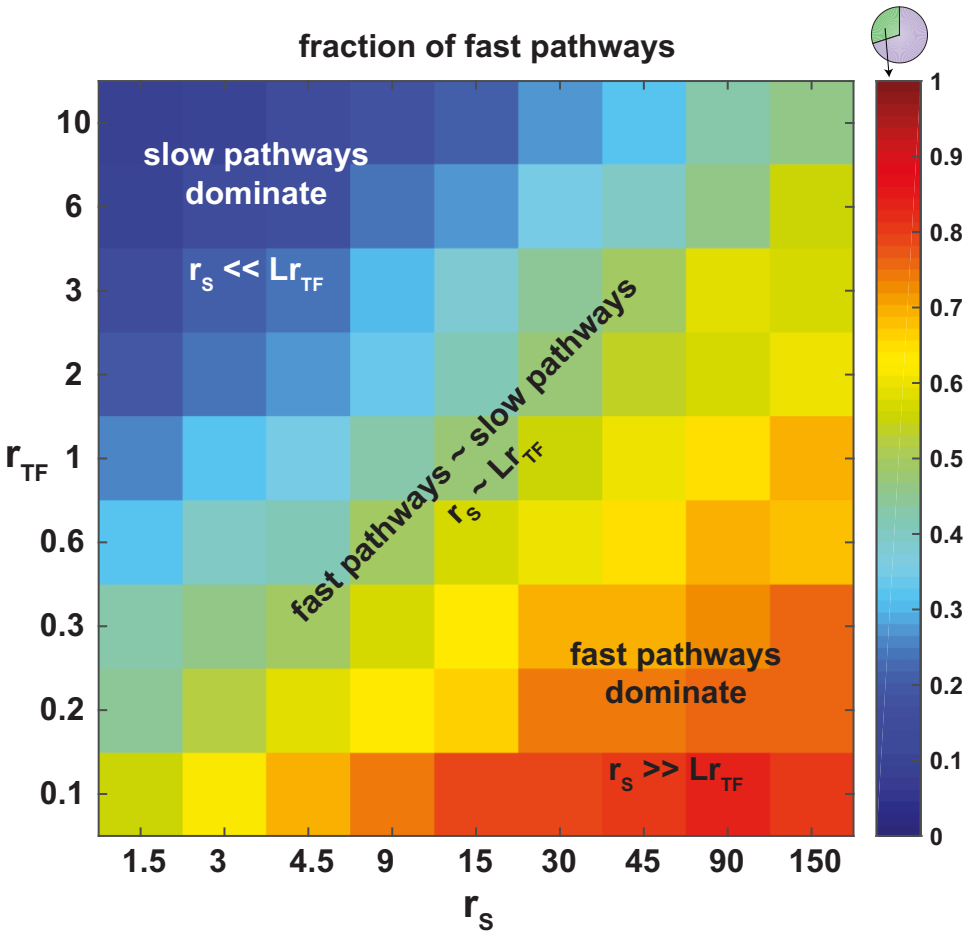


Figure 4.24: **The ratio between r_S and r_{TF} determines the dominant pathway.** We plot the fraction of fast `Partial` pathways as a function of r_S (signal sensing domain mutation rate) and r_{TF} (TF mutation rate). Other parameters remain at their baseline values (see Section 4.2). Color code denotes the fraction of fast pathways (specialization is reached via 'Partial' intermediate state).

4.7 Role of β_X, the relative fitness penalty on crosstalk

Transcription factors often bind weak secondary binding sites besides their primary target(s). This can lead to spurious activity of genes – crosstalk, i.e., deleterious activation of genes that does not happen via their primary regulatory pathway. For example, in our model a gene can be activated even if the signal to which it should respond is absent only because of (weak) binding of a transcription factor responding to another signal to its binding site. Previously, in Chapter 2, we studied the effect of crosstalk interference on gene regulation, and showed how it can place global constraints on the gene regulatory system [Friedlander *et al.*, 2016]. Here, we explore the potential role of such crosstalk interactions in shaping the evolutionary trajectories of TF specialization.

The fitness of each reduced-genotype $x \in \mathcal{G}$ depends on the difference between the actual expression pattern the genotype generates and the ideal expression pattern as defined in

Eq. (4.8).

$$F(x) = -s \sum_j \sum_m \alpha_m \beta_{jm} (p_{jm} - p^*_{jm})^2. \tag{4.27}$$

Here, β_{jm} weigh the penalties on different deviations from the desired expression level p^*_{jm}. In a certain environment m some genes should be active, $p^*_{jm} = 1$, while others should remain inactive, $p^*_{jm} = 0$. In our model, we allow for different penalties in either case. We penalize deviations from desired activity $p^*_{jm} = 1$ by setting $\beta_{jm} = 1$. We consider deviations from desired inactivity $p^*_{jm} = 0$ as less crucial and penalize them to a lesser extent $\beta_{jm} = \beta_X$, $\beta_X \in [0, 1]$. At the two extremes, if $\beta_X = 0$, no penalty on these crosstalk terms applies, while if $\beta_X = 1$, penalties on all deviations are equally important. So far, we used an intermediate value of $\beta_X = 0.5$. In this section we explore the role of β_X on the steady state distribution prior to and after TF duplication and on the evolutionary dynamics of specialization.

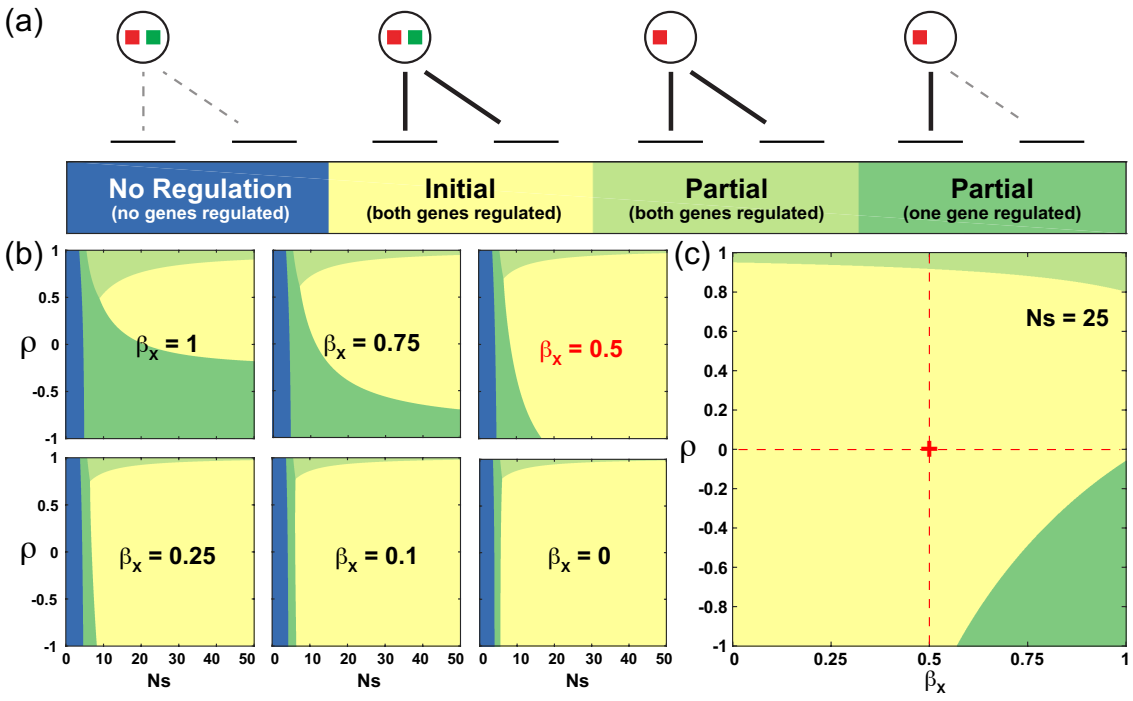


Figure 4.25: **Dominant macrostate at steady state before duplication depends on β_X (crosstalk interaction penalty). (a)** Illustration of the different macrostates when only a single TF exists. Macrostates before duplication are defined in terms of the macrostate they would result in, if a duplication occurred on those genotypes. **(b)** Most probable macrostate at steady state before duplication, as a function of selection strength, Ns, and the correlation between the two external signals, ρ, for different values of β_X, the relative weight of fitness penalties corresponding to crosstalk interactions. **(c)** The most probable macrostate at steady state before duplication, as a function of β_X and ρ at $Ns = 25$.

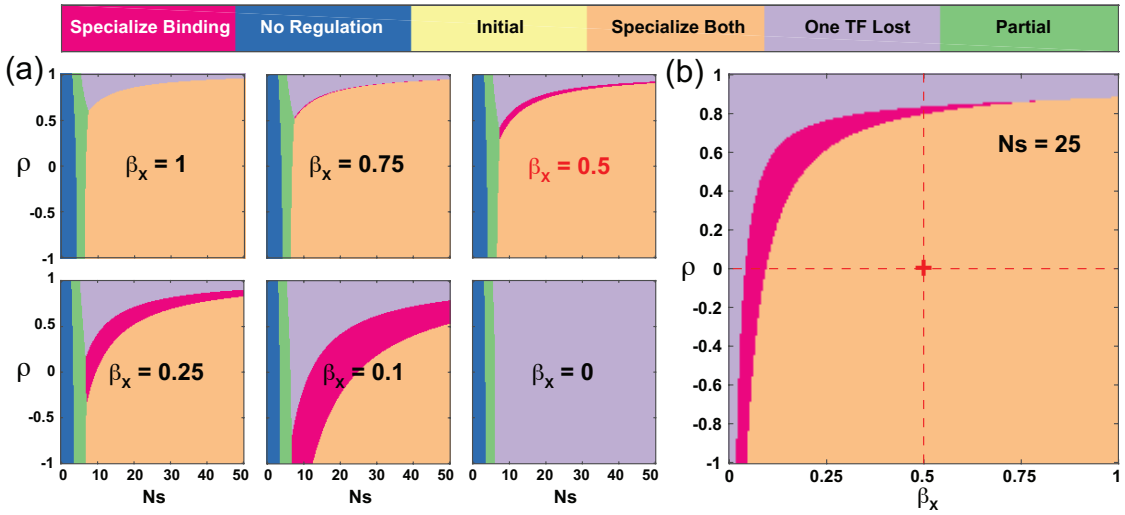


Figure 4.26: **Dependence of steady state after duplication on β_X, the fitness penalty on cross-interactions.** **(a)** The most probable macrostate at steady state after duplication, as a function of selection strength, Ns, and the correlation between the two external signals, ρ, is plotted for six different values of β_X. **(b)** The most probable macrostate at steady state after duplication, as a function of β_X and ρ at $Ns = 25$. An increase in β_X has a a similar effect to an increase in selection intensity on all interactions by varying Ns.

4.7.1 Steady state before duplication

A steady state distribution is attained before duplication, when only a single TF regulates all genes. In Fig. 4.25 we illustrate the most probable macrostate prior to duplication for different values of cross-interaction penalties β_X. The macrostates possible before duplication are `Initial` (both genes regulated), `No Regulation` (none regulated) and some (but not all) variants of `Partial` - see Fig. 4.25a for illustration. For $\beta_X \simeq 1$, the fitness penalty on mistakenly activating a gene is comparable to the fitness penalty on not fully inducing genes when needed, resulting in network configurations in which only one of the two genes is regulated (corresponding to `Partial` macrostate immediately after duplication for most $\rho < 0$). This is because, while configurations with only one gene regulated have one functional interaction and no crosstalk interactions, configurations with both genes regulated have two functional interactions and two crosstalk interactions. As β_X decreases, the selection against crosstalk interactions becomes weaker, resulting in configurations in which both genes are regulated (`Initial` macrostate immediately after duplication) even when $\rho < 0$.

4.7.2 Steady state after duplication

We proceed to observe the effect of varying β_X on the steady state after duplication, analogous to Fig. 3C of the main text where we assumed $\beta_X = 0.5$. In Fig. 4.26, we show the phase plot of the most probable outcome of duplication at steady state

for different values of β_X. The qualitative features of this phase plot are invariant to changes in β_X, as long as $\beta_X > 0$. For ρ not too close to 1, we obtain transitions from `No Regulation` to `Partial` and to `Specialize Both` as Ns increases. For large enough Ns, as ρ increases, there is a shift from `Specialize Both` to `One TF Lost`, via `Specialize Binding`, the width of which increases as β_X decreases. This is because there is reduced selection pressure on avoiding crosstalk interactions as β_X decreases. For small β_X, as ρ increases, it is sufficient that one of the TFs senses both signals while the TFs are still specialized in binding. As ρ increases even further, it is sufficient to have one TF mediating both pathways, marking the shift to the `One TF Lost` macrostate. These transitions occur very prominently for very small $\beta_X \approx 0$, where `One TF Lost` is the most probable outcome for all ρ values. Many models of duplication do not consider crosstalk interactions in their fitness function, and hence deal with the case of $\beta_X = 0$, making it important for comparison to our results. This insight that selection against crosstalk is crucial to TF specialization can be extended to the realistic case of many regulated genes by extending fitness to include positive selection for correct regulation of each gene in its corresponding environment, and negative selection against all signal-gene crosstalk. We leave a full treatment to future research.

4.7.3 Evolutionary dynamics

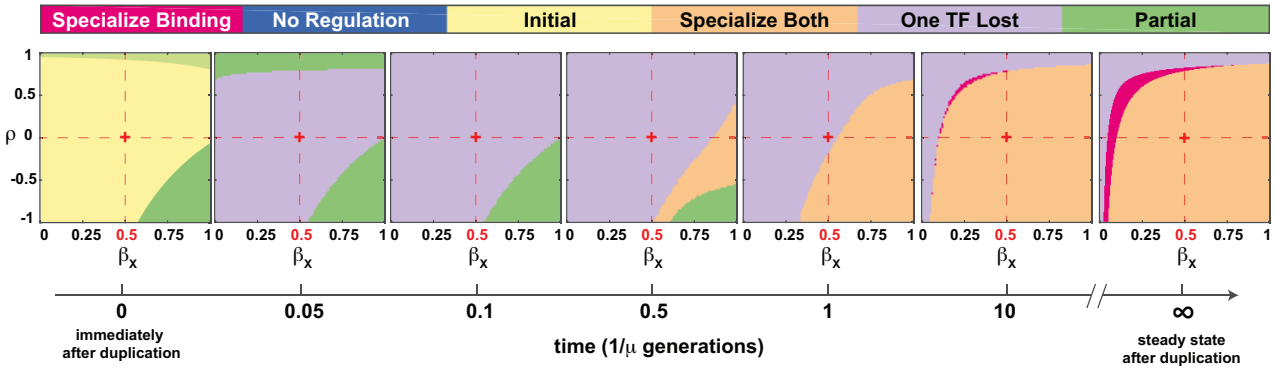


Figure 4.27: **Snapshots of the most probable macrostate at different time-points post-duplication.** The most probable macrostate as a function of signal correlation, ρ, and β_X, the relative weight of fitness penalties corresponding to crosstalk errors, for $Ns = 25$. The left-most phase plot corresponds to the time-point immediately after duplication, and the right-most phase plot corresponds to the steady state after duplication. For other parameters, the baseline values have been used. $\beta_X = 1$ corresponds to equal-magnitude selection strengths on functional as well as crosstalk interactions; $\beta_X = 0$ corresponds to no selection against crosstalk interactions. We choose $\beta_X = 0.5$ as the baseline parameter value.

To understand how β_X affects the evolutionary dynamics of specialization, we first obtained the dynamics of the most probable macrostate as a function of ρ and β_X for fixed selection intensity $Ns = 25$ (baseline parameters). In Fig. 4.27, we plot a few snapshots

of the phase diagram of the most probable macrostate at different time-points after duplication, starting from $t = 0$ (immediately after duplication), to $t = \infty$ (steady state after duplication). Specialization is faster for smaller ρ because the fitness benefit of eliminating crosstalk interactions is larger. Likewise, specialization is faster for larger β_X as the selection strength against crosstalk interactions is higher. A huge region of the (β_X, ρ) plane corresponding to small β_X or large ρ, most of which starts at `Initial` and specializes via the slow pathway of `One TF Lost`.

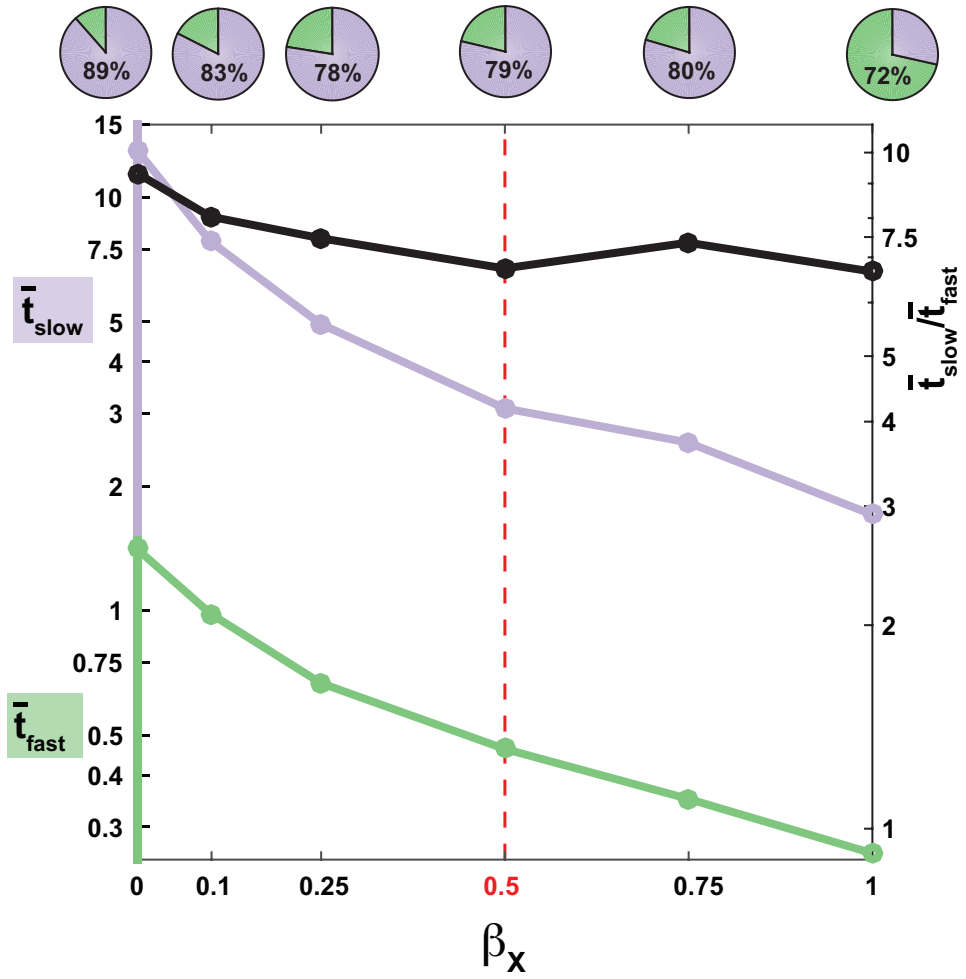


Figure 4.28: **Specialization speeds up as β_X increases.** For large β_X, the time to specialization shortens for all pathways and the fraction of trajectories to specialization taken via fast pathways (through `Partial` macrostate) increases. Pie charts illustrate the fraction of slow (lavender) and fast (green) trajectories for different values of β_X. The black line (right y-axis) shows the ratio between average specialization times, which does not significantly change with β_X. For other parameters, the baseline values were used. $\beta_X = 1$ corresponds to equal-magnitude selection strengths on functional as well as crosstalk interactions; $\beta_X = 0$ corresponds to no selection against crosstalk interactions. We choose $\beta_X = 0.5$ as the baseline parameter value.

Next we sought to understand which pathways are taken towards specialization for different β_X by running many repeats of simulations at each β_X. For each β_X, we found

the most probable genotype at steady state before duplication and ran many repeats of the simulation starting from that genotype. In Fig. 4.28, we explore the dependence on β_X of fraction of the two pathways to specialization (slow via `One TF Lost` and fast via `Partial`), and also the corresponding times to specialization. First of all, specialization becomes quicker as β_X increases from 0 to 1. This is because stronger selection against the crosstalk interactions eliminates them faster. Secondly, the relative speed of the fast pathway (compared to the slow pathway) depends only very weakly on β_X. Thirdly, about 80% of trajectories follow the slow pathway, and this depends only very weakly on β_X, till $\beta_X = 0.75$. In contrast, for $\beta_X = 1$, the fast pathways via `Partial` become predominant. This occurs because the steady state before duplication (which acts as the initial condition for the trajectories) flips from `Initial` to `Partial`.

4.8 Comparison with biallelic model

The gene duplication literature often studies models with a small number of discrete alleles, for example, binary alleles informing whether TF-BS binding occurs. Throughout this work we employ a different approach by including a biophysical description of TF/DNA interactions. Consequently, a large number of different genotypes can often realize each functional architecture (macrostate), capturing naturally the important effects of neutral processes (mutational entropy). Our framework reduces to biallelic models at $L = 1$ and alphabet size $D = 2$ (and multiallelic version with $D = 4$), so we can directly study the relationship between the results for a biophysically realistic fitness landscape and various common simplifications. We refer to these simpler models with $L = 1$ here as the biallelic-like model. The biallelic-like model cannot reproduce some of the results obtained with the biophysically-realistic model of the main text. In particular, certain important macrostates do not exist in the biallelic-like model. We also find an opposite dependence on time to specialization for the different pathways (`One TF Lost` vs. `Partial`). In Fig. 4.29 we plot the dominant macrostate at steady state for two values of D. For $D = 4$ (right panel of the figure), many qualitative features are retained from the more realistic main text model: for instance, the change from `No Regulation` to `Partial` to `Specialize Both` as Ns increases, and the change from `Specialize Both` to `Specialize Binding` to `One TF Lost` as ρ increases. For $D = 2$, we have `Partial` macrostate dominating at $Ns = 0$, because its entropy is larger than that of the `No Regulation` macrostate. Also, at large Ns and large ρ, `Partial` dominates via the genotypes in which all TF-BS links are strong but the signal sensing domain is not specialized.

Certain variants of `Partial` that exist in the general model do not exist in the biallelic-like model, as shown in Fig. 4.30. These states have intermediate fitness and they arise in the fast `Partial` pathway of the main text model, where they form a bridge between the `Initial` and the `Specialize Both` macrostates. Hence, in biallelic models, fast `Partial` pathways do not exist and instead, passing through `Partial` entails either losing a BS or specializing very fast in the signal sensing domain. These states have low fitness in

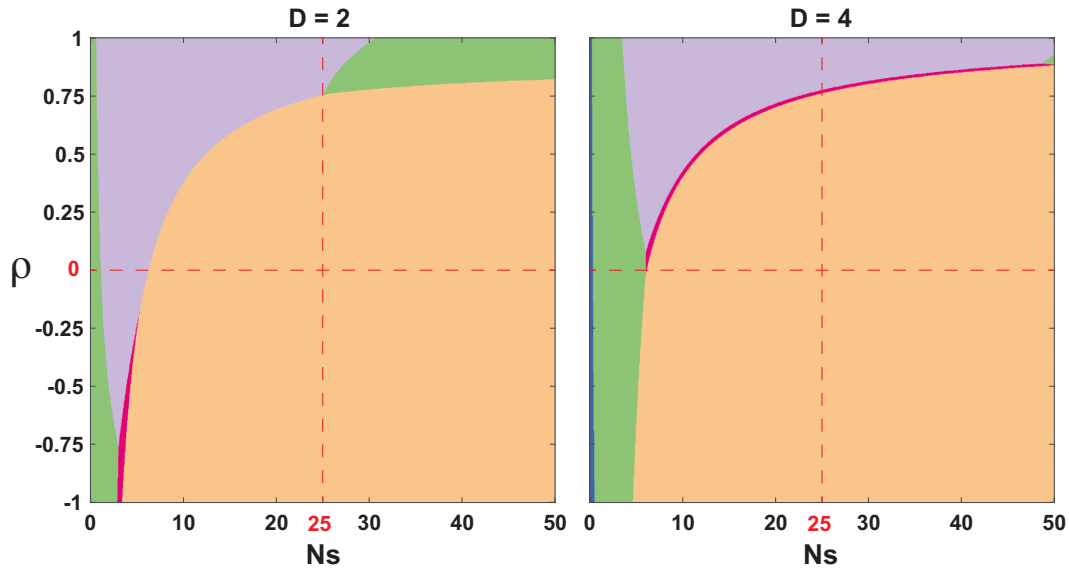


Figure 4.29: **Dominant macrostate at steady state for biallelic-like models.** Here we plot the dominant macrostate at steady state as a function of Ns and ρ for biallelic-like models with alphabet size $D = 2$ (left panel) and $D = 4$ (right panel). Color code used to indicate different macrostates is the same as in the main text.

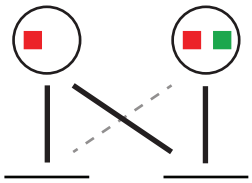


Figure 4.30: **This type of `Partial` macrostate is absent in biallelic-like models.** In biallelic-like models, strong TF-BS link means an exact match between TF and BS. Hence, the description of `Partial` states of the kind shown here is impossible.

the biallelic-like model and hence `Partial` pathway is actually slow. This is plotted in Fig. 4.31.

In summary, biallelic-like models and the biophysically realistic model share a few similarities but also differ in certain important aspects. Biallelic-like models, while being very simplistic, still capture a few key qualitative features of the steady state distribution, for example, the transitions of dominant macrostates along the ρ and Ns axes. On the other hand, biallelic-like models paint a completely different picture of evolutionary dynamics and timescales. Because they do not consider intermediate-fitness `Partial` states, unlike in the biophysically realistic model, time to specialization through `Partial` becomes slower than through `One TF Lost`.

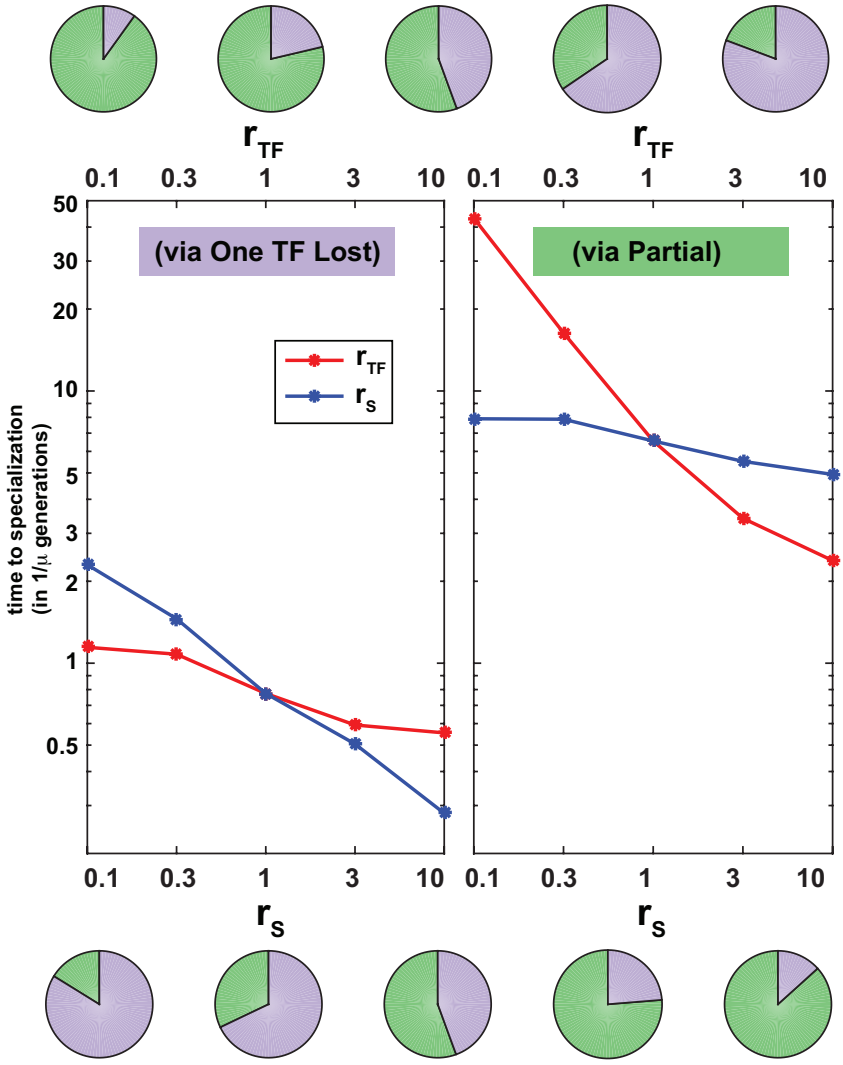


Figure 4.31: **Biallelic-like models reverse the relation between different pathways to specialization: `Partial` pathways are the slow ones and `One TF Lost` pathways are faster, in contrast to the full model.** We plot the times to specialization via `One TF Lost` (left panel) and via `Partial` (right panel), at $Ns = 100$, while changing r_{TF} (red curve) and r_S (blue curve) separately, keeping the other parameters at their baseline values in each case. We also show the fraction of these pathways as pie charts (upper pie charts refer to different r_{TF} values; lower ones to different r_S values).

4.9 Multiple target genes per TF

Typically, each TF must regulate more than one target gene. As the number of regulated genes per TF (n_G/n_{TF}) increases, intuition suggests that the evolution of the TF's consensus sequence should become more and more constrained: while a mutation in an individual binding site can lower the total fitness by increasing mismatch and thereby impeding TF-BS binding, a single mutation in the TF's consensus has the ability to simultaneously weaken the interaction with many binding sites, leading to a high fitness

penalty. Here, we analyze the biophysical fitness landscape in the presence of multiple target genes per TF, and confirm that the landscape gets progressively more frustrated as the number of regulated genes per TF increases. This is due to the explosion of constraints that TFs have to satisfy to ensure the maintenance of functional regulation, and consequently, result in extremely long times to specialization. How can it nevertheless proceed at observable rates? We provide a possible answer to this in Section 4.10.

The steady state distribution in the case of multiple target genes is

$$P(M, \{k_{ij}\}, \{\sigma_i\}) = P_0(M, \{k_{ij}\}) P_0(\{\sigma_i\}) \exp(2NF), \tag{4.28}$$

where P_0 is the neutral distribution and F is the fitness of the reduced-genotype. First, we need to account for the neutral distribution P_0 (entropic factor). This is straightforward, because for given TF consensus sequences, the probability that a particular binding site j has mismatch values (k_{1j}, k_{2j}) is independent of the state of other binding sites. Thus, we can simply factor out the probabilities for different genes:

$$P_0(M, \{k_{ij}\}, \{\sigma_i\}) = P_0(\{\sigma_i\}) P_0(M) \prod_j P_0(k_{1j}, k_{2j}|M), \tag{4.29}$$

where j enumerates the genes. Second, we need to take care of the adaptive (energy) factor $\exp(2NF)$ in the general case. Because $F = \sum_j F_j$ is linear in terms of contributions F_j from each gene j, $\exp(2NF)$ factorizes into $\prod_j \exp(2NF_j)$. Hence, we have

$$P(M, \{k_{ij}\}, \{\sigma_i\}) = P_0(M) P_0(\{\sigma_i\}) \prod_j P_0(k_{1j}, k_{2j}|M) \exp(2NF_j). \tag{4.30}$$

Now, for $\langle M \rangle$, we have,

$$\begin{aligned} \langle M \rangle &= \sum_{\{k_{ij}\}, M, \{\sigma_i\}} M P(M, \{k_{ij}\}, \{\sigma_i\}) \\ &= \sum_{\{\sigma_i\}} P_0(\{\sigma_i\}) \sum_M M P_0(M) \prod_j \sum_{k_{1j}, k_{2j}} P_0(k_{1j}, k_{2j}|M) \exp(2NF_j) \\ &= \sum_{\{\sigma_i\}} P_0(\{\sigma_i\}) \sum_M M P_0(M) \prod_j \langle \exp(2NF_j) \rangle_{P_0(\{k_{ij}\}|M)}. \end{aligned} \tag{4.31}$$

$\langle \exp(2NF_j) \rangle_{P_0(\{k_{ij}\}|M)}$ can be calculated for each gene j separately.

Evolutionary pathways

The pathways to specialization in the case of multiple regulated genes are more complex than those described in Section 4.6 for $n_G = 2$. As each TF needs to simultaneously regulate a subset of the genes while avoiding regulation of the remaining ones, the number of constraints are increased relative to the $n_G = 2$ case, and incur a diminishing number of feasible evolutionary trajectories. The fitness change due to a TF consensus sequence mutation is assessed according to its effect on the binding affinities of this TF with all existing genes. Hence, for each TF, as n_G increases, the number of constraints

also increases. This limits the number of possible substitutions a TF can access via fewer beneficial and neutral mutations. In contrast, for each binding site, the number of constraints does not change because it is only constrained by the two TFs and not by other binding sites. The following are the main pathways that are depicted in Fig. 4.32. The first proceeds via `One TF Lost` macrostate while the other pathways proceed only via `Partial` configurations.

1. The first pathway involves the `One TF Lost` macrostate, where as before one TF does not bind to any binding site. Evolving a TF-BS link to this TF entails a random walk on a neutral landscape and essentially involves regulatory evolution from scratch. After gaining a TF-BS link from a BS mutation, the system ends up on a local fitness plateau (marked with a red box in Fig. 4.32) in the `Partial` state. This is because the "lost" TF (second TF in the figure) has considerably diverged from the first TF yet has specialized only for some, but not all, of the genes associated with the green signal, but not for all of them. All of the TFs and BSs are constrained to maintain match beyond some minimal level. Hence specialization can only occur if one of the strong TF-BS links weakens. Such weakening decreases the fitness, and hence incurs crossing a fitness valley. This pathway is consequently very slow.

2. The remaining pathways do not involve `One TF Lost` macrostate and go only via `Partial` macrostate. In the second pathway, first, a TF consensus sequence mutation and a signal sensing mutation (either can occur first) lead the system to a `Partial` state with some of the signal-BS pathways specialized. Then, an additional TF consensus sequence mutation pushes the TFs further apart. This, together with BS mutations, brings the system to the local fitness plateau (in the `Partial` macrostate) described in the previous pathway. This pathway is also slow, because of the fitness valley crossing described above.

3. In the third pathway also, first, a TF consensus sequence mutation and a signal sensing mutation (either can occur first) lead the system to a `Partial` state with some of the signal-BS pathways specialized. From here, no additional TF consensus sequence mutations occur that push the TFs away. Hence, there are paths for the BSs to realign their binding preferences (to the other TF) such that fitness is always maintained and does not involving crossing any fitness valleys. Hence, this pathway is fast.

4. In the fourth and the fifth pathways, the first two mutations are signal sensing mutations that specialize the TFs' signal sensing domains. From here, a TF mutation and subsequent BS mutations can specialize without going through fitness valleys. Hence, this is a fast pathway. For a given genotype (specifying the TF and BS sequences), this fourth pathway is either possible or not. If it is not possible, then the only resort is the fifth pathway.

5. The fifth pathway comes into play when the fourth pathway is not possible. This happens when any TF mutation loses some signal-BS pathways, hence dropping

the fitness considerably. The TFs cannot diverge at all, and this involves crossing a fitness valley. Hence, this is a slow pathway.

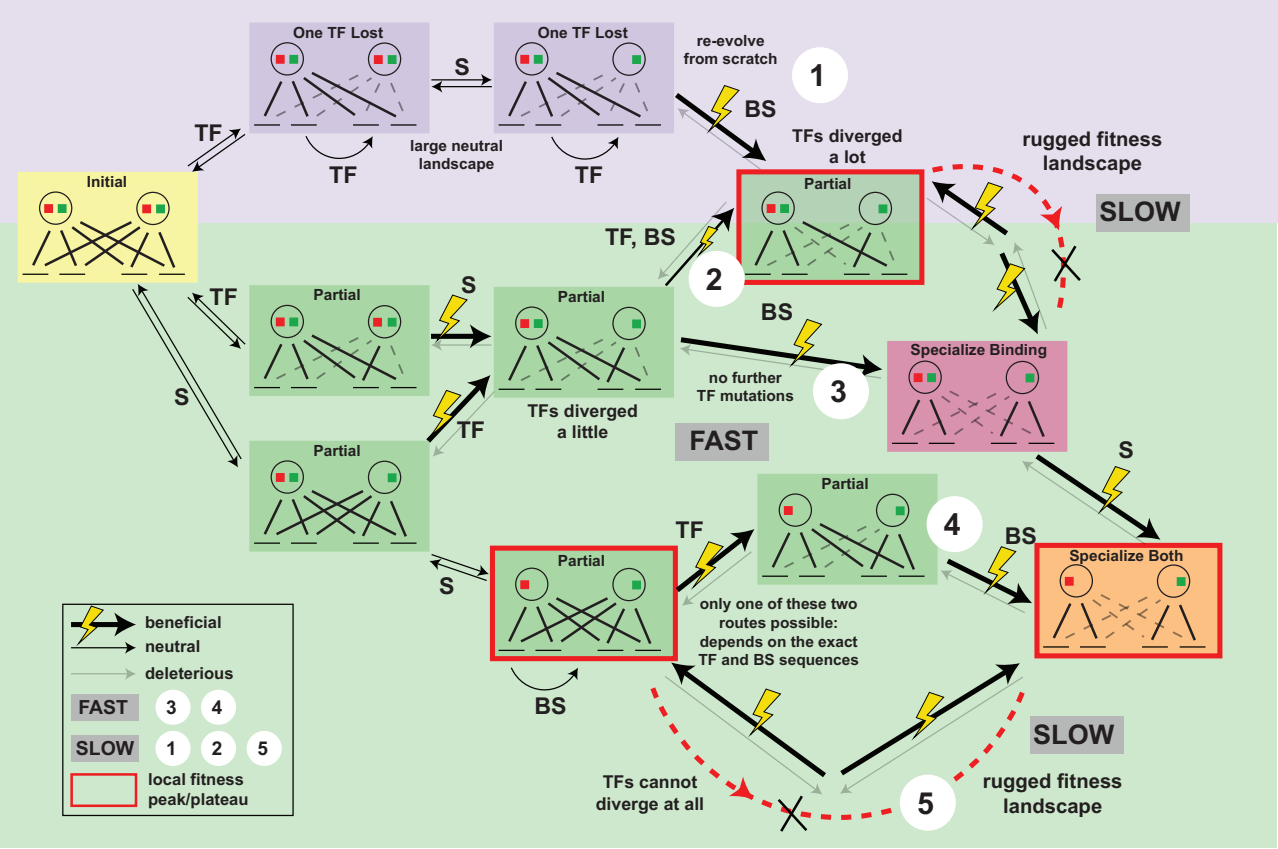


Figure 4.32: **Different pathways to specialization vary in the order and nature of mutations, and might have to cross a rugged fitness landscape for $n_G > 2$.** Here we show in detail the various mutations that occur along the different pathways (marked with numbers inside white circles) to specialization. For each mutation, we show the type of mutation (text on the arrows): TF consensus sequence mutation (TF) or binding site sequence mutation (BS), TF signal sensing domain mutation (S) and whether it is beneficial or (nearly) neutral or deleterious (style of the arrows, see legend). We also depict the macrostates along each pathway graphically, and mark local fitness peaks/plateaus with red boxes. In red dotted curved lines, we denote parts of the pathways which involve a fitness valley and hence, are very difficult to cross. Routes not involving any fitness valleys (numbered 3 and 4) are fast, while those involving a fitness valley (numbered $1, 2$ and 5) are slow. Populations often take the slower route, slowing their overall mean time to specialization.

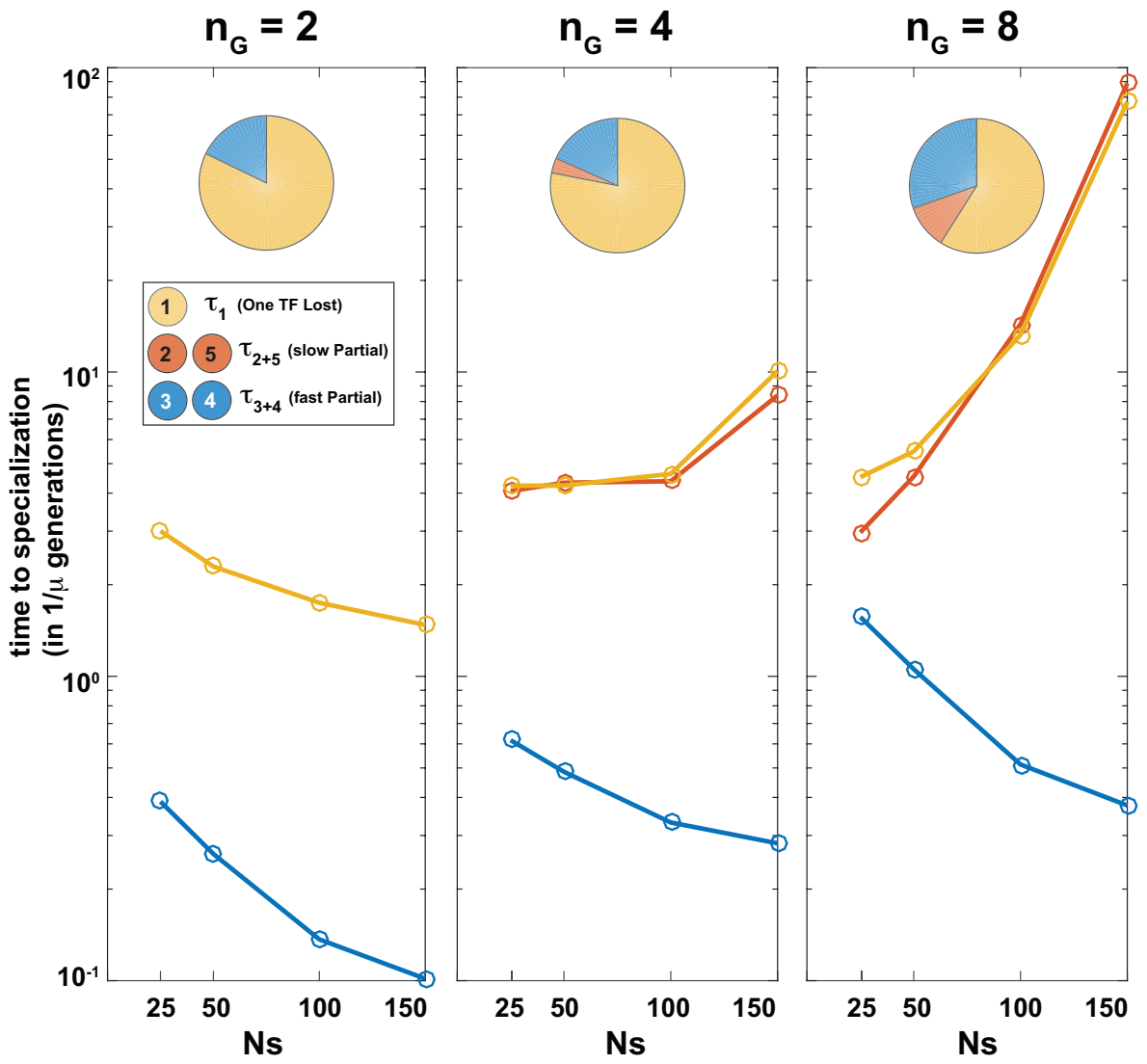


Figure 4.33: **Times to specialization via different pathways for various numbers of downstream genes.** Shown are the times to specialization via different pathways as a function of Ns for different values of n_G. We plot the times for the slow `One TF Lost` pathway (numbered 1, yellow), the slow `Partial` pathway (numbered 2 and 5, red), and the fast `Partial` pathway (numbered 3 and 4, blue). Plotted as pie charts also are the fraction of various pathways for different n_G values as pie charts; these fractions depend only very weakly on Ns. In general, the higher the n_G, the larger the fraction of fast trajectories (3 and 4) and the longer the time needed to specialize. Pathways whose time lengths with Ns, which are the slow `Partial` pathway (red) and the `One TF Lost` pathway (yellow) for $n_G > 2$, involve crossing fitness barriers.

Time to specialization

By running simulations, we calculate the time to specialization for different values of $n_G > 2$ (total number of downstream genes) via the different pathways described in the previous section. Specifically, we calculate the time to specialization, τ_1, via the

`One TF Lost` pathway (pathway 1), τ_{3+4}, via the fast `Partial` pathways (pathways 3 and 4), and, τ_{2+5}, via the slow `Partial` pathways (pathways 2 and 5). We also calculate the fractions of these pathways. These are shown in Fig. 4.33. The slow `Partial` pathway (numbered 2 and 5) is absent for $n_G = 2$. The fast `Partial` pathway (numbered 3 and 4) does not involve crossing any fitness valleys, and hence the time to specialization via this pathway decreases with increasing Ns for all n_G. The time to specialization via the slow `One TF Lost` pathway (numbered 1) decreases with increasing Ns for $n_G = 2$, and so does not involve crossing fitness valleys. For $n_G > 2$, the time to specialization via both the slow `One TF Lost` pathway and the slow `Partial` pathway increases as Ns increases. Both these pathways for $n_G > 2$ involve crossing fitness valleys. With increasing n_G, the fractions of the fast `Partial` pathway and slow `Partial` pathway increase at the expense of the slow `One TF Lost` pathway.

4.10 Promiscuity-promoting mutations

So far we considered the constant mismatch penalty model for TF-BS specificity, where each position in the TF and the binding site contributed equally to the total binding energy, depending on whether the position has a mismatch between the TF consensus sequence and the BS sequence. Let the TF consensus sequence be s^* and the binding site sequence be s, both of length L. In general, we have

$$E = \sum_i E_i, \tag{4.32}$$

where i runs over all the positions of the binding site. For each specific position i, the contribution is $E_i = 0$ if $s_i = s_i^*$ (match) and $E_i = \epsilon$ if $s_i \neq s_i^*$ (mismatch).

Experiments on TF-BS specificity, however, suggest that some TF (and binding site) positions dominate while others only have minor energetic contributions. In this section we study a simple generalization of the mismatch-energy model, where we allow for two levels of contribution: some positions are specific (favor a unique nucleotide) and have large energetic contribution while others are non-specific or promiscuous (all nucleotides are equally favorable) and have a smaller energetic contribution. For each specific position i, the contribution E_i is, as in the mismatch-energy model, ϵ if there is mismatch between the TF consensus sequence and the BS sequence in that position, and 0 if there is a match. On the other hand, for each promiscuous position i, the contribution is $E_i = \epsilon_P$ (typically $0 \leqslant \epsilon_P \leqslant \epsilon$), independent of s_i. Hence, for a TF with $L_P < L$ promiscuous positions in total, and k mismatches in the remaining $L - L_P$ specific positions, the total binding energy would be $E = \epsilon_P L_P + k\epsilon$. The different possible energy levels for specific and promiscuous TFs are illustrated in Fig. 4.34.

Promiscuity entails a cost in terms of TF-BS binding. To elucidate this cost, we consider the dependency of the free (dimensionless) concentration, C_0, of a TF, on the binding preferences of the TF.

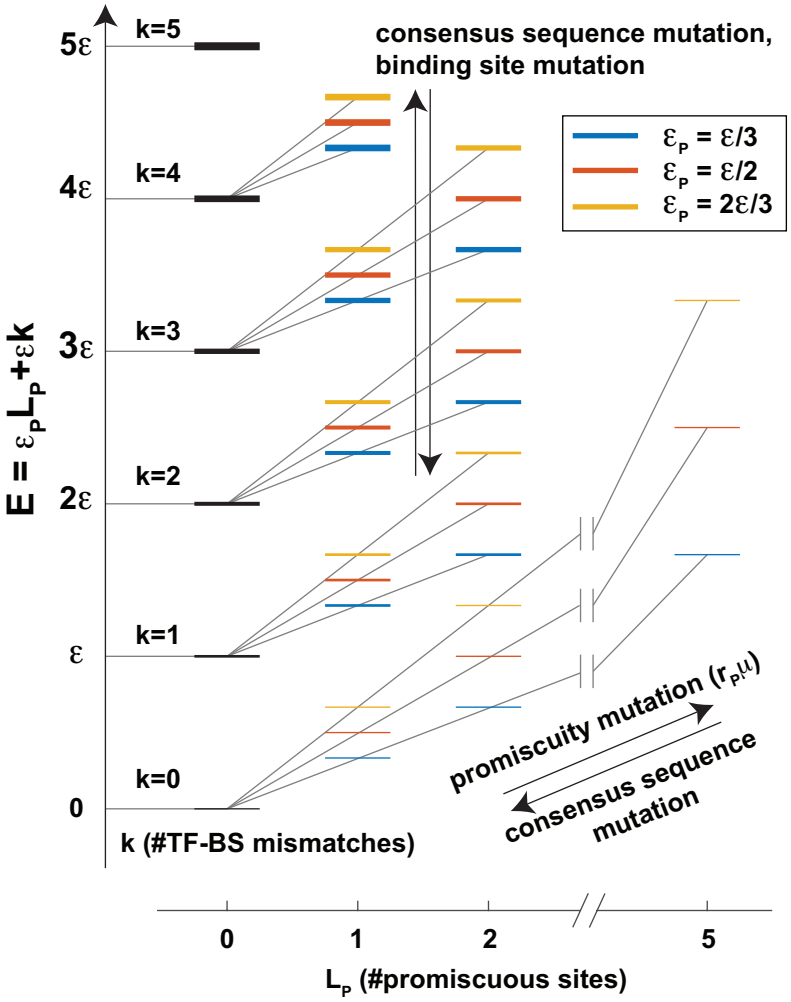


Figure 4.34: **Total TF-DNA binding energies depend on number of mismatches as well as on the number of promiscuous TF positions.** We plot the different energy levels depicting the TF-BS binding energy, $E = \epsilon_P L_P + \epsilon k$, for TFs with varying number of promiscuous positions L_P and k mismatches between the TF and BS in the remaining $L - L_P$ specific positions. Note that lower E corresponds to tighter TF-BS binding. We illustrate this for three different values of ϵ_P, the energy contribution per promiscuous position (different colors). Increasing line thickness of the energy levels represents higher mismatch values k. While promiscuity-promoting mutations increase L_P by converting a specific position to a promiscuous one, regular TF mutations that hit a promiscuous position can convert it to be specific and decrease L_P.

For a TF with no promiscuous positions, C_0 can be calculated in the grand canonical ensembl framework of Chapter 1 as

$$C_0(L_P = 0) = \frac{C}{GS(\epsilon, L) + \sum_n \exp(-E_n)}, \tag{4.33}$$

where C is the copy number of the TF, G is the number of sites on the DNA where the TF can bind in a sequence-specific manner, n enumerates other possible energy configurations

of the TF that are sequence-independent (residing in the free solution, or nonspecific binding to DNA), and E_n is the free energy in configuration n. $S(\epsilon, L) = \langle e^{-\epsilon k} \rangle_{P(k)}$ is the similarity between binding sites defined in Chapter 2, with $GS(\epsilon, L)$ acting as the Boltzmann factor for all possible specific binding configurations. This term captures the sequestration of TFs on the DNA due to spurious binding. Assuming that the DNA sequence is random, $P(k) \sim B(L, 3/4)$ is the Binomial distribution for the number of mismatches that a random DNA sequence has with a given TF consensus sequence.

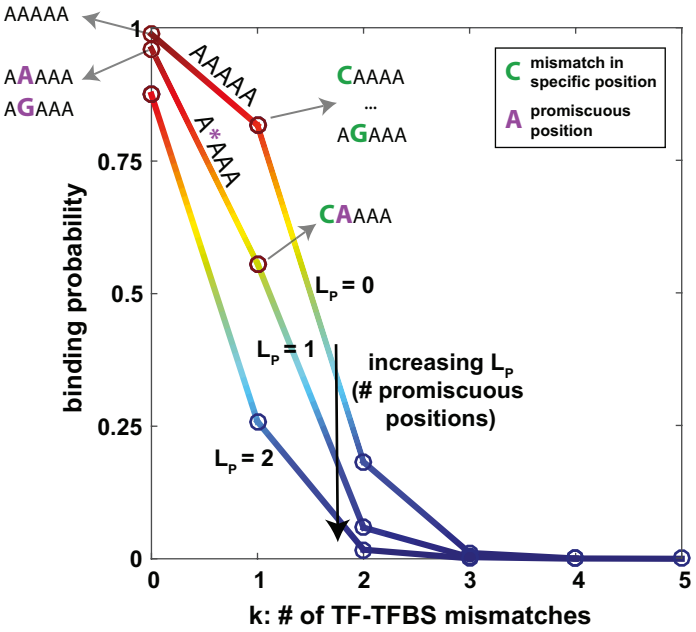


Figure 4.35: **Binding probability of the TF to DNA decreases the more promiscuous it is.** The TF-BS binding probability is plotted as a function of the number of TF-BS mismatches k among the $L - L_P$ specific positions for different values of L_P, the number of promiscuous positions in the TF. We list, as an example, different sequences that are consistent with given (L_P, k).

For a promiscuous TF with L_P promiscuous positions, we have,

$$\begin{aligned} C_0(L_P) &= \frac{C}{Ge^{-\epsilon_P L_P} S(\epsilon, L - L_P) + \sum_n \exp(-E_n)} \\ &= C_0(L_P = 0) \frac{GS(\epsilon, L) + \sum_n \exp(-E_n)}{Ge^{-\epsilon_P L_P} S(\epsilon, L - L_P) + \sum_n \exp(-E_n)} \\ &= C_0(L_P = 0) \frac{1 + A}{e^{-\epsilon_P L_P} \frac{S(\epsilon, L - L_P)}{S(\epsilon, L_P)} + A}, \end{aligned} \tag{4.34}$$

where $A = \frac{\sum_n \exp(-E_n)}{GS(\epsilon, L)}$ is an effective parameter that captures the relative contribution of the Boltzmann factor corresponding to spurious specific binding on the DNA, compared with all other Boltzmann factors. We have assumed that $A = 0.1$ is fixed in our calculations, and the results we present are fairly robust to the value of A. The probability that

a binding site is bound by a TF with $L_P > 0$ promiscuous positions and k mismatches with respect to the binding site in the remaining $L - L_P$ positions, assuming no other TF type is present, is

$$p = \frac{C_0(L_P)e^{-\epsilon k - \epsilon_P L_P}}{1 + C_0(L_P)e^{-\epsilon k - \epsilon_P L_P}}. \tag{4.35}$$

This probability is plotted in Fig. 4.35 for various k and L_p values. While $C_0(L_P)$ can be greater or lesser than $C_0(L_P = 0)$ depending on the value of ϵ_P, we have $C_0(L_P)e^{-\epsilon_P L_P} < C_0(L_P = 0)$. Hence, as the number of promiscuous positions, L_P, in the TF increases, the binding probability decreases.

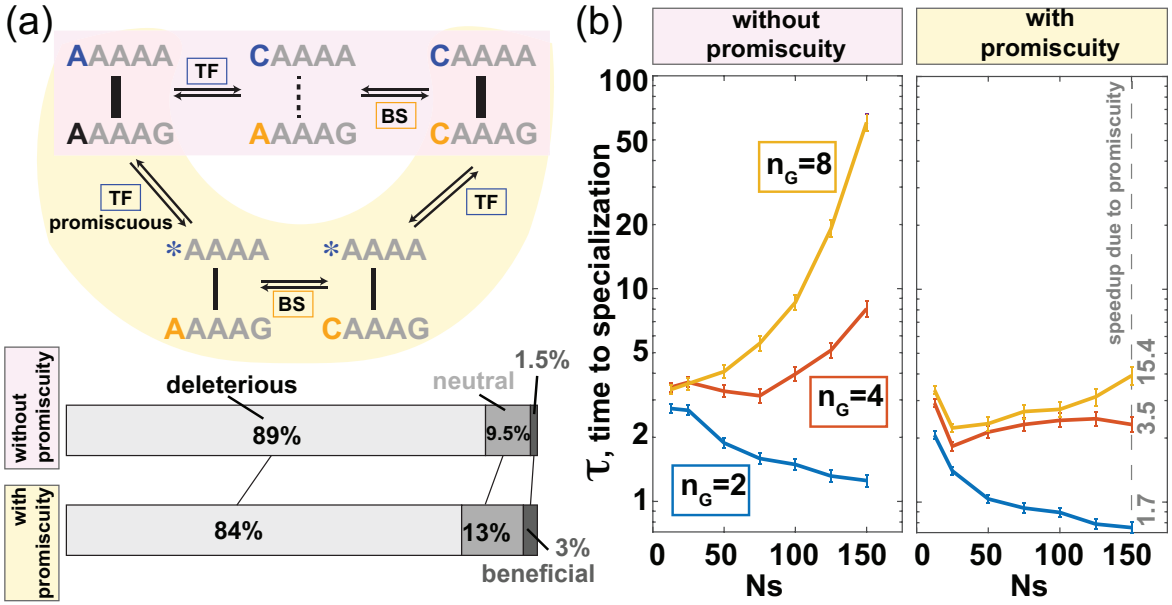


Figure 4.36: **Promiscuity-promoting mutations speed up specialization with multiple regulated genes per TF. (a)** In the absence of promiscuity-promoting mutations, a compensatory series of point mutations in the TF's consensus (upper sequence) and its binding site (lower sequence) is needed to maintain TF-BS specificity (top; light red). Alternatively, in the presence of promiscuity-promoting mutations in the TF consensus, a position in the TF's recognition sequence (marked by a star) can lose and later regain sequence specificity (middle; light yellow). Promiscuity decreases the fraction of deleterious mutations along typical pathways to specialization (bottom, computed using baseline parameters). **(b)** Time to specialization as a function of selection strength, Ns, without (left) and with (right) promiscuity promoting mutations in the TF, for different numbers of regulated genes per TF, n_G (color). Numbers in gray (right) denote the speedup ratio.

For instance, consider a TF with consensus sequence $AAAAA$ (see Fig. 4.35). This TF is specific for A's in all five positions of the binding site sequence. Each mismatch in the binding site sequence (green positions in the sequences in Fig. 4.35) with respect to $AAAAA$ decreases the binding affinity, and thereby decreases the binding probability. Now consider a promiscuous TF with consensus sequence $A * AAA$, where $*$ denotes a

promiscuous position. The second position, independent of the bp in the BS sequence (purple positions in the sequences in Fig. 4.35), decreases the binding affinity, but by a lesser amount than a specific position mismatch (green positions). Hence, the binding probabilities of the promiscuous TF to $AAAAA$, $AGAAA$, $ATAAA$ or $ACAAA$ are equal, and higher than the binding probability of the specific TF to $CAAAA$ or $AGAAA$ or other single-mismatch BS sequences.

We also introduce an additional type of mutation, called "promiscuity-promoting" mutation, that occurs at rate $r_P\mu$. These mutations convert a specific TF position in the consensus sequence to a promiscuous one. A promiscuous position can return to be specific again if it is hit by a consensus TF mutation (regular TF mutations we considered until now, happening at rate $r_{TF}\mu$). Fig. 4.36a shows how TF consensus sequence and the corresponding binding site can co-evolve using point mutations, or using the new "promiscuity-promoting" mutation type for the TF: promiscuity-promoting mutation renders one position in the recognition sequence of the TF insensitive to the corresponding DNA base in the binding site.

4.10.1 Steady state after duplication

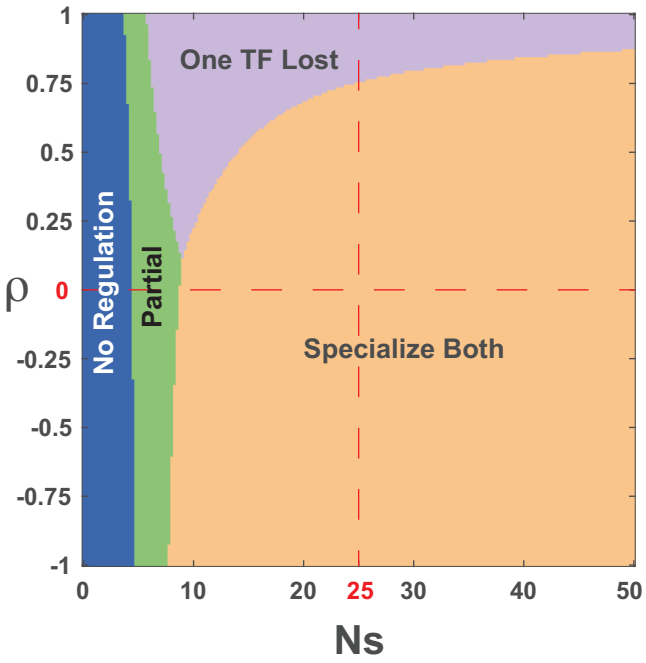


Figure 4.37: **Most probable macrostate in the presense of promiscuity-promoting mutations.** We plot the most probable macrostate at steady state, z^*_{SS}, for different ρ and Ns, for $n_G = 2$ and relative mutation rate $r_P = 3$, keeping other parameters at their baseline values. We choose $r_P = 3$ so that at each position, a specific bp has equal effective mutation rate towards a promiscuous state or another specific bp.

In the presence of promiscuity-promoting mutations, we obtain the steady state distribution over the genotypic space analytically, from which we obtain the dominant macrostate

at steady state for different ρ and Ns values (Fig. 4.37). The inclusion of promiscuity-promoting mutations does not significantly change the dominant macrostate phase plot except for a slight increase in the range of `One TF Lost` macrostate.

We also plot the mean number of promiscuous positions at steady state in Fig. 4.38. This number decreases with selection intensity, because promiscuous positions decrease the TF binding probability (see Fig. 4.35) making them less favorable once specialization has occurred.

4.10.2 Evolutionary dynamics

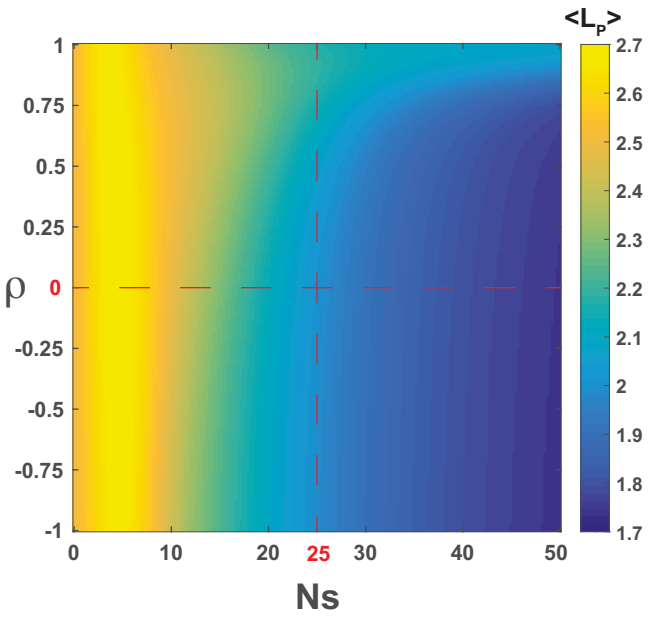


Figure 4.38: **Mean number of promiscuous TF positions at steady state decreases with selection intensity.** We plot the mean number of promiscuous positions at steady state, $\langle L_P \rangle$ (out of $L = 5$), for different values of signal correlation ρ and selection strength Ns. Steady state values of $\langle L_P \rangle$ are within a relatively small range. As selection strength increases, $\langle L_P \rangle$ decreases, yet still remains above zero. Parameter values: $n_G = 2$, $r_P = 3$; other parameters are at their baseline values.

Evolutionary pressure on the binding sites is therefore temporarily relieved, until the specificity of the TF is reestablished by a back mutation. Without promiscuity-promoting mutations, TF-BS co-evolution must proceed in a tight sequence of compensatory mutations; with promiscuity-promoting mutations, such a precise sequence is no longer required, although one extra mutation is needed to reestablish high TF-BS specificity. As shown in Fig. 4.36a, with promiscuity, the fraction of deleterious mutations along the evolutionary path towards specialization is reduced, an effect that grows stronger with increasing L. As shown in Fig. 4.36b, this has drastic effects on the time to specialization. Without promiscuity, increasing the selection strength, Ns, decreases the required

time when each TF regulates one gene, as expected for a landscape with large neutral plateaus but with no fitness barriers. For $n_G > 2$, however, the landscape develops barriers that need to be crossed, and evolutionary time starts increasing with Ns. In contrast, promiscuity enables fast emergence of TF specialization even with multiple regulated genes in a broad range of evolutionary parameters (although there are also costs due to high promiscuity).

Time to specialization

A speed-up via promiscuity mutations along various different pathways is shown in Fig. 4.39. The speedup of the fast `Partial` pathway (3 and 4) is not very large, but the speedup of the slow `Partial` (2 and 5) and the slow `One TF Lost` (1) pathways is considerable, an effect that increases with increasing Ns (see Fig. 4.32 for details of the pathways). Promiscuity-promoting mutations act by converting deleterious BS mutations into neutral or beneficial ones. By that they effectively lower or even remove fitness barriers. This effect is more significant with a large number of downstream genes, where more constraints on TF evolution exist. The fraction of different pathways does not change much if promiscuity-promoting mutations are present. Note that as a function of Ns, the fraction of fast `Partial` pathways does not change considerably, but the fraction of slow `Partial` pathways decreases while increasing the fraction of slow `One TF Lost` pathways. A reduction in N would not have a similar effect to promiscuity-promoting mutations, even though both flatten the fitness landscape. While promiscuity-promoting mutations flatten certain parts of the fitness landscape, building ridges across local fitness peaks, a reduction in N makes the overall fitness landscape flatter, making evolutionary pathways more vulnerable to meandering on huge neutral landscapes, and effectively slowing down specialization.

Typical trajectory

Promiscuity-promoting mutations play different roles in different phases of the evolutionary trajectory. While after specialization they are less favorable (because they lower binding affinity and potentially destabilize the specialized state), during adaptation they can facilitate fitness valley crossing. In Fig. 4.40, we plot the trajectory of the average number of promiscuous TF positions as a function of time. Starting with no promiscious positions in the `Initial` state, the number of promiscuous positions increases during the transient `One TF Lost` state, and then decreases to reach its steady state value after reaching the `Specialize Both` state. The speedup of evolution is mainly during the transient `One TF Lost` phase, where the number of promiscuous positions peaks.

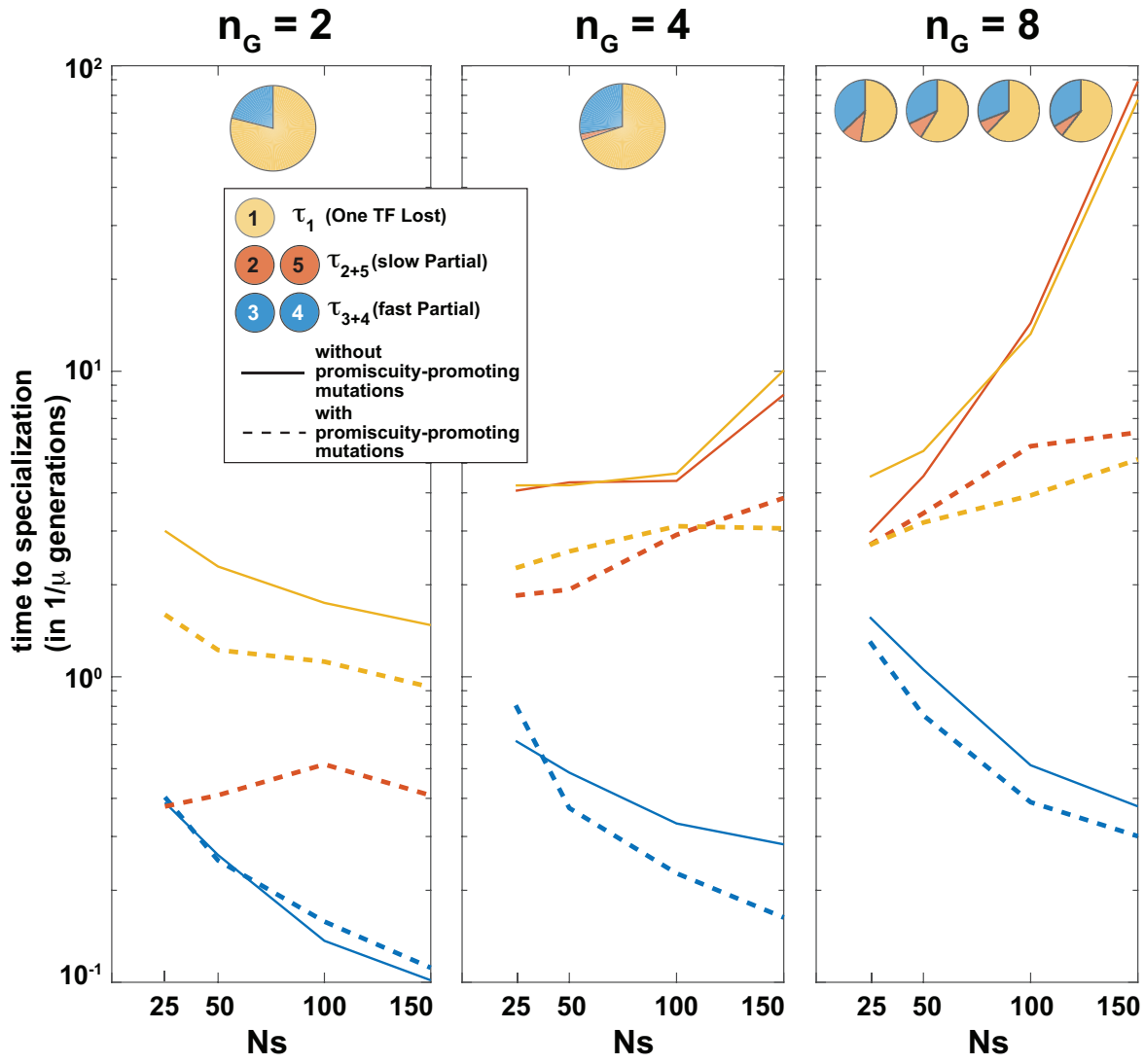


Figure 4.39: **Promiscuity-promoting mutations accelerate specialization.** We plot the times to specialization via different pathways that are depicted in Fig. 4.32, as a function of Ns for different values of n_G (the number of downstream genes per TF), in the absence (solid lines) and presence (dotted lines) of promiscuity-promoting mutations. Specialization times are shown for the slow `One TF Lost` pathway (numbered 1, yellow), the slow `Partial` pathway (numbered 2 and 5, red), and the fast `Partial` pathway (numbered 3 and 4, blue). In general, promiscuity-promoting mutations shorten evolutionary specialization times. This effect is particularly marked for the slow pathways (`One TF Lost` and slow `Partial`) and for large numbers of downstream genes n_G. The pie charts illustrate the fraction of the various pathways at each n_G value. For $n_G = 8$, we plot the pie charts for the different Ns values marked on the x-axis.

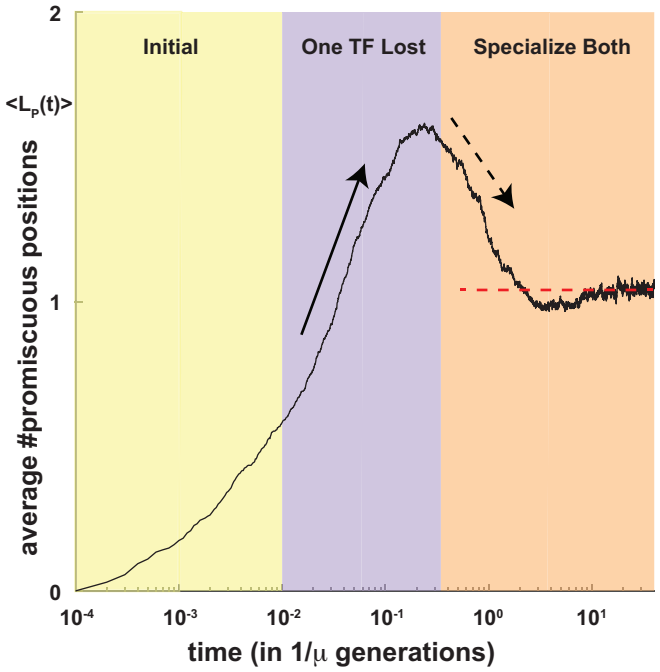


Figure 4.40: **Number of promiscuous positions transiently peaks during adaptation and relaxes after specialization to an intermediate steady state value.** We plot the average number of promiscuous positions $\langle L_P(t) \rangle$ as a function of time for $L = 5, n_G = 4, Ns = 250$ and $r_P = 10$; other parameters are at baseline values. Solid black arrow indicates the increase in the number of promiscuous positions in the transient `One TF Lost` phase, while the dotted black arrow indicates their decrease after specializing. The red dotted line indicates the steady state value of $\langle L_P \rangle$.

4.11 Discussion

The concept of fitness landscapes has been a dominant driving force for decades behind framing and answering questions in population genetics. While inferring the complete fitness landscape is impractical and has not been addressed in the field, this concept has given rise to a large body of theoretical work into evolution on toy-model like fitness landscapes in which interesting features have been artificially put in by hand [Kauffman and Levin, 1987; Kryazhimskiy *et al.*, 2009]. This has also led to recent efforts to map out empirical fitness landscapes albeit only around the wild-type on a small scale. For biological systems involving molecular recognition, biophysical constraints acting on these interactions are informative enough to permit a specification of the fitness landscape, and thereby allow us to computationally explore them. A few examples are, the secondary structure of RNA [Schuster *et al.*, 1994], antibody-antigen interactions [Adams *et al.*, 2016], protein-protein interactions [Podgornaia and Laub, 2015], and transcription factor-DNA binding [Aguilar-Rodríguez *et al.*, 2017], that we explored here.

In this chapter, we exploited this prior knowledge on the biophysics of TF-DNA binding and gene regulation, and their connection to function and fitness, to construct a fitness

landscape for a key evolutionary event of TF duplication by which regulatory networks grow in size and organisms gain new TFs. In contrast to toy model landscapes, the fitness landscape that we construct bottom-up from the underlying biophysics contains not only the essential empirical features, but also complex features like tuneable ruggedness that have been only of theoretical interest so far. Other essential concepts - the fact that specialization is driven by avoidance of regulatory crosstalk; the importance of the mutational entropy; the dependence on number of downstream genes; the existence of transient network configurations preceding specialization, which crucially impact dynamics; and the importance for evolutionary outcomes of the statistical properties of the signals that TFs respond to - emerge naturally out of such a construction. Crucially, this does not come at an increased modeling cost; while complex and containing many key features of interest, the fitness landscape is determined only by a few underlying parameters, most of which are known well. Also, the typical problem of an exponentially large space of genotypes can be coarse-grained to a small set of functional phenotypes that allow easy computation and biological interpretation. Moreover, this combination of biophysical and co-evolutionary approaches is applicable generally to the evolution of any biological system involving molecular recognition.

First, we computed the evolutionary steady state, which showed that correlation between upstream environmental signals, and in general their presence/absence statistics, act as a key determinant of whether the duplicate TFs specialize in their function (Fig. 4.11a-b). We showed that one TF duplicate will be lost due to neutral drift (and mutational entropy) unless the signals are sufficiently uncorrelated from each other. As a consequence, the *effective dimensionality of environmental signals* dictates the complexity of genetic regulatory networks [Friedlander *et al.*, 2015], reminiscent of information-theoretic tradeoffs in sensory neuroscience [Tkačik *et al.*, 2010]; in evolutionary terms, selection to maintain complex regulation needs to withstand the mutational flux into vastly more numerous but less functional network phenotypes ("survival of the flattest"). In chapter 2, I showed that finite biochemical specificity in molecular recognition events limits the complexity of genetic regulatory networks [Friedlander *et al.*, 2016]; an interesting direction for future research is to understand how the balance between regulatory crosstalk, environmental signal statistics, and evolutionary constraints together ultimately determine the number of TFs that can be stably maintained.

A clinching support for our complex biophysically realistic fitness landscape, as compared to simpler allelic models that neglect the topology of the sequence space, comes from the evolutionary dynamics towards specialized states. Timescales and pathways to specialization are completely shaped by the properties of the biophysical fitness landscape, and offer us important insights into the contexts in which specialization might occur. Specifically, we show that the fast pathway to specialization transitions through `Partial` states where neither of the two TFs completely loses binding and compensate for each other while transiting through intermediate evolutionarily transient states. Interestingly, some of this mutual compensation occurs due to the existence of crosstalk interactions, hence permitting fast adaptation via these transient states, by maintaining the network function through one TF, while the other is free to diverge in a series of

mutations to the TF and its future binding site [Shultzaberger *et al.*, 2012]. So while crosstalk thus enables some amount of network plasticity during early adaptation, it is ultimately selected against, when TFs become fully specialized [Rowland and Deeds, 2014; Eldar, 2011], a situation that we explored in chapter 2. In the protein-protein-interaction literature, `Partial` states are sometimes referred to as promiscuous states, and they have been suggested as evolutionarily accessible intermediaries that relieve the two interacting molecules of the need to evolve in a tight (and likely very slow) series of compensatory mutations [Aakre *et al.*, 2015]. In contrast to the fast pathway, the slow pathway involves a complete loss of TF-BS binding interactions; the long timescale emerges from long dwell times while the TF and the binding sites evolve in a nearly neutral landscape before TF-BS specificity is reacquired. Long binding sites and (perhaps counter-intuitively) fast TF mutation rates favor the slow pathway, while fast sensing domain mutation rates favor the fast pathway.

When a more realistic case of each TF regulating multiple target genes is considered, the situation changes qualitatively [Sengupta *et al.*, 2002]. On the one hand, entropy makes pathways that pass through the `One TF Lost` state dynamically uncompetitive, as multiple binding sites would have to emerge *d*e novo to reestablish interactions with a diverged TF. This would favor fast pathways through `Partial` states. On the other hand, because of increased constraints on the TFs, the biophysical fitness landscape develops frustration (or sign epistasis) as $n_G > 2$ and the timescales to specialization lengthen with increasing selection strength when passing through `Partial` states. Such a situation arises when one TF together with a set of binding sites (but not all) coevolve away to regulate one pathway, leaving behind a few other BSs that now have to cross a fitness valley to be bound by the diverged TF. We demonstrate that frustration is relieved by promiscuity-promoting mutations in the transcription factor, which increase the flexibility in the TF's binding preferences letting the BS lagging behind to also catch up, enabling fast emergence of specialization even with multiple regulated genes.

That coevolution is important to understand TF evolution has been attested by recent experimental studies that have demonstrated how a combination of *cis* and *trans* mutations have the potential to rewire gene regulatory networks. Such a coevolution allows for the emergence of new functions via transient and promiscuous configurations, in accordance with our model [Pougach *et al.*, 2014]. While we focused on a specific evolutionary scenario involving TF duplication, gene regulatory networks can rewire in numerous other ways. For example, Sayou *et al.* studied the evolution of TF-DNA binding specificity while the TF remains present in a single copy [Sayou *et al.*, 2014]. Duplicated TFs can also be re-used in ways that are different from what we considered [Pérez *et al.*, 2014]. Our results do, however, make predictions for expected timescales to reach different network configurations after gene duplication, which can be compared to bioinformatic data; alternatively, genomic data on TF duplication events could be used to infer selection pressures favoring regulatory divergence.

In summary, our study suggests that TF specialization proceeds through intermediate states that make use of crosstalk to maintain system functionality. A typical such state is

that in which one TF has already specialized for its input signals but not yet for the target genes, while the other TF is not yet specialized for the input signals but only regulates one gene. In the presence of multiple target genes, these intermediate states are likely to be inhabited by promiscuous TFs that are flexible in their BS binding preferences, with the promiscuity vanishing at the end of specialization in the steady states. Such a picture is qualitatively different from the accepted idea of a simple and sequential progression of compensatory mutations in the TF and its binding sites [Poelwijk *et al.*, 2006; de Vos *et al.*, 2015]. As we showed in this chapter, it depends fundamentally on the underlying biophysical model of TF-BS interactions and gene regulation function, predicts faster specialization times that overcome the ruggedness of the fitness landscape coming from coevolutionary constraints, and conveys the importance of promiscuity in TF evolution.

Chapter 5

Bioinformatic analysis of the evolution of Zn-finger TFs

5.1 Introduction

In the previous chapter, we investigated the evolution of duplicate TFs in a joint framework of TF-DNA based biophysical model of gene regulation and population genetics. This rich theoretical framework, which is based only on a few assumptions about the functional role of TFs in signal transmission and gene regulation, predicts the conditions under which duplicate TFs specialize to perform different functions. Motivated by these results, in this chapter, I will describe a preliminary bioinformatic analysis of the evolution of a major family of transcription factors – C2H2 Zn-finger TFs.

There are close to $\sim$ 800 genes in the human genome [Lambert *et al.*, 2018] that give rise to proteins containing a Zn-finger DNA-binding domain (DBD). Compared to other families of TFs, Zn-finger TFs are more modular owing to the presence of multiple DBDs in each protein that can have different binding specificities. At a finer scale, each DBD is very versatile in its binding modes, with mutations in a few key residues on the DBD allowing binding to the whole range of possible DNA sequences [Najafabadi *et al.*, 2015]. Apart from acting as transcription factors that are involved in the regulation of various cellular processes, Zn-finger TFs are also involved in chromatin remodeling [Kim *et al.*, 2015] and repression of transposable elements (TEs) [Yang *et al.*, 2017].

Among transcription factors, the Zn-finger family has undergone the most successful duplications, to which it owes its large numbers in various animal genomes [Lambert *et al.*, 2018]. While reasons such as the modular nature of Zn-finger DBD composition, and

their versatility in DNA binding have been proposed, we still lack a good understanding of the patterns of Zn-finger TF evolution that used these features to expand rapidly in mammalian genomes. Transposable element repression has been suggested to be a major driving force [Yang *et al.*, 2017] behind the recent expansion of KRAB domain containing Zn-finger TFs (KZNFs) in primate genomes. KZNFs recognize specific TE sequences via their Zn-finger DBDs and recruit TE repressing factors via their KRAB domains. It has been suggested that constant invasion and expansion of new TE families drives the retention of KZNF duplicates, and their subsequent adaptation to bind changing TE sequences, resulting in a coevolution between TEs and KZNFs [Jacobs *et al.*, 2014]. Further, the reuse of old KZNFs to bind new TEs, and also cooption of TEs into regulatory sequences has been suggested as an additional mechanism for the large numbers of Zn-finger TFs [Ecco *et al.*, 2017; Chuong *et al.*, 2017].

In this chapter, I will explore the recent evolution of Zn-finger TFs in terms of their Zn-finger DBDs and KRAB domains, and also comment on KZNF coevolution with TEs. I will first explore how various genic features like the number of Zn-finger DBDs, the number of exons and their length in bp, chromosomal location etc. correlate with each other for Zn-finger TFs. For instance, clustering of Zn-finger TF genes on chromosomes reveals their origin via gene duplication, and a strong correlation between the number of Zn-finger DBDs and length of one particular exon reveals that all DBDs are usually coded in a single exon.

Then I will turn to the major question that I will attempt to answer: what are the patterns of positive selection that drive the divergence of initially identical Zn-finger paralog TFs (immediately after duplication)? To answer this, I will compute dN and dS, the rates of non-synonymous and synonymous changes in the coding sequences of Zn-finger TF paralogs, and use dN/dS ratio as a signal to detect the pattern of selection. First, I will describe models that consider that all sites in the protein sequences of the Zn-finger TF paralogs have the same dN/dS ratio. While this assumption of similar selection pressure on all sites of a protein is empirically incorrect, it still offers first insights into how selection has operated on Zn-finger TF paralogs of various ages. By computing dN/dS ratios for specific domains like KRAB and Zn-finger DBDs, we obtain additional insights into varying patterns of selection on different domains across time. I will show that Zn-finger DBDs perhaps initially experience selection to diverge and bind different sequences, with KRAB domains undergoing positive selection in older paralogs.

To overcome the shortcomings of site-averaged models, I next consider site-specific models of dN/dS computation. By comparing pairs of nested models that account for evolution without positive selection (purifying + neutral) and evolution with positive selection (purifying + neutral + positive), I will show that for a few sets of KZNF TFs, a strong signal of positive selection can be observed at key residues on the DBDs, which are amino acids that contact nucleotides on the DNA via hydrogen bonds to establish sequence-specific binding. Finally, I will show that paralog KZNFs bind TEs of often similar typical ages, and that the paralogs (dated by the age of duplication) often arise at the same time as the younger among the typical ages of TEs bound by them. This indicates

that KZNF TFs are retained after duplication to immediately adapt to bind to TEs that arise newly. However, some paralogs have arisen before the younger among the typical ages of TEs, which indicates that some old KZNFs are reused from before to repress new TEs.

While this treatment of TF-TE coevolution is different from the gene regulation model developed in the previous chapter, we choose to investigate this first as there are a significant number of KZNF paralogs that seem to be involved in TE repression. The question of whether some of them become co-opted for gene regulation is one that will subsequently ask in future research. Also, the dynamics of TF-TE coevolution is different as there is no stationary state, and is always out-of-equilibrium with the invasion of new TEs and the expansion of various TE families.

5.2 Bioinformatics pipeline

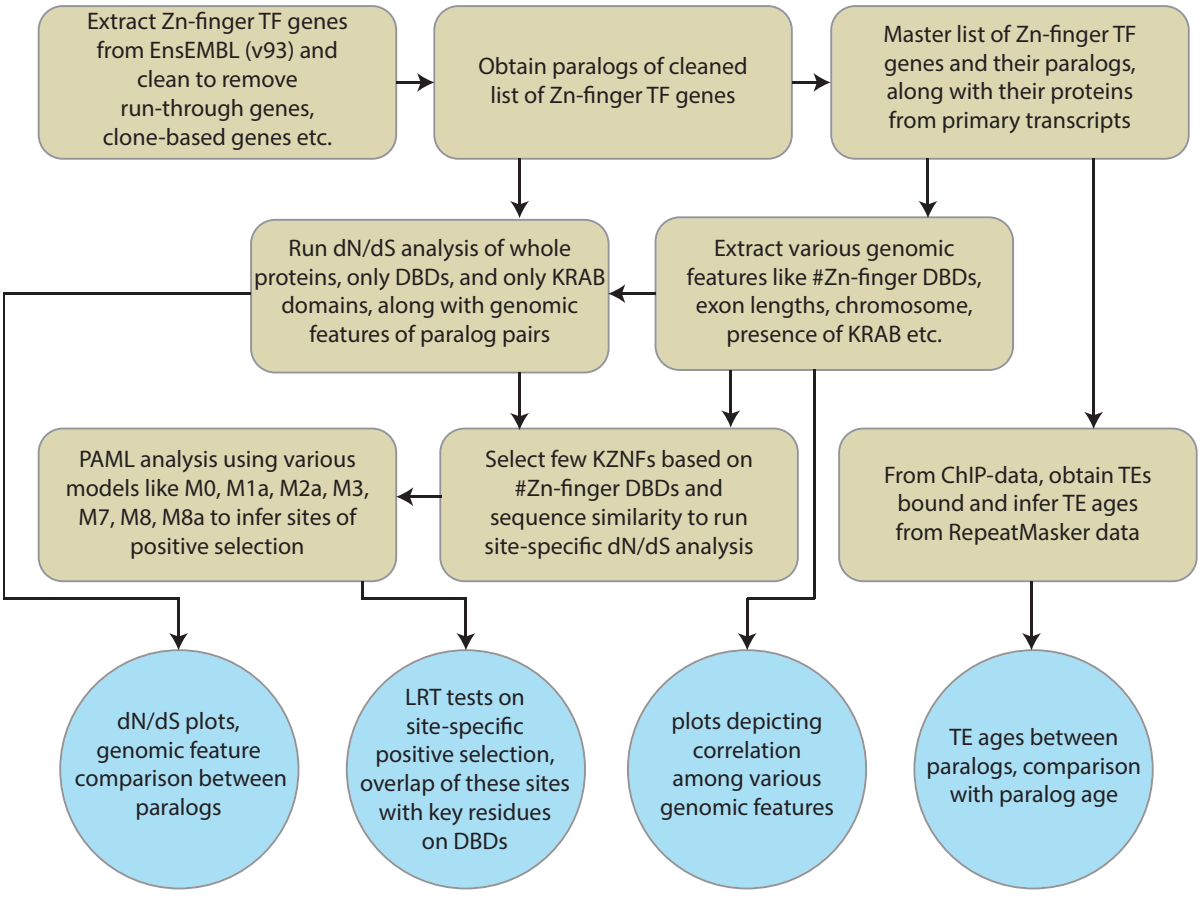


Figure 5.1: (Caption next page.)

The bioinformatics pipeline employed in this chapter is described in Fig 5.1. We use EnsEMBL Release 93 (July 2018, [Cunningham *et al.*]) RepeatMasker [Smith *et al.*, 2016], and Imbeault 2017 [Imbeault *et al.*, 2017] as the sources of genomic data, repeat

Figure 5.1: (Previous page.) **Bioinformatics pipeline followed to extract and analyze Zn-finger TF evolution.** First, we extract Zn-finger TF genes in the human genome from EnsEMBL (v93) [Cunningham *et al.*] and then remove spurious cases like run-through genes, clone-based genes etc., to obtain a cleaned list of Zn-finger TF genes. Then, we obtain a list of paralogs of the cleaned list of Zn-finger TF genes, together with the taxonomical age of each duplication event. From the paralog pairs, we obtain a master list of Zn-finger TF genes and a few other non Zn-finger TF gene paralogs, along with the proteins corresponding to their primary transcripts. Using this master list of genes, we extract from EnsEMBL various genomic features of each gene like presence of a KRAB domain (KZNF gene), number of Zn-finger DBDs in each protein, length of exon coding for Zn-finger DBDs (if they are in the same exon), chromosomal location, domain architecture of proteins etc. Apart from teasing apart the correlations among these genomic features, we use them and the paralog pair list to compute dN/dS and dS using the "MS" (model selection) method from KaKs_Calculator [Zhang *et al.*, 2006], for pairwise alignments of whole proteins, only Zn-finger DBDs and KRAB domains. From this we obtain dN/dS vs dS plots, when a single dN/dS ratio is assumed to hold across all sites in the aligned proteins of interest. To overcome the dampening of the dN/dS signal because of averaging across sites, we select a few KZNFs based on their sequence similarity and perform site-specific dN/dS analysis using various maximum likelihood models of neutrality and selection from PAML [Yang, 2007]. Apart from this, we also obtain, from ChIP-exo data on KZNFs [Imbeault *et al.*, 2017] and from TE data [Smith *et al.*, 2016], the typical ages of the TEs bound by various KZNFs, and compare the age of TEs with the age of the duplication event.

data (for TEs) and ChIP-exo data. First, from EnsEMBL, we extract a raw list of Zn-finger TF genes by using the criterion that one of the gene's proteins should have an annotated SMART domain "SM00355", which corresponds to a Zn-finger DBD [Schultz *et al.*, 2000]. Then, we clean this raw list to remove spurious cases like read-through genes and clone-based genes (allelic genes) to obtain a clean list of Zn-finger TF genes.

We query EnsEMBL for paralogs of these genes, and after cleaning the paralog list to account for spurious cases, we now have a master list of paralogs in which at least one copy comes from the cleaned list of Zn-finger TF genes. We also have the primary transcript corresponding to each of the genes from this paralog list; note that not all Zn-finger TF genes from the master list are annotated to have at least one paralog (for instance, due to large sequence divergence), and that non Zn-finger TF genes are also present in the the paralog pairs (for instance, due to loss of all Zn-finger DBDs after duplication). For this master list of all genes – Zn-finger TF genes with annotated paralogs, Zn-finger TF genes without annotated paralogs, and non Zn-finger TF genes that are annotated as Zn-finger TF gene paralogs – we extract their various genomic features like the chromosomal location of the gene, protein domain architecture – which domains are present and at what locations on the protein, number of Zn-finger DBDs, exon containing Zn-finger DBDs (if they occur on the same exon, as happens for a majority), the length of this exon, the

length of the coding part of the exon, the length of intron immediately preceding the Zn-finger exon, the peptide sequence of the proteins, the raw intronic and exonic sequences, and a few other features. We compute how correlated these features are across the whole set of genes involved to get first clues about the features that evolve as part of Zn-finger TF evolution.

Next, we use these features, primarily the identities and locations of various domains on the proteins, and the peptide and coding DNA sequences, to compute the dN/dS and dS values of each paralog pair for pairwise alignment of the whole protein sequences, of only the Zn-finger DBD sequences, and of only KRAB (SMART domain "SM00349") sequences. We use the "MS" (model selection) method from KaKs_Calculator [Zhang *et al.*, 2006] that uses maximum likelihood techniques and AIC to infer the best underlying DNA and codon evolution model. After understanding a few broad patterns of Zn-finger TF evolution from dN/dS rations, to overcome the shortcomings of averaging across all the sites – MS model assumes that dN/dS is same across all sites of the protein chunk considered, we then employ site-specific models of dN/dS computation.

We first select a few KZNF (Zn-finger TFs with a KRAB domain) that share medium to large sequence similarity based on dS values and share the same number of Zn-finger DBDs to form four different sets of KZNF genes. Then on each set, we first run ClustalW2 multiple sequence alignment [Thompson *et al.*, 1994] and use PAL2NAL [Suyama *et al.*, 2006] to construct a multiple codon alignment from the protein alignment. Then we run maximum likelihood models M0, M1a, M2a, M3, M7, M8 and M8a from PAML [Yang, 2007], corresponding to various scenarios of nearly neutrality and positive selection, to obtain likelihoods and information on which specific sites are positively selected for. As various pairs of these models are nested, we use likelihood ratio tests on these likelihoods to infer whether rejecting the null model of (near) neutrality against the alternative model of positive selection is statistically significant.

We also compare the sites inferred to be positively selected in each set with the key residues on the DBDs (those amino acids that contact nucleotides on the DNA via hydrogen bonds) of the underlying KZNF TFs, and infer that KZNFs have undergone selection at key residues that change their binding specificities towards DNA.

Next, motivated by evidence that a major function of KZNFs is TE repression [Najafabadi *et al.*, 2015; Yang *et al.*, 2017], and that they have coevolved with TE sequences [Jacobs *et al.*, 2014], we query RepeatMasker for the copy numbers of various families of TEs in representative species from a set of hierarchical taxonomical nodes (based on *Homo sapiens*) starting from Homininae (subfamily comprising humans, chimpanzees and bonobos) through the order Primates, to the class Mammalia comprising mammals, and the clade Amniota comprising reptiles, birds and mammals. From these we obtain a rough estimate of the age of various TEs of interest on a taxonomical scale. We then compare the typical ages of the TEs bound by the paralogs, and also ask if TE ages are correlated with paralog age, which we consider to be the taxonomical age of the duplication event inferred from a reconciliation of gene trees with the reference species tree.

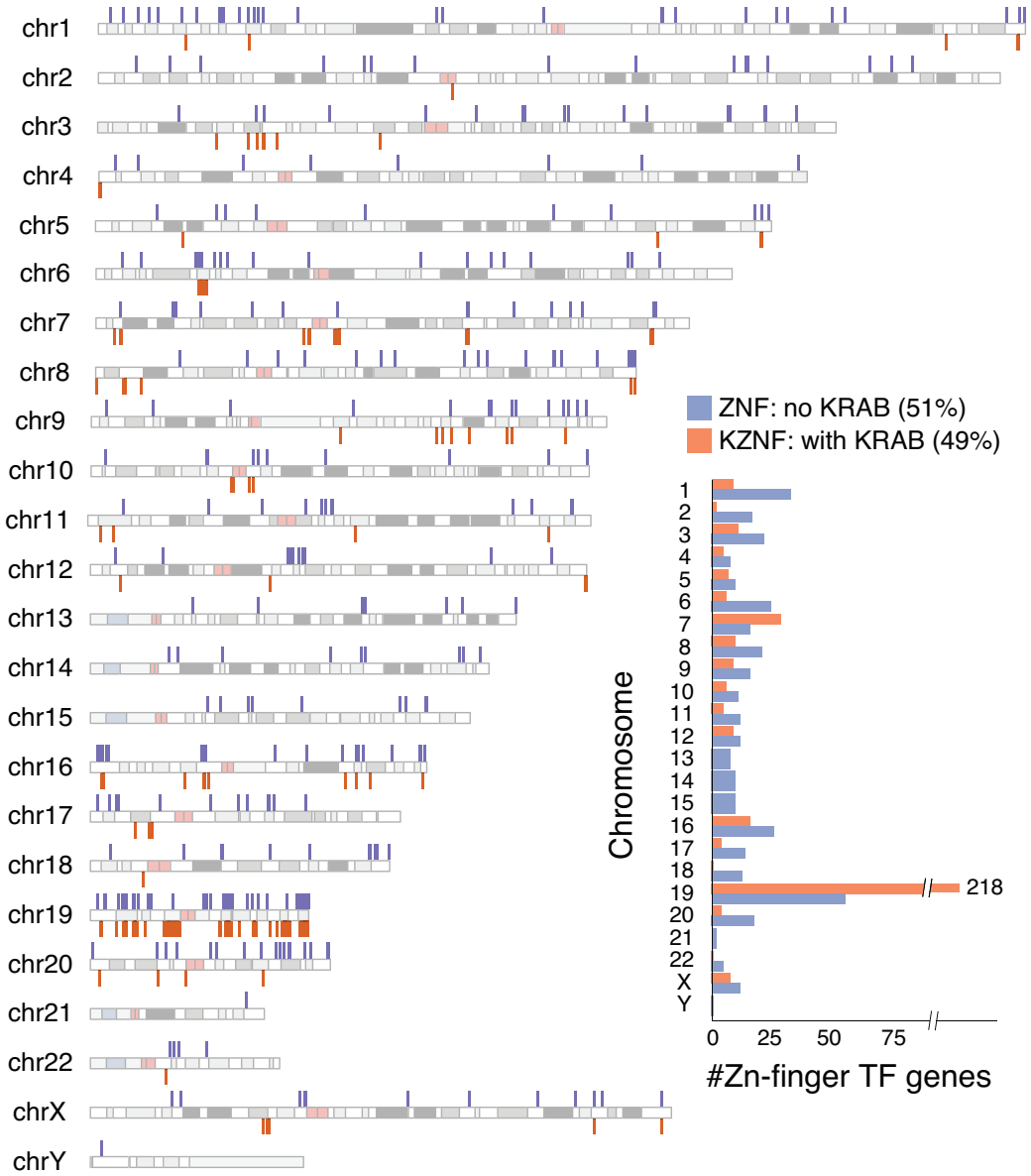


Figure 5.2: **Zn-finger TF genes are distributed heterogeneously on the human chromosomes, and are often present in clusters.** The spatial positions of the Zn-finger TF gene repertoire, classified by the absence (violet, ZNF genes) and presence (orange, KZNF genes) of a KRAB domain, from the human genome are marked on the human chromosomes. Also, in the inset, we show a histogram of the the ZNF genes on the various chromosomes. Both ZNF and KZNF genes are distributed heterogeneously on the various chromosomes, and are often found in clusters, which probably originate via gene duplication events. Many of these clusters are found on chromosome 19. Pearson's correlation indicated that there is no significant correlation between chromosome lengths and the number of either ZNFs, KZNFs, or all Zn-finger TFs on them – p-values $p = 0.4511$ for ZNFs, $p = 0.1899$ for KZNFs and $p = 0.5303$ for all Zn-finger TFs.

5.3 Genomic features and their correlation

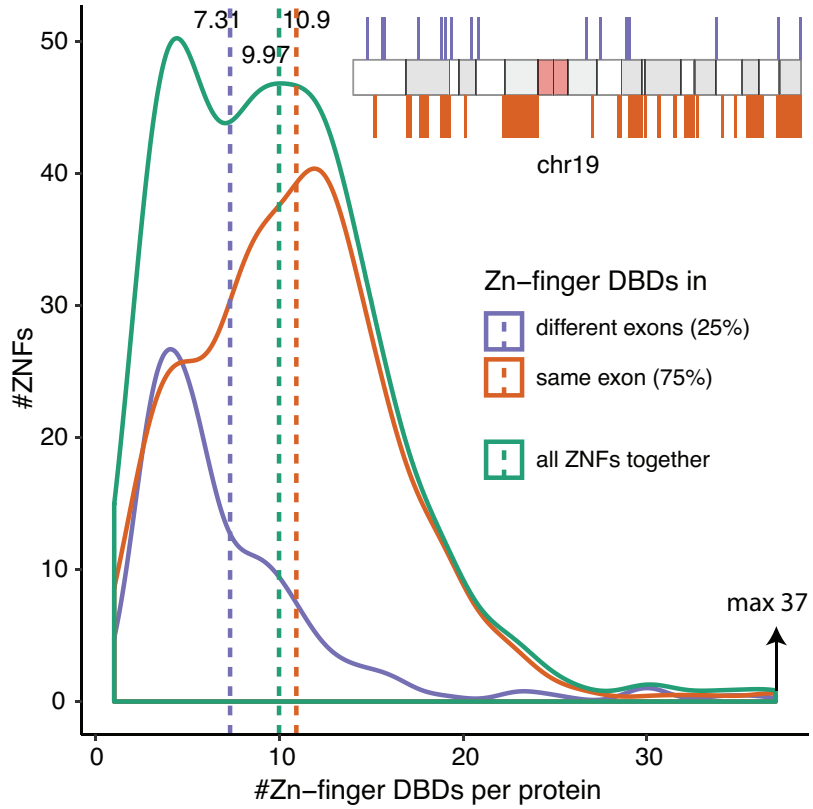


Figure 5.3: **Zn-finger TFs have varying numbers of Zn-finger DBDs.** Different Zn-finger TFs have varying number of Zn-finger DBDs in their proteins. Zn-finger TFs whose Zn-finger DBDs are in the same exon (orange) have a larger number of DBDs – mean 10.9, compared to those whose Zn-finger DBDs are in different exons (violet) – mean 7.31 (means significantly different according to Mann-Whitney U test, $p < 2.2e-16$) The mean number of Zn-finger DBDs in all Zn-finger TFs is 9.97. Also, the spatial positions of these two classes of Zn-finger TFs are marked on chromosome 19 in the inset . Zn-finger TFs whose DBDs are on different exons tend to cluster less and are located separately from those whose DBDs are in the same exon.

In Fig. 5.2, we mark the chromosomal positions (no strand information) of the various Zn-finger TF genes, colored according to the presence (orange, KZNF) and absence (violet, ZNF) of a KRAB domain in the protein product of the gene. Both ZNFs and KZNFs, each of which make up about 50% each of the repertoire of Zn-finger TFs, occur in clusters on the various chromosomes, with chromosome 19 inhabiting a major fraction of these clusters. This further indicates the gene duplication origin of ZNFs and KZNFs.

Next, we probe the distribution of the number of Zn-finger DBDs in different Zn-finger TFs. We plot empirical density plots (Gaussian kernel on histograms) of the number of Zn-finger DBDs, grouped by their presence in the same exon (orange: same exon, violet:

different exons, green: all Zn-finger TFs together) in Fig. 5.3. Zn-finger TFs with all their DBDs in the same exon tend to have a larger number of DBDs (mean 10.9), compared to those that have DBDs in different exons (mean 7.31) – means significantly different according to Mann-Whitney U test, $p < 2.2e-16$. Overall, the average number of Zn-finger DBDs per protein is close to 10, with a large variance. The maximum number of DBDs in a Zn-finger protein is 37. Each DBD contacts about 3 nucleotides on the DNA with a large affinity, meaning the average length of DNA sequences bound by Zn-finger TFs is 30bp. Such large binding site sequences allow Zn-finger TFs to specifically target locations on the DNA, and reduces spurious binding to off-target sites. This, together with the modularity offered via the chopping and changing of each DBD separately, make the Zn-finger TF family versatile and flexible compared to other major families of TFs.

We then ask how the number of exons and number of coding exons compare with each other among Zn-finger TF genes. As shown in Fig. 5.4, where we plot the total count of Zn-finger TFs with different numbers of exons and coding exons, these are strongly correlated (Kendall's tau coefficient, $\tau_b = 0.79, p < 2.2e-16$) with each other, meaning that most exons in Zn-finger TF genes act as coding exons. We use Kendall's tau because the underling data is not normally distributed (Multivariate Shapiro-Wilk test, $p < 2.2e-16$). We also see that there are 70 ZNF (no KRAB) and 3 KZNF genes that have only 1 coding exon, even though many of them have more than 1 exon. These might have arisen either by losing exons, for instance, by losing the KRAB-coding exons, leading to only the Zn-finger DBD coding exon – ZNF gene → KZNF gene, or by retro-transposition based duplication.

We already saw in Fig. 5.3 that in a majority of Zn-finger TFs, all the DBDs are coded in the same exon. This might imply that the number of Zn-finger DBDs might have little to do with the total number of exons (and thereby the number of coding exons), which is what we observe (Kendall's tau coefficient, $\tau_b = -0.02, p > 0.05$) in Fig. 5.5.

Next, we summarize all correlations among various genomic features in Fig. 5.6. The strand on which the Zn-finger TF genes are found does not correlate with any of the other features, as expected. Notice that the number of exons and coding exons are highly correlated ($\tau_b = 0.96, p < 2.2e-16$), as observed in Fig. 5.4. The number of Zn-finger DBDs does not share a significant correlation with the either the number of exons or the number of coding exons, but is highly correlated with the length of the coding part of the exon containing DBDs ($\tau_b = 0.53, p < 2.2e-16$). This is because all the DBDs are often placed on the same exon, and make up a significant fraction of the coding part of the exon. The loss of a few DBDs by degenerate mutations and/or the lack of their annotation, or the presence of other domains in the same exon, might be the reasons behind a correlation of $\tau_b = 0.53$ and not higher. The number of Zn-finger DBDs shares a smaller but significant correlation ($\tau_b = 0.16, p < 0.001$) with the total exon length, because of the presence of $3'$ UTRs in the exon.

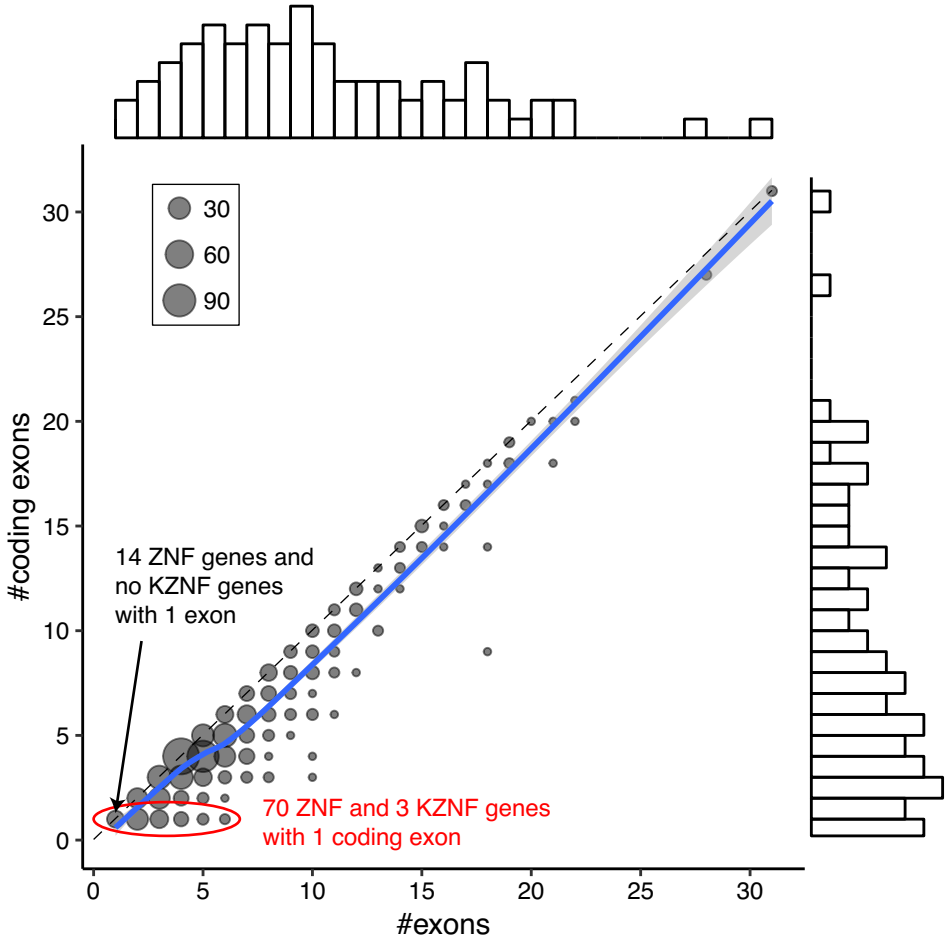


Figure 5.4: **The number of exons and the number of coding exons in Zn-finger TF genes are strongly correlated.** Here we plot the numbers (counts) of Zn-finger TF genes (ZNF and KZNF genes together) that have different numbers of exons (x-axis) and coding exons (y-axis) in them. The number of exons and the number of coding exons are very strongly correlated (Kendall's tau coefficient, $\tau_b = 0.79, p < 2.2e-16$), meaning that most exons in Zn-finger TF genes are coding exons. In blue, we plot a local regression curve. There are 70 ZNF genes and 3 KZNF genes with only 1 coding exon (even though many of them have more than 1 exon). It is possible that these arose either by losing exons (loss of KRAB-coding exons, leading to ZNF genes from KZNF genes) or via a retro-transposition based duplication. Also, there are 14 ZNF genes and no KZNF genes with only 1 exon in total.

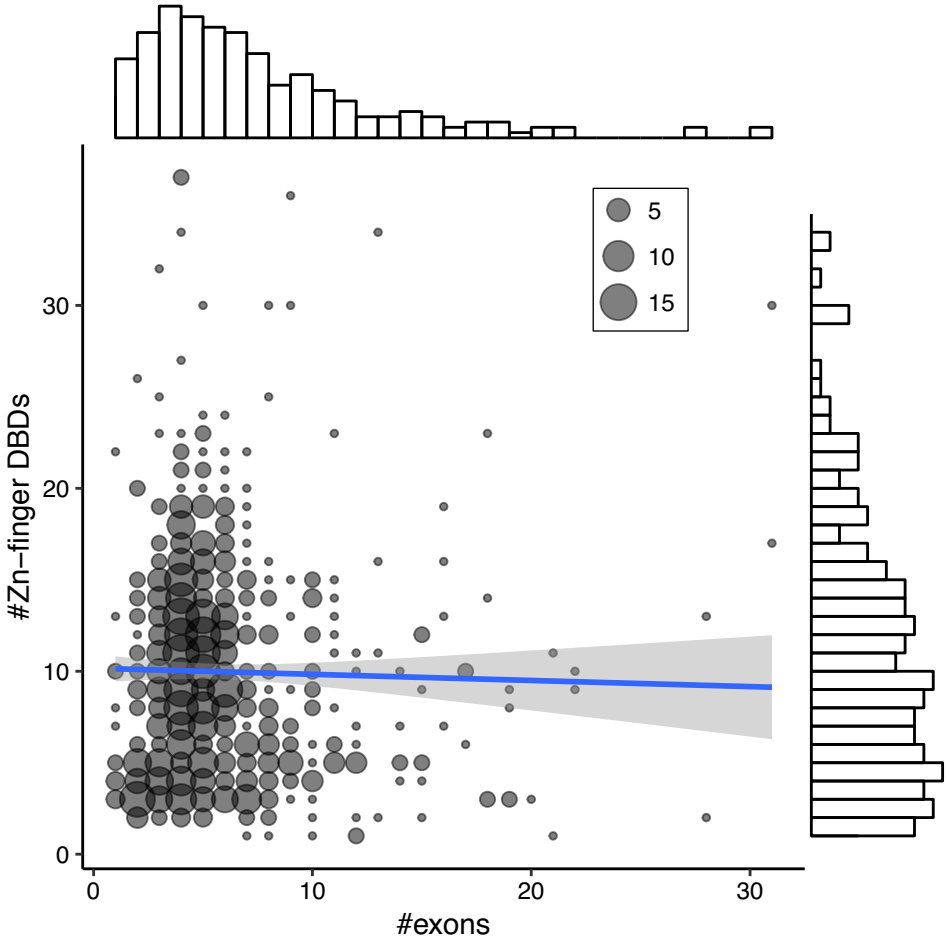


Figure 5.5: **The number of Zn-finger DBDs has no significant correlation with the number of exons in Zn-finger TF genes.** Here we plot the numbers (counts) of Zn-finger TF genes (ZNF and KZNF genes together) that have different numbers of exons (x-axis) and Zn-finger DBDs (y-axis) in them. The number of exons and the number of Zn-finger DBDs do not have any significant correlation (Kendall's tau coefficient, $\tau_b = -0.02, p > 0.05$), meaning that a larger number of exons does not mean a larger number of Zn-finger DBDs. This is also because all the Zn-finger DBDs are usually packed in a single exon. We plot in blue a linear regression curve that verifies this independence.

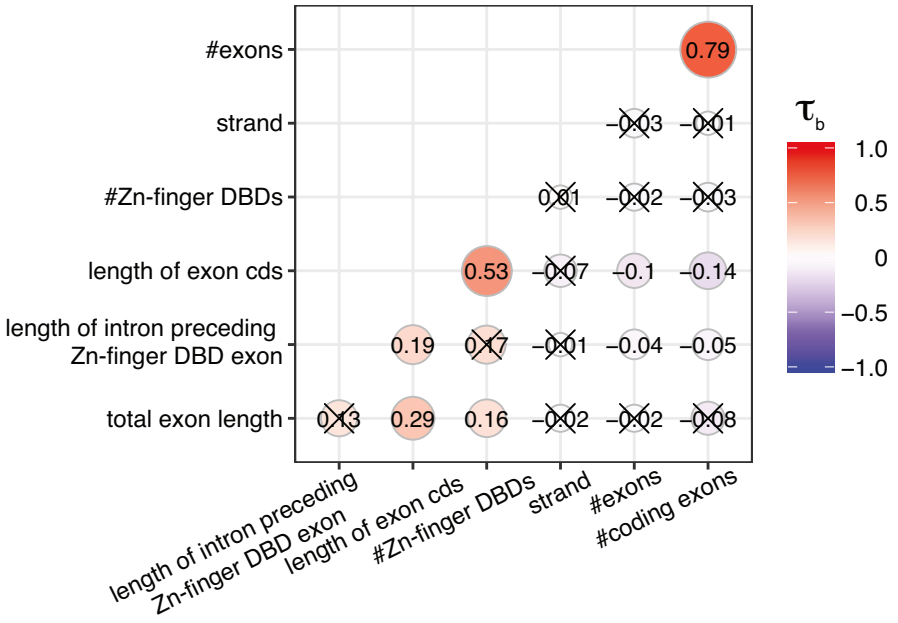


Figure 5.6: **Correlations among various features of Zn-finger TF genes inform us of their intron-exon-DBD structure.** We plot Kendall's tau coefficients, τ_b, among various features of the Zn-finger TF genes, and also mark with a "x" (cross) those pairs that do not share a significant correlation (significance level 0.05), The following features are considered – "#coding exons": the number of coding exons, "#exons": the number of exons, "strand": directionality of the DNA strand (+1 or -1), "#Zn-finger DBDs": the number of Zn-finger DBDs, "length of exon cds": the length of the coding part of the exon (without the UTRs), "total exon length": the total length of the exon, including possible UTRs, "length of intron preceding Zn-finger DBD exon": length of the intron immediately before the exon that codes for the Zn-finger DBD. None of the other genomic features are significantly correlated with "strand". As pointed out in Fig. 5.4, "#exons" and "#coding exons" significantly strongly correlated, $\tau_b = 0.79$. Also, as pointed out in Fig. 5.5, "#exons" and "#Zn-finger DBDs" are not significantly correlated. However, "#Zn-finger DBDs" is strongly correlated with "length of exon cds", indicating that all DBDs are placed on a single exon, and they make up a significant fraction of the coding part of the exon. The loss of a few DBDs by degenerate mutations and/or the lack of their annotation might be the reason behind a correlation of $\tau_b = 0.53$ and not higher.

5.4 Paralogs: genomic features and averaged dN/dS ratios

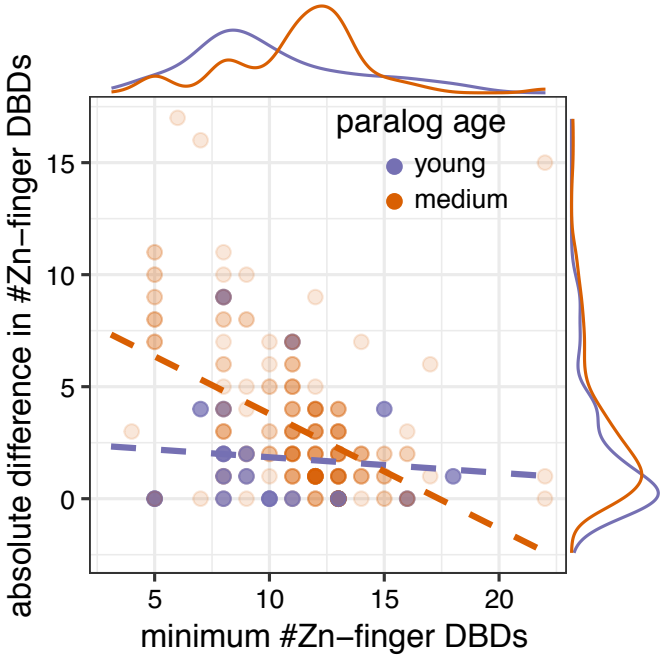


Figure 5.7: **Zn-finger TF genes lose (and gain) Zn-finger DBDs over time.** Along paralogous pairs of Zn-finger TF genes, there is often a difference in the number of Zn-finger DBDs. Let us say that x and y are the numbers of Zn-finger DBDs in the paralogs, and let us assume that $x \leqslant y$ wlg. Here, we show how $|x-y|$ depends on x for paralogs that are young (violet, $dS \leqslant 0.15$) and of medium age (orange, $0.15 < dS \leqslant 0.3$). For younger paralogs, x ranges across from low to high, but $|x-y|$ is typically low (regression line: violet, dashed), meaning that the younger paralogs share a similar number of Zn-finger DBDs. On the other hand, for paralogs of medium age, $|x-y|$ decreases as x increases (regression line: orange, dashed). While there exist TFs with higher number of Zn-finger DBDs that tend to have a paralog partner with a higher number of DBDs, there are many pairs in which one TF has a low number of DBDs with the other TF having a large number of them. This implies that DBDs are lost over time from TFs with a large number of them. If the number of Zn-finger DBDs between paralogs are correlated, we expect $|x-y|$ to not be correlated with x, which is what we observe from a non-significant Kendal's tau coefficient of $\tau_b = 0.29, p > 0.05$ for young paralogs. On the other hand, if the numbers of Zn-finger DBDs are not correlated between paralogs, we expect $|x-y|$ to be negatively correlated with x, which is what we observe from a significantly negative Kendall's tau coefficient, $\tau_b = -0.42, p < 3.353e-15$.

Next, we use the master paralog list obtained by querying EnsEMBL and subsequent cleaning to remove spurious gene annotations, to compute site-independent dN/dS ratios and dS values for all the paralog pairs. We use the "MS" (Model Selection) method

from KaKs_Calculator [Zhang *et al.*, 2006] which compares various maximum likelihood models of underlying DNA and codon evolution, to pick the best model according to AIC. We perform these calculations for pairwise alignment of whole proteins of the paralogs, and also compute domain-specific dN/dS ratios by using information about the domain architecture of proteins (Zn-finger DBDs and KRAB domains' locations on the protein) to form a domain-specific pairwise alignment of the corresponding protein chunks.

Before describing dN/dS ratios, I will describe how various genomic features compare across different paralog pairs. First, we focus on the relationship between the number of Zn-finger DBDs in paralogs, say x and y. As x and y are not bivariate normal, we use the Kendall's tau coefficient to check for correlations between x and y. When the number of data points are not large, any biases (for instance, if $x < y$ always) in the ordering of x and y might affect these coefficients. Hence, we randomized the data to overcome this bias and generated multiple datasets ($N_{perm} = 300$ in total), on each of which we computed Kendall's tau coefficient. For young paralogs ($dS <= 0.15$), over these permuted datasets, we obtain a mean Kendall's tau coefficient of $\langle \tau_b \rangle = 0.77$, with all of them significant at $p < 0.0001$. On the other hand, for paralogs of medium age ($0.15 < dS \leqslant 0.3$), we find no significant correlation ($p > 0.05$) for all the datasets.

We also use another approach to check the correlation between the number of Zn-finger DBDs between paralogs, and understand any particular patterns that might exist. We reorder the data for each paralog pair such that, $x < y$ always, and use the absolute difference in the number of Zn-finger DBDs ($|x - y|$) and the minimum number of DBDs (x) in either of the paralog TF as the variables of interest. Such an approach does not have the biases that raw data might have as described in the previous paragraph.

In Fig. 5.7, we plot $|x - y|$ vs x and the linear regression lines for each age group – young and medium – separately. Note that there is often a difference in the number of Zn-finger DBDs between paralogs ($|x - y|$ sometimes large), meaning that Zn-finger paralogs diverge by the loss/gain of DBDs. For young paralogs (violet, $dS \leqslant 0.15$), while x varies across a broad range of values, $|x-y|$ is usually low (seen also from the regression curve, violet dashed line). But for paralogs of medium age (orange, $0.15 < dS \leqslant 0.3$), $|x - y|$ is large at small x, but low at large x (seen also from the regression curve, orange dashed line). Moreover, when one considers the maximum number of DBDs, $\max(x, y)$, for each paralog pair, $\max(x, y)$ has a smaller variance for medium age paralogs than for younger paralogs. This paints a picture in which paralog pairs with large number of DBDs diverge such that one of the TF in the pair loses a few DBDs over time. This is a mode of evolution that complements evolution via point mutations at specific sites. We also compute Kendall's tau coefficient between $|x-y|$ and x to check if these relationships are significant. If the number of Zn-finger DBDs between paralogs are correlated, we expect $|x - y|$ to not be correlated with x, which we verify from a non-significant Kendal's tau coefficient of $\tau_b = 0.29, p > 0.05$ for young paralogs. On the other hand, if the numbers of Zn-finger DBDs between paralogs are not correlated, we expect $|x - y|$ to be negatively correlated with x. For paralogs of medium age, we find a significantly negative Kendall's tau coefficient, $\tau_b = -0.42, p < 3.353e-15$, verifying that the numbers of Zn-finger DBDs

between medium age paralogs are largely uncorrelated.

Next, as shown in Fig. 5.8, we find that, between paralogs, there is no significant correlation between the absolute difference in the number of Zn-finger DBDs and the absolute difference in the total length of exon sequences that contain the DBD (Kendall's tau, $p > 0.05$ for both age groups). However, for both young and medium age paralogs, we find a significant strong correlation ($\tau_b = 0.77, p < 0.0001$) between the absolute difference in the number of Zn-finger DBDs and the absolute difference in the lengths of the coding part of the exons that contain the DBD. This indicates changes in the coding part of the exons are the primary cause for the difference in the number of Zn-finger DBDs. A few possible mechanisms behind this could be recombination, shifts in intron-exon boundaries, or premature stop codons.

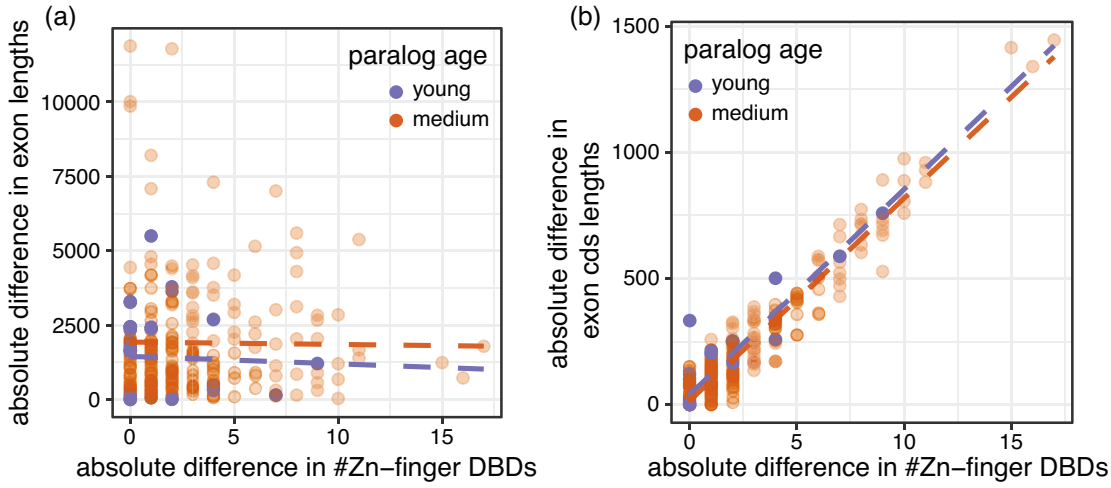


Figure 5.8: **Between paralog pairs, difference in the number of Zn-finger DBDs correlates well with the difference in the length of the coding part of exon sequences, while being uncorrelated with the difference in the total length of exon sequences.** For various paralog pairs, we show how the absolute difference in the number of Zn-finger DBDs between the two paralogs (x-axis), varies with the absolute difference in the **(a)** total lengths of the exons that contain the DBD, and **(b)** lengths of the coding part of the exons that contain the DBD. For both young (violet, $dS \leqslant 0.15$) and medium age paralogs (orange, $0.15 < dS \leqslant 0.3$), we find that there is no significant correlation between the absolute difference in the number of Zn-finger DBDs and the absolute difference in the total length of exon sequences that contain the DBD ($p > 0.05$ for both, Kendall's tau). On the other hand, we find a significant positive correlation between the absolute difference in the number of Zn-finger DBDs and the absolute difference in the lengths of the coding part of the exons that contain the DBD ($\tau_b = 0.77, p < 0.0001$ for both young and medium age).

Next, in Fig. 5.9, we compare the lengths of the exon coding for the DBDs (if they all are on the same exon) between the two paralog Zn-finger TFs, using the total exon lengths in Fig. 5.9a and only the length of the coding part of the exon in Fig. 5.9b. We use

the same method of permuting ($N_{perm} = 300$) to randomize any bias in the data, and compute mean Kendall's tau coefficient, $\langle \tau_b \rangle$. The total exon lengths between paralogs are not significantly correlated for both age groups (young: $\langle \tau_b \rangle = 0.02, p > 0.05$ for all permuted datasets, medium age: $\langle \tau_b \rangle = 0.13, p > 0.05$ for all permuted datasets). However, the length of the coding part of the exons is strongly correlated between young paralogs ($\langle \tau_b \rangle = 0.82, p < 0.00001$ for all permuted datasets), with those in medium age group not significantly correlated ($\langle \tau_b \rangle = 0.005, p > 0.05$ for all permuted datasets). Again, this suggests evolution by gain and loss of DBDs.

The overall picture conveyed by the above analysis is that of an interplay between evolution by the loss of DBDs, and evolution via point mutations (as described next). Duplication of a TF with many DBDs results in two copies, and one of the TF copy often loses DBDs by mutational/recombination mechanisms different from point mutations. This results in more flexibility with respect to the TF's binding and allows a quicker evolution to find a new target. In future research, we will investigate this interplay further, ask questions about the processes that shape the distribution of the number of DBDs, and also infer the phylogenetic trees connecting various DBDs.

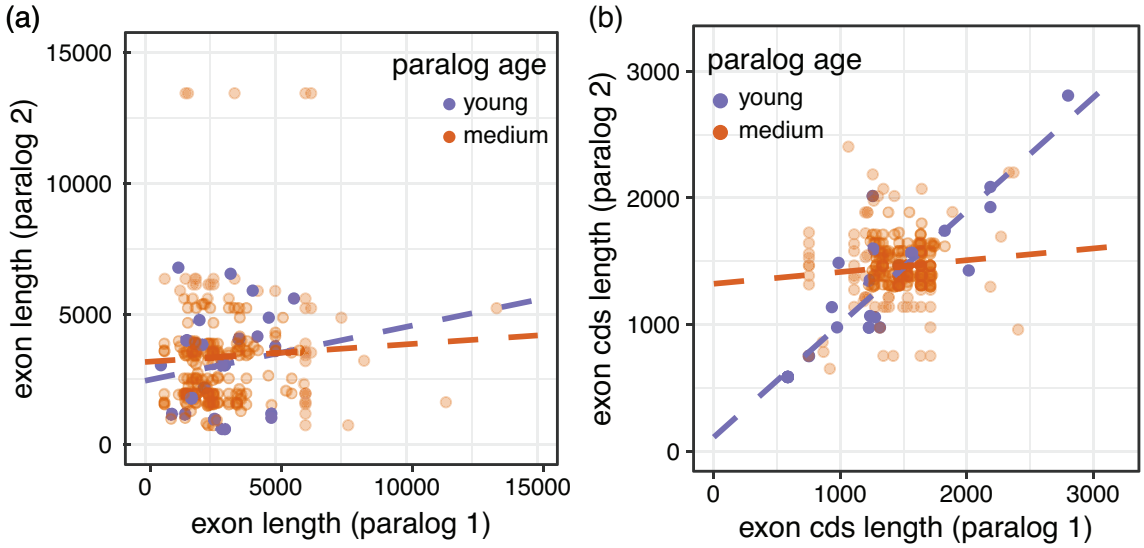


Figure 5.9: **Exon lengths between paralogs are typically only weakly correlated, but coding sequence exon lengths between paralogs have a higher correlation at younger ages.** In **(a)**, we plot the exon (that containing Zn-finger DBDs) lengths of the two paralogs for various paralog pairs, and in **(b)**, we plot the lengths of the coding parts of the exons of the two paralogs for various paralog pairs, colored according to their ages (young – violet, $dS \leqslant 0.15$, medium age – orange, $0.15 < dS \leqslant 0.3$). While the total exon lengths are not significantly correlated across both the age groups (young: $\langle \tau_b \rangle = 0.02, p > 0.05$ for all permuted datasets, medium age: $\langle \tau_b \rangle = 0.13, p > 0.05$ for all permuted datasets), lengths of the coding parts of the exon are strongly correlated for younger pairs ($\langle \tau_b \rangle = 0.82, p < 0.00001$ for all permuted datasets), with the medium pairs uncorrelated ($\langle \tau_b \rangle = 0.005, p > 0.05$ for all permuted datasets).

Finally, we turn our attention to dN/dS measures. We compute this by using the "MS"

(Model Selection) model from the KaKs_Calculator package, which selects best maximum likelihood model based on AIC, from a set of underlying DNA and codon evolution models. A crucial assumption in this model is that all sites of the proteins considered have the same value of dN/dS. While this assumption is very often violated empirically, it is nevertheless a powerful tool to gain insights into patterns of selection. Starting from the time of duplication, dN/dS, as a single number, measures the relative cumulative rate (over time) of non-synonymous changes, compared to synonymous changes. Non-synonymous changes, even though they change amino acid make up of the proteins, can often vary in terms of their effect – they can be neutral, deleterious or advantageous. In addition, this might, and often does, depend on the particular site in the protein. Heterogenous patterns of selective forces on different sites, and heterogenous patterns across sites, might both be averaged out in the final single dN/dS value, masking out interesting patterns. If all sites are evolving neutrally continuously, one would expect dN/dS to be close to 1, and dN/dS to be negatively correlated with dS. If sites are continuously under purifying selection, then dN/dS would be very low, and if sites are continuously under positive selection, $dN/dS > 1$, with dN/dS largely uncorrelated with dS in both cases. A mixture of positive and purifying selection on different sites would result in dN/dS values around or less than 1, making the distinction with neutral evolution difficult [Hahn, 2018]. While we perform a domain-specific analysis, current available softwares of maximum likelihood analysis do not let the user fix the divergence time and calculate conditional dN/dS ratios, which would make a domain-specific analysis complete. We will perform this in future research.

In Fig. 5.10, we plot dN/dS against dS across different age groups (young – violet, $dS \leqslant 0.15$, medium age – orange, $0.15 < dS \leqslant 0.3$, and old – green, $0.3 < dS \leqslant 0.5$) for whole-protein alignments (Fig. 5.10a) and domain-specific alignments (KRAB – Fig. 5.10b, DBDs – Fig. 5.10c). Also, in Fig. 5.10d, we plot Kendall's tau (in color) between dN/dS and dS for paralogs of different age classes, for alignments of while proteins, KRAB domains and Zn-finger DBDs, in a 3×3 cell grid, and also show average values $\langle dN/dS \rangle$, averaged over all paralog pairs inside the age class, inside each cell. For young paralogs (first column, Fig. 5.10d), dN/dS and dS are significantly negatively correlated for whole protein alignment ($\tau_b = -0.38, p < 0.005$) (which could also result from dS in the denominator) but the negative correlations are not significant for KRAB alignment ($\tau_b = -0.27, p = 0.07$) and DBD alignment ($\tau_b = -0.2, p = 0.124$). This indicates that the KRAB domain and the DBD are probably under purifying or positive selection, while the rest of the protein is under neutral selection. A quick look at $\langle dN/dS \rangle$ reveals that for whole and DBD alignments, $\langle dN/dS \rangle \approx 1$, and is less than 1 for KRAB.

So together, for young paralogs, this points to neutral evolution at some sites in both the DBDs and KRAB, mixed with positive selection at some sites in the DBDs, and purifying selection on some sites in the KRAB domain.

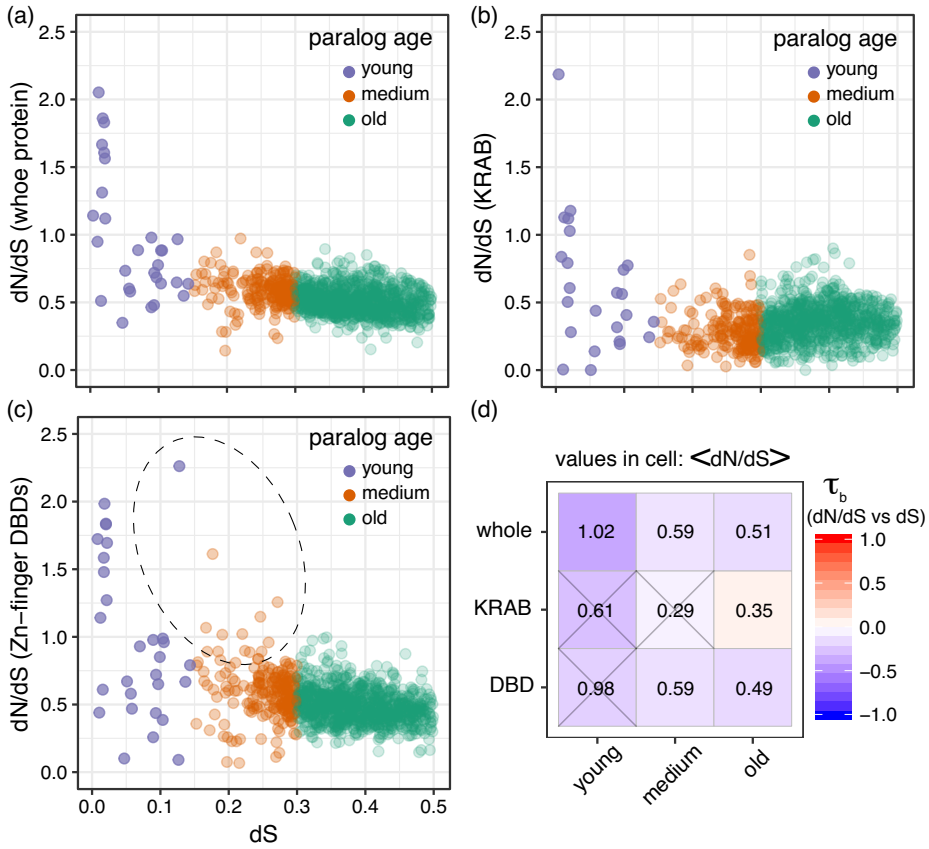


Figure 5.10: dN/dS **vs** dS **reveals possible selection on Zn-finger DBDs and KRAB domains at different time points.** Here we plot dN/dS ratios vs dS, computed for various paralog pairs using the "MS" (model selection) method that uses a AIC criterion to select the best underlying DNA and codon evolution model, using **(a)** a whole protein pairwise alignment, **(b)** an alignment of only KRAB domains of the paralog pair, whenever both TFs have a KRAB domain, and **(c)** an alignment of the sets of Zn-finger DBDs in both the TFs. We classify different paralog pairs by their age: young – violet, $dS \leqslant 0.15$, medium age – orange, $0.15 < dS \leqslant 0.3$, and old – green, $0.3 < dS \leqslant 0.5$. **(d)** We plot Kendall's tau coefficient (in color) between dN/dS and dS for whole protein alignment, KRAB alignment and DBD alignment, for young, medium and old paralogs in a 3×3 cell grid. Correlations that are not significant are marked with a "x" (cross) in the cells. Average values $\langle dN/dS \rangle$, averaged over all pairs in the age class, are written inside each cell. In general, $\langle dN/dS \rangle$ is high for young paralogs, and drops as the paralog age increases, indicating either a mixture of positive+purifying selection, or neutral evolution initially, with purifying selection and neutral evolution in the later stages. This negative relation is also an artefact of dS in the denominator. A larger $\langle dN/dS \rangle$ for DBDs compared to KRAB for young paralogs points to an initial adaptation of the DBD, while an increase in $\langle dN/dS \rangle$ for KRAB from medium to old paralogs indicates a selection on KRAB in the later stages.

Note that $\langle dN/dS \rangle$ decreases for medium age paralogs (second column, compared with the first, Fig. 5.10d), compared to their younger counterparts, marking an increase in the number of sites under purifying selection. Paralogs in this age group perhaps al-

ready diverged to perform specific functions, requiring their amino acids to be maintained. While for KRAB alignment, correlation between dN/dS and dS is not significant ($\tau_b = -0.06, p = 0.22$), for whole protein ($\tau_b = -0.12, p < 0.05$) and DBD alignments ($\tau_b = -0.12, p < 0.05$), the negative correlations, though weak, are significant ($\tau_b = -0.13, p < 0.01$). $\langle dN/dS \rangle < 1$ is low. Also, $\langle dN/dS \rangle$ is lower for KRAB, compared to DBDs, pointing to purifying selection on most sites in KRAB, with a mixture of purifying selection and neutral evolution on most sites in DBDs. However, note that a few pairs of medium age group have large dN/dS values greater than 1 for the DBD alignment (black dashed ellipse in Fig. 5.10c), indicating that some pairs might be under positive selection. On the other hand, for old paralogs, while for whole protein ($\tau_b = -0.15, p < 0.05$) and DBD alignments ($\tau_b = -0.16, p < 0.05$), there are weak but significant negative correlations, for KRAB alignment, the correlation is weakly positive and significant ($\tau_b = 0.07, p < 0.05$). This, together with an increase in $\langle dN/dS \rangle$ for KRAB from medium to old paralogs, perhaps points to a selective phase for the KRAB domain in the older paralogs. This could related to the cooption of old KZNF TFs from transposable element repression to transcriptional regulation.

While such conclusions can be drawn from the site-averaged dN/dS analysis, this might still drown signals of positive selection on a few sites in the sea of numerous other residues that are either under neutral evolution or purifying selection. Hence, we consider site-specific models next.

5.5 Site-specific dN/dS ratios using PAML

To detect positive selection signals from specific functional sites on the Zn-finger proteins, we used site-specific models from the PAML package [Yang, 2007]. First, we selected a few sets of KZNFs based on sequence similarity and the numbers of Zn-finger DBDs. Set 1 with 12 Zn-finger DBDs in each – ZNF695, ZNF723, ZNF626, ZNF117, ZNF430, ZNF431, ZNF479, and ZNF680; Set 2 with 11 Zn-finger DBDs in each – ZNF682, ZNF253, ZNF730, ZNF100, ZNF718, and ZNF141; Set 3 with 13 Zn-finger DBDs in each – ZNF714, ZNF257, ZNF92, ZNF273, ZNF492, ZNF98, ZNF727, and ZNF675; and Set 4 with 15 Zn-finger DBDs in each – ZNF267, ZNF732, ZNF726, ZNF254, ZNF85, ZNF429, ZNF90, and ZNF66. On each set, we first ran ClustalW2 multiple sequence alignment [Thompson *et al.*, 1994] and used PAL2NAL [Suyama *et al.*, 2006] to construct a multiple codon alignment from the protein alignment. We also construct NJ (neighbour-joining) trees from the dS values obtained from KaKs_Calculator analysis.

Using the program "codeml" from PAML, and the multiple alignments and NJ trees, we compute site-specific dN/dS ratios using various underlying models M0, M1a, M2a, M3, M7, M8 and M8a, and also compute the likelihoods of the each model. These models assume that sites belong to various categories, with sites in each category being assigned a single $\omega = dN/dS$ ratio to be estimated. The models differ in the number of categories considered and the constraints on the ω in various categories.

M0 is a basic model that considers that the dN/dS ratio is the same across all sites, M1a is a nearly neutral model with two categories – $\omega_1 < 1, \omega_2 = 1$. M2a is a selection model with three categories – $\omega_1 < 1, \omega_2 = 1, \omega_3 > 1$. M3 is a discrete model with three different categories – $\omega_1, \omega_2, \omega_3$ with no other constraints. M7 is a nearly neutral model with ten categories $\omega_i \sim \texttt{Beta}(\alpha, \beta)$. M8 is a selection model with eleven categories, ten following beta distribution like in M7, $\omega_i \sim \texttt{Beta}(\alpha, \beta)$, with another category $\omega > 1$. M8a is nearly neutral model with eleven categories, with then following beta distribution, $\omega_i \sim \texttt{Beta}(\alpha, \beta)$, with the last category fixed at $\omega = 1$. Various pairs of these models are nested and hence a likelihood ratio test can be used to reject the null hypothesis of neutrality over alternative hypothesis of selection – M0 vs M3, M1a vs M2a, M7 vs M8 and M8a vs M8. In each of these two-model comparisons, the model on the left (M0, M1a, M7 and M8a) assumes neutrality (no selection) while the model on the right (M3, M2a, and M8) additionally allows for positive selection. Because of the nested nature of these pairs of model, we use the test statistic, $-2(\ln \lambda)$, where λ is the likelihood ratio, and compare it with chi-squared distribution (χ^2) with degrees of freedom equal to the additional number of parameters in the model on the right. We show these results in Table 5.1. In all the four sets, for all four model comparisons, we reject the null hypothesis of neutrality in favor of selection with a very low p-value $p < 0.0001$. Further, a modified version of the Empirical Bayes method, called the Bayes Empirical Bayes method [Yang *et al.*, 2005], reveals the specific sites under selection along with their statistical significance. Also, in Figures 5.11 and 5.13, we plot the multiple alignments of the different sets of genes, together with information on the KRAB and Zn-finger DBD domains. We also mark key residues, which are those that contact nucleotides on the DNA via hydrogen bonds to establish specific binding, on the DBDs. A major fraction of positively selected sites inferred from the positive selection model M8 overlap with key residues, signifying that these KZNFs have undergone positive selection to alter DNA binding preferences.

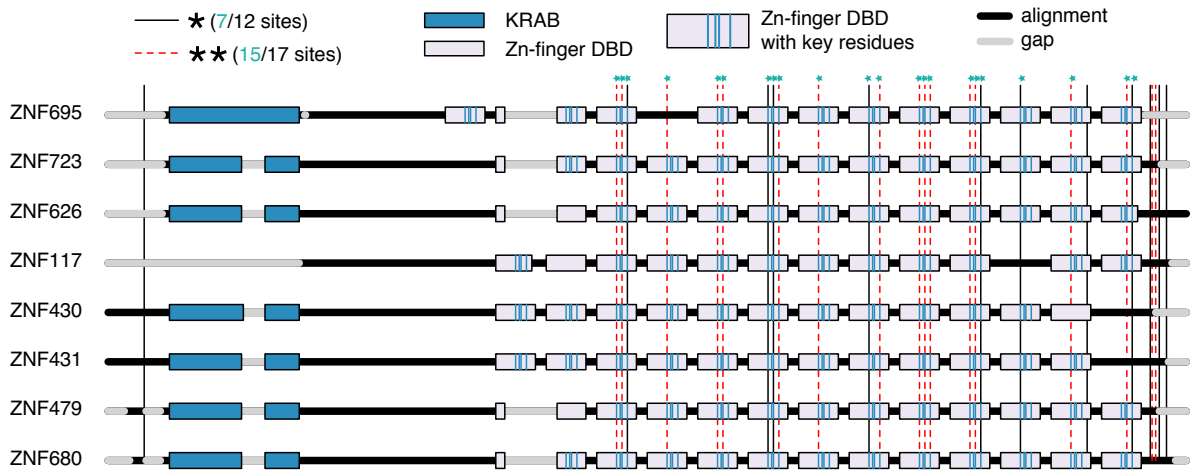


Figure 5.11: (Caption next page.)

Figure 5.11: (Previous page.) **Key amino acid residues on the DBDs, which contact nucleotides on DNA, have undergone positive selection.** Multiple alignments of a few (K)ZNF proteins are shown, together with their domain architecture, indicating the identities and locations of the KRAB and Zn-finger DBD domains. On each Zn-finger DBD, we also show the key amino residues (in blue) that contact the nucleotides on the DNA to establish TF-DNA binding. Site-specific maximum likelihood estimation of dN/dS ratios reveals that a few sites underwent positive selection in their recent evolutionary history, and that a large majority of these sites overlap with the key residues on the Zn-finger DBDs. Significance levels of positively selected sites from Bayes Empirical Bayes analysis [Yang *et al.*, 2005] ("$*$" : $P > 95\%$ (black solid); "$**$" : $P > 99\%$ (red dashed)) are also shown. Out of the 12 positively selected sites at $P > 95\%$, 7 sites exactly overlap with key residues on DBDs (green stars), and out of the 17 positively selected sites at $P > 99\%$, 15 sites exactly overlap with key residues on DBDs (green stars).

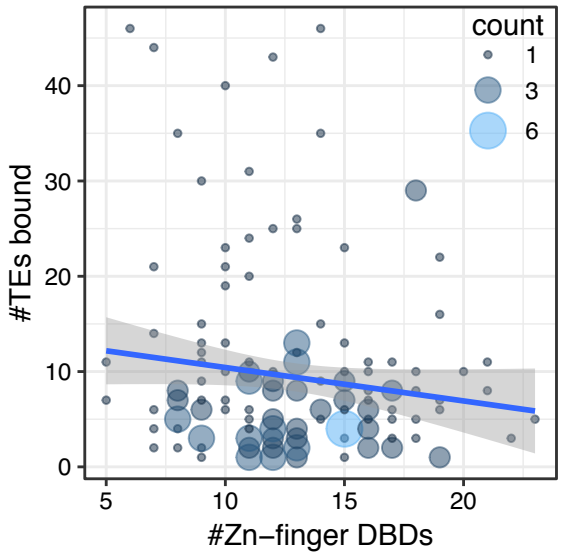


Figure 5.12: **The number of TEs bound is not correlated with the number of Zn-finger DBDs in the TFs.** For a set of KZNF and ZNF TFs, we plot the number of TEs (of various types) bound against the number of Zn-finger DBDs. Somewhat surprisingly, there is no significant correlation between them (Kendall's tau coefficient, $\tau_b = -0.05, p > 0.05$). A possible reason is a strong dependence on the evolutionary history of different sets of TF-TE family pairs.

Data set	Comparison – H0: purifying+neutral vs H1: purifying+neutral +positive selection	$-2(\ln\lambda)$	d.f.	χ^2	Sites under selection according to M8
Set 1 (ZNF695) Fig. 5.11	M0 vs M3	332.96	4	$p < 10^{-4}$, reject H0: only neutral	→ multiple-alignment: 29 sites positively selected → 7 key residues out of 12 with significance $p < 0.05$ → 15 key residues out of 17 with significance $p < 0.001$ → ZNF267: 23/515 (4.47%) selected sites
	M1a vs M2a	90.60	2		
	M7 vs M8	114.75	2		
	M8a vs M8	85.78	1		
Set 2 (ZNF682) Fig. 5.13a	M0 vs M3	161.98	4	$p < 10^{-4}$, reject H0: only neutral	→ multiple-alignment: 13 sites positively selected → 8 key residues out of 10 with significance $p < 0.05$ → 2 key residues out of 3 with significance $p < 0.001$ → ZNF267: 13/498 (2.6%) selected sites
	M1a vs M2a	29.30	2		
	M7 vs M8	33.85	2		
	M8a vs M8	28.89	1		
Set 3 (ZNF714) Fig. 5.13b	M0 vs M3	337.74	4	$p < 10^{-4}$, reject H0: only neutral	→ multiple-alignment: 23 sites positively selected → 6 key residues out of 10 with significance $p < 0.05$ → 12 key residues out of 13 with significance $p < 0.001$ → ZNF714: 23/555 (4.14%) selected sites
	M1a vs M2a	105.87	2		
	M7 vs M8	125.17	2		
	M8a vs M8	107.42	1		

Table 5.1: (Continued in the next page.)

Set 4 (ZNF267) Fig. 5.13c	M0 vs M3	388.45	4	$p < 10^{-4}$, reject H0: only neutral	→ multiple-alignment: 25 sites positively selected
	M1a vs M2a	86.58	2		→ 10 key residues out of 16 with significance $p < 0.05$
	M7 vs M8	112.14	2		→ 7 key residues out of 9 with significance $p < 0.001$
	M8a vs M8	87.24	1		→ ZNF267: 25/743 (3.36%) selected sites

Table 5.1: **Summary of site-specific dN/dS analysis on four different sets of KZNF genes.** M0 is a basic model that considers that the dN/dS ratio is the same across all sites, M1a is a nearly neutral model with two categories – $\omega_1 < 1, \omega_2 = 1$. M2a is a selection model with three categories – $\omega_1 < 1, \omega_2 = 1, \omega_3 > 1$. M3 is a discrete model with three different categories – $\omega_1, \omega_2, \omega_3$ with no other constraints. M7 is a nearly neutral model with ten categories $\omega_i \sim \texttt{Beta}(\alpha, \beta)$. M8 is a selection model with eleven categories, ten following beta distribution like in M7, $\omega_i \sim \texttt{Beta}(\alpha, \beta)$, with another category $\omega > 1$. M8a is nearly neutral model with eleven categories, with then following beta distribution, $\omega_i \sim \texttt{Beta}(\alpha, \beta)$, with the last category fixed at $\omega = 1$. While M0, M1a, M7 and M8a are null models, which assume purifying selection and neutrality (no positive selection), their partners M3, M2a and M8 are the alternative models, additionally allowing for positive selection. Various pairs of these models are nested and hence a likelihood ratio test, by comparing $-2(\ln \lambda)$ with a χ^2 distribution (λ is the likelihood ratio), can be used to reject the null hypothesis H0 of neutrality (M0, M1a, M7, M8a) over alternative hypothesis H1 of selection (M3, M2a, M8) – M0 vs M3, M1a vs M2a, M7 vs M8 and M8a vs M8. Further, a modified version of the Empirical Bayes method, called the Bayes Empirical Bayes method [Yang *et al.*, 2005], reveals the specific sites under selection along with their statistical significance. In all the four sets, for all four model comparisons, we reject the null hypothesis of neutrality in favour of selection with a very low p-value $p < 0.0001$. Further, a major fraction of positively selected sites overlap with key residues, signifying that these KZNFs have undergone positive selection to alter DNA binding preferences.

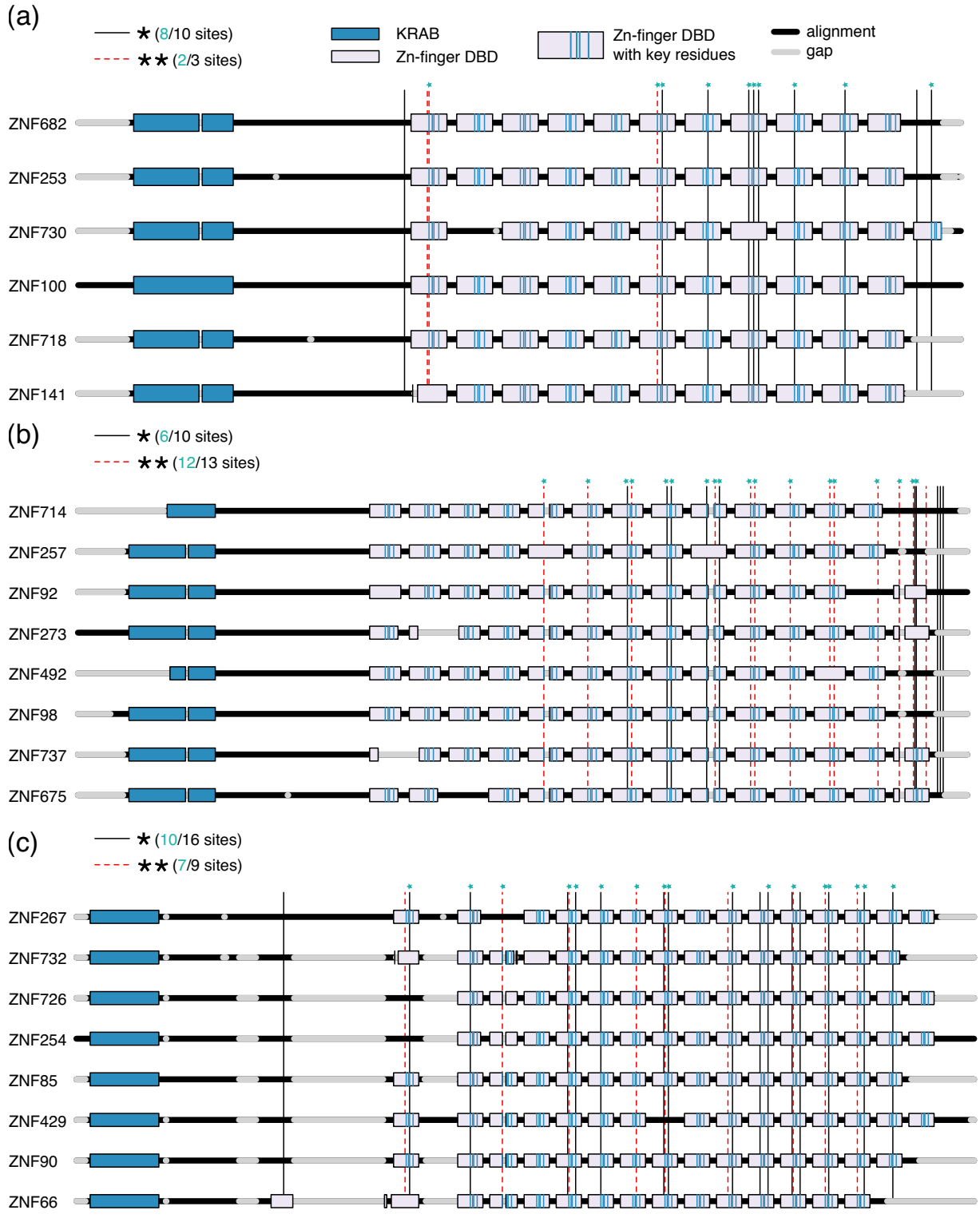


Figure 5.13: (Caption next page.)

Figure 5.13: (Previous page.) **A few other groups of (K)ZNF proteins also reveal a picture of significant positive selection at a few sites, primarily on the key amino acid residues on the DBDs.** Multiple alignments of a few other sets of (K)ZNF proteins are shown in **a, b** and **c**, together with their domain architecture, indicating the identities and locations of the KRAB and Zn-finger DBD domains. On each Zn-finger DBD, we also show the key amino residues (in blue) that contact the nucleotides on the DNA to establish TF-DNA binding. Site-specific maximum likelihood estimation of dN/dS ratios reveals that a few sites underwent positive selection in their recent evolutionary history, and that a large majority of these sites overlap with the key residues on the Zn-finger DBDs. Significance levels of positively selected sites from Bayes Empirical Bayes analysis [Yang *et al.*, 2005] ("$*$" : $P > 95\%$ (black solid); "$**$" : $P > 99\%$ (red dashed)) are also shown. Positively selected sites overlapping with key residues on the DBDs are marked by green stars. **(a)** Out of the 10 positively selected sites at $P > 95\%$, 8 sites exactly overlap with key residues on DBDs, and out of the 3 positively selected sites at $P > 99\%$, 2 sites exactly overlap with key residues on DBDs. **(b)** Out of the 10 positively selected sites at $P > 95\%$, 6 sites exactly overlap with key residues on DBDs, and out of the 13 positively selected sites at $P > 99\%$, 12 sites exactly overlap with key residues on DBDs. **(c)** Out of the 16 positively selected sites at $P > 95\%$, 10 sites exactly overlap with key residues on DBDs, and out of the 9 positively selected sites at $P > 99\%$, 7 sites exactly overlap with key residues on DBDs, with a few of the other positively selected sites occurring just adjacent to key residues on the DBDs.

5.6 KZNFs and TEs

A few studies have pointed out that KZNFs, Zn-finger TFs with a KRAB domain, are involved in repression of transposable elements [Yang *et al.*, 2017], and that they have coevolved together with TEs [Jacobs *et al.*, 2014]. We investigate this possibility by obtaining the TEs bound by various KZNFs using ChIP-exo data from Imbeault 2017 [Imbeault *et al.*, 2017], and estimating their age by querying from RepeatMasker, the copy number data of TEs in various representative species from different taxonomical nodes. While such an ageing of TEs is not very fine-grained and might contain errors as we use only representative species that might have specially gained or lost a TE family, we undertake it as a first step in discerning KZNF-TE coevolutionary patterns.

First, in Fig. 5.12, we ask if the number of TEs (of different types) bound by KZNFs depend on the number of DBDs in them. It might be the case that KZNFs with smaller number of DBDs bind to more variety of TEs. Somewhat surprisingly, we find that it is not the case, and that there is no significant correlation (Kendall's tau, $\tau_b = -0.05, p > 0.05$) between the number of TEs bound and the number of DBDs. One possible reason for this counter-intuitive finding is a strong dependence on the evolutionary histories of KZNF-TE binding patterns, restricting the usage of a new KZNF to a related TE, independent of whether it has lost or gained a few DBDs.

	On the diagonal	Upper triangle	Lower triangle
Data	43	17	57
Expected	9.75	49.56	49.56

Table 5.2: **Age of duplication is rarely older than the age of the typical new TE to which the TFs adapted.** In this table, we show the number of paralog pairs in Fig. 5.15 that fall exactly on the diagonal, in the upper triangle and in the lower triangle. In the first row "Data", we show the empirically observed paralog counts, and in the second row "Expected", we show the expected numbers if there was no relation at all. Diagonal corresponds to the situation where the age of duplication (age of TF) coincides with the age of the typical new TE on the taxonomic scale. Upper triangle corresponds to the situation in which the typical new TE is older than the TFs, and the lower triangle corresponds to the situation in which the typical new TE is younger than the TFs. The diagonal is over-represented in the data, while the upper triangle is under-represented, with the lower triangle being slightly over-represented.

Next, in Fig. 5.14, we compare the typical ages of the TEs bound by the pairs of paralogous KZNFs. These typical ages of TEs very often fall close to the diagonal, meaning that paralogous pairs share binding to TEs of a similar age. A Fisher's Exact Test indicates that the typical ages of the TEs bound by paralogs are related to each other ($p < 0.001$). However, as the ages are ranked, we also compute Kendall's tau ($\tau_b = 0.28, p < 0.001$), which reveals a strong correlation between the typical ages of TEs bound by paralogs, meaning they adapted to TE binding on similar timescales. However, there are some pairs of paralogs that are off-diagonal, perhaps pointing to KZNF adaptation to newer TEs. Out of 117 pairs, 43 exactly fall on the diagonal (36.75%, compared to expected 8.33%).

In Fig. 5.15, we compare the typical TE ages with the age of the duplication event (paralog age), obtained from a reconciliation of the gene tree with the species tree. We consider the younger among the typical ages of the TEs bound by the two paralogs, and compare it with paralog age. The younger among the typical ages of the TEs often denotes a major set of new TEs. First, a Fisher's Exact Test reveals that the younger among the typical ages of the TEs is related to the age of duplication ($p < 0.0001$). Again, as the ages are ranked, we also compute Kendall's tau coefficient, ($\tau_b = 0.19, p < 0.05$), from which we see that the younger among the typical ages of the TEs is mostly either of a similar age or is newer compared to the paralog age. This points strongly to KZNF adaptation to bind new novel TFs that arise in organisms, both by using new paralogs (close to the diagonal), as well as by reusing old TFs (lower triangle). See Table 5.2 for actual numbers.

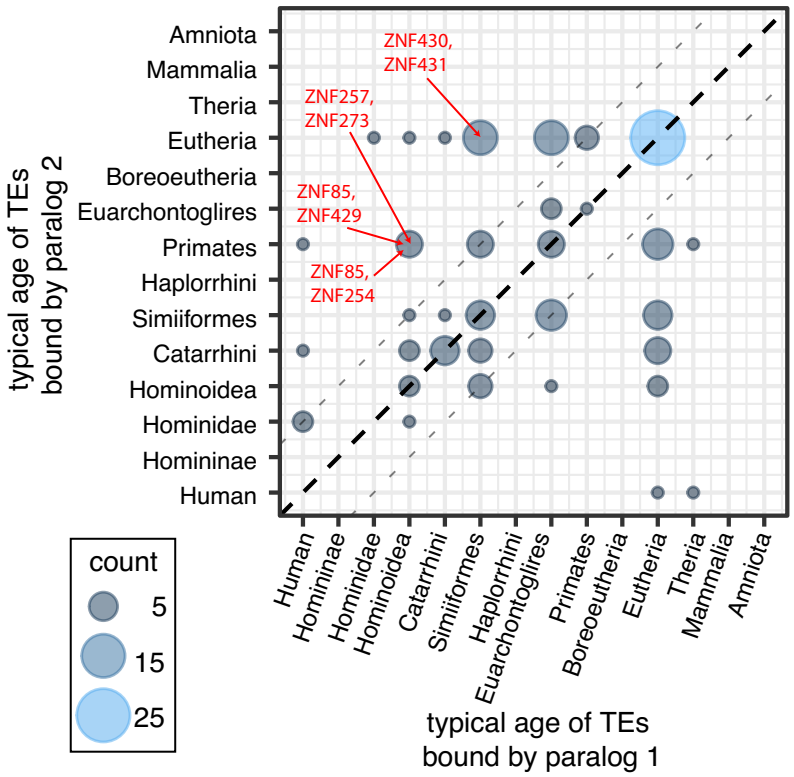


Figure 5.14: **The typical ages of TEs bound by paralog TFs are positively correlated.** For various pairs of paralog KZNF TFs, we plot the typical ages of TEs bound by each of the TFs against each other. We use a taxonomical scale for the age of TEs, estimated by the species that they are found in. A significant number of paralog pairs fall on and around the diagonal (within the grey small dashed lines around the thick black dashed diagonal) – the typical ages of the TEs bound by paralog TFs is similar (Kendall's tau, $\tau_b = 0.28, p < 0.001$). However, there are some paralog pairs that are located off the diagonal, meaning that one (or both) of the paralogs underwent adaptation to bind different TEs than its partner. A few pairs from the sets considered in Sec. 5.5 are pointed out in red.

5.7 Discussion

The question of the relative importance of transcription factor evolution in comparison with regulatory sequence evolution, towards phenotypic divergence, is a long-standing one that has attracted diverse views [King and Wilson, 1975; Gilad *et al.*, 2006; Carroll, 2005; Wray *et al.*, 2003]. Recent studies have painted a picture of coevolution of transcription factors and regulatory sequences as the correct framework to resolve this problem [Gordon and Ruvinsky, 2012; Friedlander *et al.*, 2017]. In Chapter 4, we explored a theoretical framework of the evolution of a simple regulatory network that involves transcription factors sensing specific upstream signals and responding by activating specific required target genes. In particular, we focused on the evolution of such a network after a duplication event of the gene coding for the transcription factor, and uncovered the evolutionary steady states reached by the system under various conditions, and the evolutionary

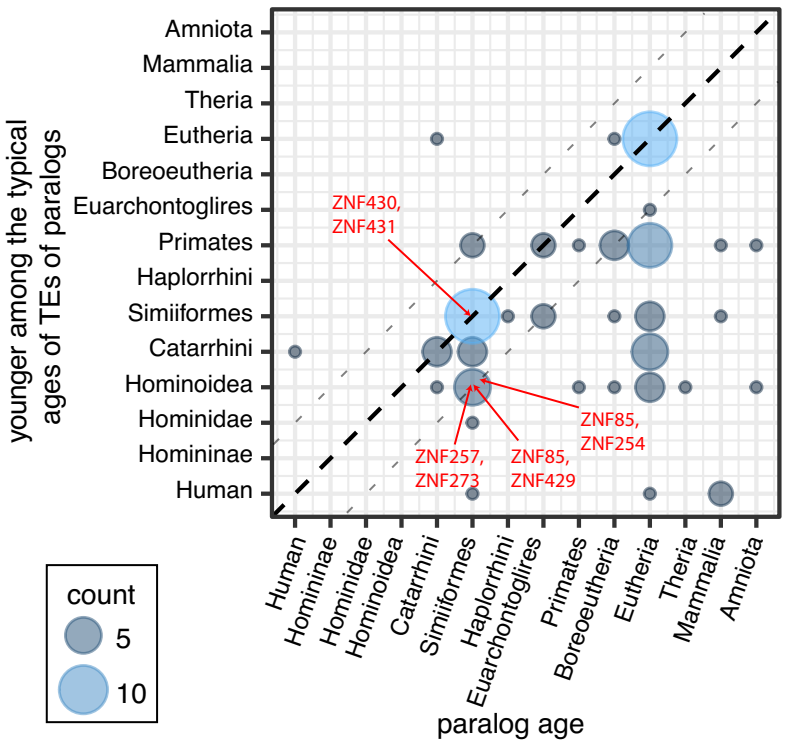


Figure 5.15: **Paralog KZNF TFs adapt to bind TEs that newly arose.** For various pairs of paralog KZNF TFs, we plot the younger among the typical ages of TEs (y-axis) bound by each of the TFs against the age of the duplication event (x-axis). We use a taxonomical scale for both the age of TEs, estimated by the species that they are found in, and for the age of duplication, obtained by comparing the gene-tree with the species-tree. A significant number of paralog pairs fall around the diagonal and below it. Kendall's tau has a value of $\tau_b = 0.19$ ($p < 0.05$). This signifies that duplication age is at least older than the age of the typical new TE, pointing to KZNF adaptation to bind novel TEs that arise in organisms, and a potential reuse of old TFs to repress new TEs. A few pairs from the sets considered in Sec. 5.5 are pointed out in red.

pathways involved and their timescales. This particular question of transcription factor evolution after duplication is biologically relevant because in all organisms, the repertoire of transcription factors is comprised in a few families of paralogous transcription factors (those that arose by gene duplication). As the fraction of space occupied by functional proteins (and DNA sequences in general) is infinitesimally small in the entire space of possible proteins (DNA sequences), gene duplication has been a dominant force in, not just the expansion of transcription factor families, but also the expansion of the genetic repertoire of organisms in general. Gene duplication provides the organism with new genetic material that is already shaped to be functionally relevant, and because of the redundancy with two copies of the genes, the organism can now evolve a new gene with different functions.

It is this advantage with the mechanism of gene duplication that has led to an expansion in many families of transcription factors, resulting in a complex network of gene regulatory interactions that result in rich phenotypes. A paradigmatic example is the Hox gene

family, which codes for homeodomain transcription factors that control the body plan of metazoan embryos. The largest family of transcription factors in humans and other animals is the C2H2 Zn-finger family of transcription factors [Lambert *et al.*, 2018]. They are involved in various gene regulatory pathways, chromatin remodeling [Kim *et al.*, 2015] and repression of transposable elements (TEs) [Yang *et al.*, 2017]. Each Zn-finger TF contains an array of variable number of Zn-finger DNA-binding domains, each of which has a certain binding specificity towards DNA sequences depending on its amino acid composition. Studies have also shown that by changing certain amino acids in the DBD, Zn-finger DBDs can be tuned to bind the whole range of possible DNA sequences [Najafabadi *et al.*, 2015]. However, the process by which new Zn-finger TFs arise, and take up their functions, is not well understood. What are the selective pressures that act on duplicated Zn-finger TFs and how do they undergo selection to evolve towards their new functions? In the case of KZNFs, what are the selection pressures acting on their DBDs and KRAB domains, that enable them to adapt to the constantly changing TE landscape? Do old KZNFs along with their dormant TE partners get coopted to involve in transcriptional regulation?

In this chapter, we used a bioinformatics approach to scratch the surface of such questions. By using sources like EnsEMBL v93 [Cunningham *et al.*], RepeatMasker [Smith *et al.*, 2016], ChIP-exo data [Imbeault *et al.*, 2017], and softwares such as KaKs_Calculator [Zhang *et al.*, 2006], ClustalW2 [Thompson *et al.*, 1994], PAL2NAL [Suyama *et al.*, 2006] and PAML [Yang, 2007], we built a pipeline (Fig. 5.1) to probe the evolution of Zn-finger TFs in humans, with a focus on pairs of annotated paralog TFs. After teasing apart the correlations among various genomic features of Zn-finger TFs, we computed dN/dS ratios to infer selection patterns on the paralogs. Finally, by connecting the ages of TEs and TFs, we discovered signs of KZNFs being retained after duplication to bind to new TEs.

First, we extracted a list of annotated Zn-finger TF genes and their proteins from EnsEMBL, along with various genomic features like their chromosomal location, number of Zn-finger DBDs etc. and looked at the correlations among these genomic features. We found that Zn-finger TFs are present in clusters on chromosomes, and that chromosomal lengths have no significant correlation with the number of Zn-finger genes on them (Fig. 5.2), pointing out their origin via gene duplication. We found that Zn-finger TFs have a variable number of DBDs in them (Fig. 5.3), starting from only 1 DBD till a maximum of 37 (which would bind DNA sequences longer than 100bp), with a mean of 10 DBDs (average DNA sequence bound $\sim$ 30bp) per protein. We then found a strong correlation between the number of exons and the number of coding exons (Fig. 5.4), indicating that most exons act as coding exons in Zn-finger TFs. We also found that a sizeable number of Zn-finger TFs have only 1 coding exon, some of which could have arise by retrotransposition-based duplication. Next, we found no significant correlation between the number of exons and the number of DBDs (Fig. 5.5), meaning that DBDs are most often coded in a single exon. All pairs of correlations are depicted in Fig. 5.6, which also indicates that the number of DBDs correlates strongly with the length of the coding part of the exon that codes for the DBDs. Further, we found that young paralogs

have similar number of DBDs, but paralogs of medium age differ in the number of DBDs (Fig. 5.7). However, the maximum number of DBDs in the pair of paralogs, has a smaller variance for medium age paralogs than young paralogs, implying evolution by the loss of DBDs. Further evidence is seen in Fig. 5.8b, where the difference in number of DBDs between paralogs is strongly correlated with the difference in lengths of the coding regions of the exons. In summary, this paints a picture of evolution not only by point mutations in the DBDs that alter DNA binding preferences, but also by loss (or sometimes gain) of DBDs by their removal from the coding part of the exon. The molecular mechanistic pathways by which this is achieved is not clear; a few possibilities are recombination, premature stop codons and shifting of intron-exon boundaries.

Next, we computed dN/dS, together with dS for various paralog pairs of different age groups, using alignments of whole proteins, only DBDs, and only KRAB domains. We used "MS" (Model Selection) method from KaKs_Calculator to select the best maximum likelihood model (AIC) of underlying DNA and codon evolution. Correlations between dN/dS and dS for paralogs of different age groups, and the average values $\langle dN/dS \rangle$ point to an initial (after duplication) selection on the Zn-finger DBDs, followed by a period of purifying selection, before a possible selection on the KRAB domain (Fig. 5.10). The initial adaptation of the DBD helps the new TF copy to bind to its new regulatory sequence, and in the case of a KZNF, allows it to coevolve with a new TE and repress it successfully. The later selection on KRAB could be related to the cooption of the KZNF-TE (if dormant) pair to act as a transcriptional regulatory link. As these models assume that all sites across the protein subsequence aligned have the same dN/dS, they miss any further heterogenous selection patterns that might exist between sites.

Hence, we then used site-specific models from the PAML package on a few KZNF sets, and found (a) that there is significant evidence of selection acting on KZNF TFs, and (b) that positive selection primarily acted on the key residues (amino acids contacting DNA to establish sequence-specificity) on the DBD, implying a quick adaptation to bind new sequences, possibly TEs. To check for this latter possibility, we inferred taxonomic ages of various TE types from their copy number data in various representative species at various taxonomical levels. By combining ChIP-exo data, TE age data and paralog age data (dS, as well as taxonomic age from EnsEMBL via reconciliation of gene trees and the reference species tree), we inferred (a) that the typical ages of TEs bound by paralog KZNFs are correlated with each other, and (b) that the paralogs are at least as old as the younger among the typical ages of the TE bound by them. This implies that often new KZNF TFs are often retained after duplication to immediately adapt to bind to new TEs, but that old KZNFs are also sometimes reused to repress new TEs.

Together, this informs us that Zn-finger TFs that are retained as functional copies in the human genome, have undergone specific changes that played a role in their retention. They evolve, apart from loss (and gain) of DBDs, but also probably by an early selection in their DBDs to bind new sequences. In particular, KZNFs have undergone bursts of positive selection at key residues on the DBDs, presumably to track and coevolve with the constantly changing landscape of TEs. There is also evidence for reuse of old KZNF

TEs in repression of new TEs, and also possible cooption of older KZNFs in roles other than TE repression by a late selection on the KRAB domain.

In future research, we will undertake multiple directions from this preliminary analysis. First, while most recent ZNF paralogs in the form of KZNF paralogs seem to be involved in TE repression, we will form bridges between a TF-TE coevolutionary analysis with a gene regulation based TF-BS coevolutionary analysis to relate to the predictions from the coevolutionary theory of the previous chapters. We will investigate the interplay between TF evolution via DBD loss and point mutations to elucidate the possibility of promiscuous pathways as lower number of DBDs might result in more promiscuous TFs. Further, we will use ChIP-seq to infer DBD divergence, and RNA-seq data to understand upstream expression, to test specific predictions of the TF-BS coevolutionary model and to understand genome-wide crosstalk?

Chapter 6

Coevolution of transcription factors and their binding sites in sequence space

In this final chapter, I will summarize the main questions tackled and the main findings of this thesis. This chapter is written in such a way as to allow two kinds of readers understand the thesis better - (a) those who do not have enough time to read the entire thesis on its own, and hence are looking for a concise version that refers to particular important results, and (b) those who prefer reading a concise and compact version of the thesis first, with a focus on the main results, before jumping into the little details and further sub-results. In fact, as I wrote in the introductory Chapter 0, any reader might be benefitted maximally by choosing to read this chapter after the introductory chapter, occasionally referring to particular results if necessary, and then going through the thesis in detail, starting from Chapter 1.

6.1 TFs recognize specific DNA sequences

The biological cell is an extremely crowded dynamic environment with atoms and molecules of multiple types moving around inside the cell and constantly interacting with each other. However, the cell's functioning and survival depend on precise biochemical pathways that result from some of these interactions. How do we reconcile the picture of a crowded cell with these precise schemes of biochemical reactions? Biological molecules use "molecular recognition" to ensure that only correct pairs of biological molecules (cognate pairs) interact to result in successful outcomes, while ensuring that numerous possible incor-

rect pairs of biological molecules (non-cognate pairs) result in unsuccessful interactions. Cognate pairs of molecules recognize each other using structural and biochemical features that result in hydrogen bonds, hydrophobic forces, van der Waals forces, π–π interactions and electrostatic interactions – much like a lock and its correct key that "fit" together well. Molecular recognition is ubiquitous, and plays an important role in conceivably every cellular process. For instance, molecular recognition is present at all levels in a typical signalling pathway: ligand–receptor interactions at the cell membrane that help the cell sense correct signals when needed, protein-protein interactions between signalling molecules in the cytoplasm that help the cell transmit signalling information to specific parts of the cell, protein-protein and protein-DNA interactions inside the nucleus that help integrate these various signals and regulate gene expression.

A particular class of DNA-binding proteins called transcription factors (TFs) recognize specific sequences of DNA, and upon binding to genomic loci containing these specific sequences, regulate the expression of nearby associated genes (intuitively, switching ON or OFF of genes). TFs are crucial players in transcription, the conversion of genetic information from DNA to RNA. RNAs often get translated into proteins when the transcribed loci correspond to genes; these RNAs and protein products are used by the cell in various cellular processes. Some of these protein products are themselves TFs, resulting in a set of complex interactions among various genes. Hence, TFs, via these sets of complex interactions among themselves and with other genes, help the cell control the spatiotemporal expression of various genes. These complex interactions can be pictured as a gene regulatory network (GRN) – a network of genetic interactions in which "nodes" correspond to various genes, a subset of which result in TF proteins, and "edges" between nodes are defined by genetic interactions. If two genes are connected by an edge, then the TF protein product of one gene is involved in the regulation of the expression of the other gene.

Transcription factors contain special protein domains called DNA-binding domains (DBDs) that establish sequence-specific DNA-binding via the specificity of hydrogen bonds between particular amino acids in the DBD and the nucleotides on the DNA. Each transcription factor, depending on the set of amino acids in its DBD, makes maximal hydrogen bonds with a specific DNA sequence, called the consensus sequence. Different TFs, depending on the amino acids in their DBDs, have different consensus sequences. Also, each TF is not perfectly specific in DNA-sequence binding – apart from the consensus sequence, each TF binds a variety of other DNA sequences with varying affinities, with the affinities roughly depending on the similarity of the DNA sequence with the consensus sequence. Given these properties of TF-DNA binding, different TFs can often bind to the same DNA sequence with non-negligible affinities – the consensus sequences of different TFs can be similar to each other, TFs can be less specific in their DNA-sequence preference. ChIP data has shown that TFs *in vivo* bind to a large number of DNA sequences at various loci with varying affinities. The few strongly bound DNA sequences tend to be transcribed and are present in known regulatory loci, while a massive number of weakly bound DNA sequences tend to be present in closed chromatin and are not transcribed. However, this can lead to segregation of TF molecules at non-functional

sites and hence prevent their binding to functional loci. On the other hand, a substantial number of weakly bound DNA sequences are often weakly transcribed, leading to substantial spurious transcription. These spuriously transcribed RNA can interfere with cellular functions, and get translated into unwanted proteins if they happen to be transcribed from the coding region of an unwanted gene.

Such non-cognate binding of TFs to DNA sequences is termed crosstalk, and can often result in the expression of unwanted genes, leading to interference with cellular processes. Even if unwanted genes are not expressed, such non-cognate binding of TFs can result in spurious RNA from other DNA loci, and also alter the chromatin state of the genome by recruiting chromatin remodelers to spurious loci. Analytical analysis (Chapter 1) reveals that the fraction of long DNA sequences that do not contain spurious binding sites within them is indeed very low, and decreases as both the length of DNA sequence and the number of TFs to be avoided increase (Fig. 1.6). To understand the role of crosstalk in transcriptional regulation, we used the thermodynamic mismatch model of TF-DNA binding to quantify such crosstalk between TFs and binding sites of target genes, and investigated the limits it places on the design of gene regulatory networks in Chapter 2.

In our basic setup, we have M genes in total, each of which is associated with a binding site corresponding to a particular TF. In different environments faced by the cell, different sets of Q out of these M genes are required to be ON, with the corresponding particular Q TF regulators present in each of these environments. Over time, the cell faces different environments, and in each environment, it deploys the correct set of Q TFs that are required to switch the necessary genes ON, thereby ensuring that the cell navigates into the correct cellular state (Fig. 4.1). However, in each of these environments, because of two factors – the limited specificity of TFs and the similarity of consensus sequences of different TFs, there is often non-cognate binding between TFs and binding sites, leading to crosstalk. By defining crosstalk states (Fig. 2.4a) as those in which (a) the binding sites of required genes are not bound by their cognate TF (either due to being unbound, or being bound by one of the $Q-1$ non-cognate TFs), and (b) the binding sites of unwanted genes are bound by one of Q non-cognate TFs, we use the thermodynamic model of TF-DNA binding, and a mean-field like assumption over various possible environments (Sec. 2.4), to compute the overall crosstalk, X (Eq. 2.8) – average fraction of time a randomly selected binding site in the cell is in a crosstalk state.

Our formulation of the model allows us to compute crosstalk, X, as a function of various parameters that have biological relevance, like binding site length L, total number of genes M, number of genes to be ON Q, energetic mismatch penalty (TF specificity) ϵ, and TF concentration C. We compute optimal crosstalk, X^*, which is defined as crosstalk at optimal TF concentration C^* obtained by minimizing X over C, as the latter is hard to estimate empirically and is highly variable (Eq. 2.10-2.12). By doing this, we establish a lower bound on crosstalk as a function of the other parameters. For a fixed total number of genes M, we find that optimal crosstalk, X^*, depends on two important parameters – the number of genes to be ON, Q, and the binding site similarity, S. The latter parameter,

S, is an effective parameter that captures how similar the various binding sites in the genome are with each other. Binding site similarity S depends on L and ϵ (Fig. 2.2), and on how the binding site sequences are arranged in the sequence space – with random arrangement leading to higher similarity than optimal arrangements that ensure binding site sequences are sufficiently dissimilar from each other (Fig. 2.12).

The phase diagram of X^* against Q and S (Fig. 2.4b, c) shows three distinct phases, depending on whether C^* is 0 (phase I), ∞ (phase II) or finite (phase III). By focusing on phase III, we see that X^* increases as binding site similarity S increases. This is intuitive as the non-cognate binding increases as S increases, leading to larger crosstalk. This also means that, for a given L and ϵ, optimal arrangement of binding site sequences in sequence space results in lesser crosstalk than a random arrangement. On the other hand, X^*, as a function of Q, exhibits more complex behaviour, by taking a maximum at some intermediate value of Q. We see that crosstalk is larger when there are many TF molecules available and many unwanted binding sites present, resulting in extensive non-cognate binding. This does not occur at both small and large Q, but occurs at intermediate Q, resulting in a maximum at some intermediate value. Further, we considered complex variants of the basic model in which regulation is not 1-to-1 (1 TF for 1 gene), but is 1-to-many – every TF regulates Θ genes (Sec. 2.9), or is either many-to-many (combinatorial) – different combinations of TFs regulate different genes (Sec. 2.8, Fig. 2.18). Because of reduced number of TF regulators in both these models, there is often a significant reduction in crosstalk.

Where do real organisms fall on this phase diagram? By assuming random arrangement of binding site sequences, we empirically estimated binding site similarities of the genomes of various organisms (Fig. 2.16) by using available position-weight matrices (PWMs). The overall picture conveyed by this basic model is that crosstalk is high in eukaryotes, often exceeding 0.25, meaning that a randomly selected binding site is in crosstalk states at least 25% of the time. However, as shown by the comparison between random arrangement of binding sites and their optimal arrangement, evolution can fine-tune TF binding preferences and binding site sequences to ensure that they are sufficiently dissimilar to result in reduced crosstalk. Motivated by this, by building on a general model of TF-BS coevolution from Chapter 3, we investigated a model that corresponds to a typical occurrence of such co-evolution of TFs and binding sites in Chapter 4.

We considered the following scenario to investigate TF-BS coevolution, and the role of crosstalk in the corresponding evolutionary dynamics. Motivated by the presence of TFs in paralogous families, we considered the coevolution of TFs and their binding sites after a gene duplication event of gene coding for the transcription factor. We consider a framework in which TFs transmit information from upstream signals to downstream target genes (Fig. 4.1a), by sensing the presence of particular signals via their signal sensing domain, and binding to specific binding site sequences via their DNA binding domains. We have two upstream signals, whose individual presence requires the switching ON of two specific target genes respectively. Initially, we have a single TF gene whose TF protein product senses both the signals and by binding to the binding sites of the

target genes, switches ON both the genes. After a gene duplication event, there are now two identical copies of the TF, both of which sense both signals and regulate both genes (Fig. 4.1b). Subsequent mutations occur either in the binding sites, the TF consensus sequences, or the TF signal sensing domains (Fig. 4.1c), and selection acts to ensure correct signalling transmission between upstream signals and downstream target genes (Fig. 4.2). The main question we asked is how the evolutionary dynamics proceed after duplication as the two copies of the TFs and the binding sites coevolve together, and in particular, when and how quickly the two TFs specialize in function – with each regulating one of the two signalling pathways separately.

We consider the coevolution problem in sequence space (Fig. 4.1d). By specifying as the genotype – the binding site sequences, TF consensus sequences, and the TF signal sensing preferences via discrete alleles, we end up with a huge state space that presents computational challenges. However, the thermodynamic mismatch model of TF-DNA binding allows us to coarse-grain (Fig. 4.3) and treat the coevolution problem in the space of mismatches (between TFs and BSs, and between TFs). To answer questions about the probability and timescales of evolutionary outcomes like TF specialization, we also map the whole set of genotypes onto a few functional "macrostates", depending on how the TFs jointly transmit information from the upstream signals to downstream target genes (Sec. 4.2.5 and (Fig. 4.10a). Each of these macrostates is composed of different numbers of underlying genotypes, and vary in fitness across different environments (Fig. 4.10b). Note that successful specialization involves the coordinated coevolution of both TFs, in both their signal sensing domains and their DNA binding domains.

At evolutionary steady state, we find that the dominant evolutionary outcome (macrostate) varies as a function of overall selection strength to maintain regulation, Ns, and signal correlation, ρ, which quantifies how correlated the signals are in the set of environments the cell faces over its lifetime (Fig. 4.11a). At low selection strength, the cell fails to evolve a working regulatory network and no information is passed from the signals to genes. As selection strength increases, the dominant macrostate is that of partial regulation – networks in which information is partially transmitted correctly between signals and genes. At strong selection strength, TFs specialize in information transmission – one TF per one signalling pathway, with no crosstalk. However, whether specialization is the dominant macrostate also depends on the signal correlation ρ, with specialization occurring as long as the signals are not too strongly correlated. At strong signal correlations, networks in which one TF "is lost" evolve – only one TF transmits information from signals to genes, like in the pre-duplicated state. This is because at the strong signal correlations, either both signals are only rarely present individually, the need for two specialized pathways does not arise. In summary, specialization occurs at strong selection strengths when signals are not too strongly correlated – the evolutionary outcome depends on the effective dimensionality of the signal space. Further, importantly, we found that specialization depends crucially on selection against crosstalk – it occurs only when there is strong selection against crosstalk. In the absence of any selection against crosstalk, networks with only one functioning TF evolve (Sec. 4.7 and Fig. 4.26).

An investigation into the evolutionary pathways (Sec. 4.6) taken up from the post-duplicated state to specialized states reveals the presence of two predominant pathways (Fig. 4.21). The slower pathway involves a transition through those networks in which one TF is transiently lost – meaning that specialization from those states effectively involves evolving a regulatory link from scratch. This is slow, and in terms of fitness landscape intuition, this corresponds to meandering on a huge neutral landscape. The faster pathway involves transition through networks implementing partial regulation – the two TFs partially compensate for each other, with one TF specializing in signal sensing, with the other specializing in DNA binding (Fig. 4.23). In our construction of the model, we consider the signal sensing domain to be modular – the parts of the signal sensing domain that result in a sensitivity for each signal are separate. However, the DNA-binding domain is not modular – there is no uncoupling of the components that "tell" the TF which binding site to bind to. While this assumption might not always be realistic, it reveals to us that the more modular part of the TF drives specialization – specialization often occurs faster when the first mutation is a signal sensing mutation, and often occurs slower when the first mutation is a DNA-binding mutation (Fig. 4.22 and Fig. 4.24).

Considering a more realistic scenario of multiple target genes per signal, we found that the fitness landscape becomes more rugged, and hence, often results in extremely long specialization times (Fig. 4.33) – the population is stuck at a local fitness peak that does not correspond to complete specialization (Fig. 4.32). Motivated by recent experimental findings [Pougach *et al.*, 2014; Sayou *et al.*, 2014; Aakre *et al.*, 2015], we introduced a new type of TF mutation, called "promiscuity-promoting" mutation in its DNA-binding domain, that makes the TF less specific at one of its nucleotide positions – it does not prefer any particular nucleotide on the binding site after this mutation (Fig. 4.34 and Fig. 4.35). This mutation relieves the frustration of the fitness landscape, and helps the population escape local fitness peaks more easily, drastically reducing specialization times (Fig. 4.39).

Motivated by these theoretical findings, we looked towards bioinformatics data (Chapter 5) on TF duplication to understand TF coevolution after gene duplication in real organisms. We sought to validate our model by looking for bioinformatics data that follows some predictions of the model. We considered the largest paralogous family in the human genome – the C2H2 Zn-finger TFs, as the focus of our attention. Site-independent dN/dS analysis (Fig. 5.10) of paralogous pairs does not reveal a lot of information about evolutionary patterns and selection strengths, mainly because of the spatial averaging over all protein sites. However, we do see some weak signatures of positive selection in the DNA-binding domains in medium age paralogs, and in the KRAB domains in older paralogs. Site-specific selection models (Fig. 5.11) significantly reveal that among many paralog sets, DNA-contacting amino acid residues in the DNA-binding domains have undergone positive selection more often than other amino acids, indicating strong selection pressure on paralogs to diverge in DNA-binding. Apart from this point-mutation based evolution, we also found that Zn-finger paralogs differentially lose entire DNA-binding domains and evolve on this DBD-level indicating their modular nature (Fig. 5.7). Such a

joint evolution by using both point mutations and DBD loss is probably a strong reason behind the prevalence of C2H2 Zn-finger TFs, a feature that is not present in other TF families. While further investigation into the mechanism behind DBD loss, the phylogenetic trees of DBDs in paralogs, and their interplay with point mutations is needed to gain a fuller picture of Zn-finger TF evolution, such a picture corresponds to the promiscuous pathway of the theoretical model of TF duplication coevolution. Loss of DBDs means binding shorter sequences, and hence might correspond to binding a larger repertoire of binding sites, a scenario that we will investigate in future research. KRAB-domain containing Zn-finger TFs (KZNFs) have been implicated in transposable element (TE) repression, but their coevolutionary dynamics is not well understood yet, and it is not known how paralogous KZNFs evolve and adapt to bind new TEs. We found promising preliminary results in this context – that paralogous KZNFs are retained and adapt to bind new TEs, and are sometimes again reused to bind new TEs (Fig. 5.15). This is another direction of future research we plan to undertake – the coevolution of Zn-finger TFs together with transposable elements.

Bibliography

[Aakre *et al.*, 2015] Christopher D. Aakre, Julien Herrou, Tuyen N. Phung, Barrett S. Perchuk, Sean Crosson, and Michael T. Laub, "Evolving new protein-protein interaction specificity through promiscuous intermediates," *Cell*, 163(3):594–606, October 2015.

[Achim and Arendt, 2014] Kaia Achim and Detlev Arendt, "Structural evolution of cell types by step-wise assembly of cellular modules," *Current Opinion in Genetics & Development*, 27:102–108, August 2014.

[Ackers *et al.*, 1982] G. K. Ackers, A. D. Johnson, and M. A. Shea, "Quantitative model for gene regulation by lambda phage repressor," *Proceedings of the National Academy of Sciences*, 79(4):1129–1133, February 1982.

[Adams *et al.*, 2016] Rhys M. Adams, Thierry Mora, Aleksandra M. Walczak, and Justin B. Kinney, "Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves," *eLife*, 5:e23156, December 2016.

[Aguilar-Rodríguez *et al.*, 2017] José Aguilar-Rodríguez, Joshua L. Payne, and Andreas Wagner, "A thousand empirical adaptive landscapes and their navigability," *Nature Ecology & Evolution*, 1:0045, January 2017.

[Akira *et al.*, 2006] Shizuo Akira, Satoshi Uematsu, and Osamu Takeuchi, "Pathogen Recognition and Innate Immunity," *Cell*, 124(4):783–801, February 2006.

[Allshire and Madhani, 2018] Robin C. Allshire and Hiten D. Madhani, "Ten principles of heterochromatin formation and function," *Nature Reviews Molecular Cell Biology*, 19(4):229–244, April 2018.

[Arendt, 2008] Detlev Arendt, "The evolution of cell types in animals: emerging principles from molecular studies," *Nature Reviews Genetics*, 9(11):868–882, November 2008.

[Baker *et al.*, 2011] Christopher R. Baker, Brian B. Tuch, and Alexander D. Johnson, "Extensive DNA-binding specificity divergence of a conserved transcription regulator," *Proceedings of the National Academy of Sciences*, 108(18):7493–7498, May 2011.

[Barton and Coe, 2009] N.H. Barton and J.B. Coe, "On the application of statistical physics to evolutionary biology," *Journal of Theoretical Biology*, 259(2):317, June 2009.

[Berg *et al.*, 2004] Johannes Berg, Stana Willmann, and Michael Lässig, "Adaptive evolution of transcription factor binding sites," *BMC Evolutionary Biology*, 4:42, 2004.

[Berg and von Hippel, 1987] Otto G. Berg and Peter H. von Hippel, "Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters," *Journal of Molecular Biology*, 193(4):723–743, February 1987.

[Bintu *et al.*, 2005] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips, "Transcriptional regulation by the numbers: models," *Current Opinion in Genetics & Development*, 15(2):116–124, April 2005.

[Bird, 1995] Adrian P. Bird, "Gene number, noise reduction and biological complexity," *Trends in Genetics*, 11(3):94–100, March 1995.

[Blackwood and Kadonaga, 1998] Elizabeth M. Blackwood and James T. Kadonaga, "Going the Distance: A Current View of Enhancer Action," *Science*, 281(5373):60–63, July 1998.

[Boehning *et al.*, 2018] Marc Boehning, Claire Dugast-Darzacq, Marija Rankovic, Anders S. Hansen, Taekyung Yu, Herve Marie-Nelly, David T. McSwiggen, Goran Kokic, Gina M. Dailey, Patrick Cramer, Xavier Darzacq, and Markus Zweckstetter, "RNA polymerase II clustering through carboxy-terminal domain phase separation," *Nature Structural & Molecular Biology*, 25(9):833–840, September 2018.

[Britten and Davidson, 1969] Roy J. Britten and Eric H. Davidson, "Gene Regulation for Higher Cells: A Theory," *Science*, 165(3891):349–357, July 1969.

[Burda *et al.*, 2010] Z. Burda, A. Krzywicki, O. C. Martin, and M. Zagorski, "Distribution of essential interactions in model gene regulatory networks under mutation-selection balance," *Physical Review E*, 82(1):011908, 2010.

[Burger *et al.*, 2010] Anat Burger, Aleksandra M. Walczak, and Peter G. Wolynes, "Abduction and asylum in the lives of transcription factors," *Proceedings of the National Academy of Sciences*, 107(9):4016–4021, March 2010.

[Carroll, 2005] Sean B Carroll, "Evolution at Two Levels: On Genes and Form," *PLoS Biol*, 3(7):e245, July 2005.

[Cepeda-Humerez *et al.*, 2015] Sarah A. Cepeda-Humerez, Georg Rieckh, and Gašper Tkačik, "Stochastic proofreading mechanism alleviates crosstalk in transcriptional regulation," *arXiv:1504.05716 [q-bio]*, April 2015.

[Chen *et al.*, 2018] Hongtao Chen, Michal Levo, Lev Barinov, Miki Fujioka, James B. Jaynes, and Thomas Gregor, "Dynamic interplay between enhancer-promoter topology and gene activity," *Nature Genetics*, 50(9):1296–1303, September 2018.

[Cho *et al.*, 2018] Won-Ki Cho, Jan-Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I. Cisse, "Mediator and RNA polymerase

II clusters associate in transcription-dependent condensates," *Science*, page eaar4199, June 2018.

[Chothia and Janin, 1975] Cyrus Chothia and Joël Janin, "Principles of protein-protein recognition," *Nature*, 256(5520):705–708, August 1975.

[Chuong *et al.*, 2017] Edward B. Chuong, Nels C. Elde, and Cédric Feschotte, "Regulatory activities of transposable elements: from conflicts to benefits," *Nature Reviews Genetics*, 18(2):71–86, February 2017.

[Conant *et al.*, 2014] Gavin C Conant, James A Birchler, and J Chris Pires, "Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time," *Current Opinion in Plant Biology*, 19:91–98, June 2014.

[Coulon *et al.*, 2013] Antoine Coulon, Carson C. Chow, Robert H. Singer, and Daniel R. Larson, "Eukaryotic transcriptional dynamics: from single molecules to cell populations," *Nature Reviews Genetics*, 14(8):572–584, August 2013.

[Cunningham *et al.*] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M. Ridwan Amode, Irina M. Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Carla Cummins, Claire Davidson, Kamalkumar Jayantilal Dodiya, Astrid Gall, Carlos García Girón, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, José C. Marugán, Thomas Maurel, Aoife C. McMahon, Benjamin Moore, Joannella Morales, Jonathan M. Mudge, Michael Nuhn, Denye Ogeh, Anne Parker, Andrew Parton, Mateus Patricio, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Eloise Stapleton, Marek Szuba, Kieron Taylor, Glen Threadgold, Anja Thormann, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nick Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Andrew D. Yates, Daniel R. Zerbino, and Paul Flicek, "Ensembl 2019," *Nucleic Acids Research.*

[de Vos *et al.*, 2015] Marjon G. J. de Vos, Alexandre Dawid, Vanda Sunderlikova, and Sander J. Tans, "Breaking evolutionary constraint with a tradeoff ratchet," *Proceedings of the National Academy of Sciences*, 112(48):14906–14911, December 2015.

[Desai and Fisher, 2007] Michael M. Desai and Daniel S. Fisher, "Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection," *Genetics*, 176(3):1759–1798, July 2007.

[Dubuis *et al.*, 2013] Julien O. Dubuis, Gašper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek, "Positional information, in bits," *Proceedings of the National Academy of Sciences*, 110(41):16301–16308, October 2013.

[Ecco *et al.*, 2017] Gabriela Ecco, Michaël Imbeault, and Didier Trono, "KRAB zinc finger proteins," *Development*, 144(15):2719–2729, August 2017.

[Eldar, 2011] Avigdor Eldar, "Social conflict drives the evolutionary divergence of quorum sensing," *Proceedings of the National Academy of Sciences*, 108(33):13635–13640, August 2011.

[Force *et al.*, 2005] Allan Force, William A. Cresko, F. Bryan Pickett, Steven R. Proulx, Chris Amemiya, and Michael Lynch, "The Origin of Subfunctions and Modular Gene Regulation," *Genetics*, 170(1):433–446, May 2005.

[Force *et al.*, 1999] Allan Force, Michael Lynch, F. Bryan Pickett, Angel Amores, Yilin Yan, and John Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, 151(4):1531–1545, April 1999.

[François and Hakim, 2004] Paul François and Vincent Hakim, "Design of genetic networks with specified functions by evolution in silico," *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):580–585, 2004.

[Friedlander and Brenner, 2008] T. Friedlander and N. Brenner, "Cellular properties and population asymptotics in the population balance equation," *Physical review letters*, 101(1):18104, 2008.

[Friedlander and Brenner, 2011] T. Friedlander and N. Brenner, "Adaptive response and enlargement of dynamic range," *Mathematical Biosciences and Engineering*, 8(2):515–528, 2011.

[Friedlander *et al.*, 2015] Tamar Friedlander, Avraham E. Mayo, Tsvi Tlusty, and Uri Alon, "Evolution of bow-tie architectures in biology," *PLoS Comput Biol*, 11(3):e1004055, 2015.

[Friedlander *et al.*, 2017] Tamar Friedlander, Roshan Prizak, Nicholas H. Barton, and Gašper Tkačik, "Evolution of new regulatory functions on biophysically realistic fitness landscapes," *Nature Communications*, 8(1):216, August 2017.

[Friedlander *et al.*, 2016] Tamar Friedlander, Roshan Prizak, Călin C. Guet, Nicholas H. Barton, and Gašper Tkačik, "Intrinsic limits to gene regulation by global crosstalk," *Nature Communications*, 7:12307, August 2016.

[Gama-Castro *et al.*, 2011] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, and others, "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)," *Nucleic Acids Research*, 39(suppl 1):D98–D105, 2011.

[Gerland *et al.*, 2002] Ulrich Gerland, J. David Moroz, and Terence Hwa, "Physical constraints and functional characteristics of transcription factor DNA interaction," *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, 2002.

[Gilad *et al.*, 2006] Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed, and Kevin P. White, "Expression profiling in primates reveals a rapid evolution of human transcription factors," *Nature*, 440(7081):242–245, March 2006.

[Gillespie, 1976] Daniel T Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, 22(4):403–434, December 1976.

[Gillespie, 1983] John H. Gillespie, "A simple stochastic gene substitution model," *Theoretical Population Biology*, 23(2):202–215, April 1983.

[Gillespie, 1984] John H. Gillespie, "Molecular Evolution Over the Mutational Landscape," *Evolution*, 38(5), 1984.

[Gillespie, 1994] John H. Gillespie, *The Causes of Molecular Evolution*, Oxford University Press, May 1994, Google-Books-ID: 257cnXAoREwC.

[Gillespie, 2004] John H. Gillespie, *Population Genetics: A Concise Guide*, The Johns Hopkins University Press, 2nd edition, July 2004.

[Gordon and Ruvinsky, 2012] Kacy L. Gordon and Ilya Ruvinsky, "Tempo and Mode in Evolution of Transcriptional Regulation," *PLoS Genet*, 8(1):e1002432, January 2012.

[Govern and ten Wolde, 2014] Christopher C. Govern and Pieter Rein ten Wolde, "Energy dissipation and noise correlations in biochemical sensing," *Physical Review Letters*, 113(25):258102, December 2014.

[Hahn, 2018] Matthew W. Hahn, *Molecular Population Genetics*, Sinauer Associates is an imprint of Oxford University Press, New York : Sunderland, MA, 1 edition edition, April 2018.

[Hahn *et al.*, 2003] Matthew W. Hahn, Jason E. Stajich, and Gregory A. Wray, "The Effects of Selection Against Spurious Transcription Factor Binding Sites," *Molecular Biology and Evolution*, 20(6):901–906, June 2003.

[He *et al.*, 2010] Xin He, Md. Abul Hassan Samee, Charles Blatti, and Saurabh Sinha, "Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression," *PLoS Comput Biol*, 6(9):e1000935, September 2010.

[Hilbert *et al.*, 2018] Lennart Hilbert, Yuko Sato, Hiroshi Kimura, Frank Jülicher, Alf Honigmann, Vasily Zaburdaev, and Nadine Vastenhouw, "Transcription organizes euchromatin similar to an active microemulsion," *bioRxiv*, page 234112, August 2018.

[Hillenbrand *et al.*, 2016] Patrick Hillenbrand, Ulrich Gerland, and Gašper Tkačik, "Beyond the French Flag Model: Exploiting Spatial and Gene Regulatory Interactions for Positional Information," *PLOS ONE*, 11(9):e0163628, September 2016.

[Hnisz *et al.*, 2017] Denes Hnisz, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp, "A Phase Separation Model for Transcriptional Control," *Cell*, 169(1):13–23, March 2017.

[Hobert *et al.*, 2010] Oliver Hobert, Ines Carrera, and Nikolaos Stefanakis, "The molecular and gene regulatory signature of a neuron," *Trends in neurosciences*, 33(10):435–445, 2010.

[Hobert and Westphal, 2000] Oliver Hobert and Heiner Westphal, "Functions of LIM-homeobox genes," *Trends in Genetics*, 16(2):75–83, February 2000.

[Hopfield, 1974] J. J. Hopfield, "Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity," *Proceedings of the National Academy of Sciences*, 71(10):4135–4139, October 1974.

[Imbeault *et al.*, 2017] Michaël Imbeault, Pierre-Yves Helleboid, and Didier Trono, "KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks," *Nature*, 543(7646):550–554, March 2017.

[Innan and Kondrashov, 2010] Hideki Innan and Fyodor Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," *Nature Reviews Genetics*, 11(2):97–108, February 2010.

[Itzkovitz *et al.*, 2006] Shalev Itzkovitz, Tsvi Tlusty, and Uri Alon, "Coding limits on the number of transcription factors," *BMC Genomics*, 7(1):239, September 2006.

[Jacob and Monod, 1961] François Jacob and Jacques Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, 3(3):318–356, June 1961.

[Jacobs *et al.*, 2014] Frank MJ Jacobs, David Greenberg, Ngan Nguyen, Maximilian Haeussler, Adam D Ewing, Sol Katzman, Benedict Paten, Sofie R Salama, and David Haussler, "An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons," *Nature*, 516(7530):242–245, December 2014.

[Johnson *et al.*, 2005] Jason M. Johnson, Stephen Edwards, Daniel Shoemaker, and Eric E. Schadt, "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments," *Trends in Genetics*, 21(2):93–102, February 2005.

[Johnson and Hummer, 2011] Margaret E. Johnson and Gerhard Hummer, "Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks," *Proceedings of the National Academy of Sciences*, 108(2):603–608, 2011.

[Jones and Thornton, 1996] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proceedings of the National Academy of Sciences*, 93(1):13–20, January 1996.

[Kacser and Beeby, 1984] Henrik Kacser and Richard Beeby, "Evolution of catalytic proteins," *Journal of Molecular Evolution*, 20(1):38–51, February 1984.

[Kaessmann *et al.*, 2009] Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long, "RNA-based gene duplication: mechanistic and evolutionary insights," *Nature Reviews Genetics*, 10(1):19–31, January 2009.

[Kauffman and Levin, 1987] Stuart Kauffman and Simon Levin, "Towards a general theory of adaptive walks on rugged landscapes," *Journal of Theoretical Biology*, 128(1):11–45, September 1987.

[Kim *et al.*, 2015] Somi Kim, Nam-Kyung Yu, and Bong-Kiun Kaang, "CTCF as a multifunctional protein in genome regulation and gene expression," *Experimental & Molecular Medicine*, 47(6):e166, June 2015.

[Kimura, 1962] Motoo Kimura, "On the Probability of Fixation of Mutant Genes in a Population," *Genetics*, 47(6):713–719, June 1962.

[King and Wilson, 1975] M. C. King and A. C. Wilson, "Evolution at two levels in humans and chimpanzees," *Science*, 188(4184):107–116, April 1975.

[Kinney *et al.*, 2010] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence," *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, May 2010.

[Košmrlj *et al.*, 2008] Andrej Košmrlj, Abhishek K. Jha, Eric S. Huseby, Mehran Kardar, and Arup K. Chakraborty, "How the thymus designs antigen-specific and self-tolerant T cell receptor sequences," *Proceedings of the National Academy of Sciences*, 105(43):16671–16676, October 2008.

[Kryazhimskiy *et al.*, 2009] S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin, "The dynamics of adaptation on correlated fitness landscapes," *Proceedings of the National Academy of Sciences*, 106(44):18638–18643, November 2009.

[Lalanne and François, 2013] Jean-Benoît Lalanne and Paul François, "Principles of adaptive sorting revealed by in silico evolution," *Physical Review Letters*, 110(21):218102, May 2013.

[Lambert *et al.*, 2018] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch, "The Human Transcription Factors," *Cell*, 172(4):650–665, February 2018.

[Lan and Pritchard, 2016] Xun Lan and Jonathan K. Pritchard, "Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals," *Science*, 352(6288):1009–1013, May 2016.

[Lancet *et al.*, 1993] D. Lancet, E. Sadovsky, and E. Seidemann, "Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system," *Proceedings of the National Academy of Sciences*, 90(8):3715–3719, April 1993.

[Larroux *et al.*, 2008] Claire Larroux, Graham N. Luke, Peter Koopman, Daniel S. Rokhsar, Sebastian M. Shimeld, and Bernard M. Degnan, "Genesis and expansion of metazoan transcription factor gene classes," *Molecular Biology and Evolution*, 25(5):980–996, May 2008.

[Lässig, 2007] Michael Lässig, "From biophysics to evolutionary genetics: statistical aspects of gene regulation," *BMC Bioinformatics*, 8(6):1–21, 2007.

[Levo and Segal, 2014] Michal Levo and Eran Segal, "In pursuit of design principles of regulatory sequences," *Nature Reviews Genetics*, 15(7):453–468, July 2014.

[Lieberman-Aiden *et al.*, 2009] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, 326(5950):289–293, October 2009.

[Lim *et al.*, 2018] Bomyi Lim, Tyler Heist, Michael Levine, and Takashi Fukaya, "Visualization of Transvection in Living Drosophila Embryos," *Molecular Cell*, 70(2):287–296.e6, April 2018.

[Lin and Costello, 2004] Shu Lin and Daniel J. Costello, *Error Control Coding*, Pearson, Upper Saddle River, N.J, 2 edition edition, June 2004.

[Loehlin and Carroll, 2016] David W. Loehlin and Sean B. Carroll, "Expression of tandem gene duplicates is often greater than twofold," *Proceedings of the National Academy of Sciences*, 113(21):5988–5992, May 2016.

[Lynch and Force, 2000] Michael Lynch and Allan Force, "The probability of duplicate gene preservation by subfunctionalization," *Genetics*, 154(1):459–473, January 2000.

[Lynch and Hagner, 2015] Michael Lynch and Kyle Hagner, "Evolutionary meandering of intermolecular interactions along the drift barrier," *Proceedings of the National Academy of Sciences*, 112(1):E30–E38, January 2015.

[Madan Babu *et al.*, 2006] M. Madan Babu, Sarah A. Teichmann, and L. Aravind, "Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks," *Journal of Molecular Biology*, 358(2):614–633, April 2006.

[Maerkl and Quake, 2007] Sebastian J. Maerkl and Stephen R. Quake, "A Systems approach to measuring the binding energy landscapes of transcription factors," *Science*, 315(5809):233–237, January 2007.

[Maerkl and Quake, 2009] Sebastian J. Maerkl and Stephen R. Quake, "Experimental Determination of the Evolvability of a Transcription Factor," *Proceedings of the National Academy of Sciences*, 106(44):18650–18655, November 2009.

[Magadum *et al.*, 2013] Santoshkumar Magadum, Urbi Banerjee, Priyadharshini Murugan, Doddabhimappa Gangapur, and Rajasekar Ravikesavan, "Gene duplication as a major force in evolution," *Journal of Genetics*, 92(1):155–161, March 2013.

[Mangan and Alon, 2003] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, October 2003.

[Maslov *et al.*, 2009] Sergei Maslov, Sandeep Krishna, Tin Yau Pang, and Kim Sneppen, "Toolbox model of evolution of prokaryotic metabolic networks and their regulation," *Proceedings of the National Academy of Sciences*, 106(24):9743–9748, June 2009.

[Maston *et al.*, 2006] Glenn A. Maston, Sara K. Evans, and Michael R. Green, "Transcriptional Regulatory Elements in the Human Genome," *Annual Review of Genomics and Human Genetics*, 7(1):29–59, September 2006.

[Mathelier *et al.*, 2013] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, François Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman, "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles," *Nucleic Acids Research*, page gkt997, November 2013.

[McKeithan, 1995] T. W. McKeithan, "Kinetic proofreading in T-cell receptor signal transduction," *Proceedings of the National Academy of Sciences*, 92(11):5042–5046, May 1995.

[McKeown *et al.*, 2014] Alesia N. McKeown, Jamie T. Bridgham, Dave W. Anderson, Michael N. Murphy, Eric A. Ortlund, and Joseph W. Thornton, "Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module," *Cell*, 159(1):58–68, September 2014.

[Mitchell and Tjian, 1989] P. J. Mitchell and R. Tjian, "Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins," *Science*, 245(4916):371–378, July 1989.

[Mora, 2015] Thierry Mora, "Physical limit to concentration sensing amid spurious ligands," *arXiv:1504.07203 [physics, q-bio]*, April 2015.

[Murugan *et al.*, 2015] Arvind Murugan, James Zou, and Michael P. Brenner, "Undesired usage and the robust self-assembly of heterogeneous structures," *Nature communications*, 6, 2015.

[Myers, 2008] Christopher R. Myers, "Satisfiability, sequence niches and molecular codes in cellular signalling," *IET systems biology*, 2(5):304–312, 2008.

[Nadimpalli *et al.*, 2015] Shilpa Nadimpalli, Anton V. Persikov, and Mona Singh, "Pervasive variation of transcription factor orthologs contributes to regulatory network evolution," *PLOS Genet*, 11(3):e1005011, March 2015.

[Najafabadi *et al.*, 2015] Hamed S. Najafabadi, Sanie Mnaimneh, Frank W. Schmitges, Michael Garton, Kathy N. Lam, Ally Yang, Mihai Albu, Matthew T. Weirauch, Ernest Radovani, Philip M. Kim, Jack Greenblatt, Brendan J. Frey, and Timothy R. Hughes, "C2H2 zinc finger proteins greatly expand the human regulatory lexicon," *Nature Biotechnology*, 33(5):555–562, May 2015.

[Necsulea and Kaessmann, 2014] Anamaria Necsulea and Henrik Kaessmann, "Evolu-

tionary dynamics of coding and non-coding transcriptomes," *Nature Reviews Genetics*, 15(11):734–748, November 2014.

[Nguyen and Saier, 1995] Chuck Chuong Nguyen and Milton H. Saier, "Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors," *FEBS Letters*, 377(2):98–102, December 1995.

[Ohno, 1970] Susumu Ohno, *Evolution by gene duplication.*, Springer-Verlag, 1970.

[Olson, 2006] Eric N. Olson, "Gene Regulatory Networks in the Evolution and Development of the Heart," *Science*, 313(5795):1922–1927, September 2006.

[Ouldridge and ten Wolde, 2014] Thomas Ouldridge and Pieter Rein ten Wolde, "The robustness of proofreading to crowding-induced pseudo-processivity in the MAPK pathway," *Biophysical Journal*, 107(10):2425–2435, November 2014.

[Payne and Wagner, 2014] Joshua L. Payne and Andreas Wagner, "The robustness and evolvability of transcription factor binding sites," *Science*, 343(6173):875–877, February 2014.

[Pennacchio *et al.*, 2013] Len A. Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano, "Enhancers: five essential questions," *Nature Reviews Genetics*, 14(4):288–295, April 2013.

[Pérez *et al.*, 2014] J. Christian Pérez, Polly M. Fordyce, Matthew B. Lohse, Victor Hanson-Smith, Joseph L. DeRisi, and Alexander D. Johnson, "How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes," *Genes & Development*, 28(12):1272–1277, June 2014.

[Phillips, 2015] Rob Phillips, "Napoleon is in equilibrium," *Annual Review of Condensed Matter Physics*, 6(1):85–111, 2015.

[Podgornaia and Laub, 2015] Anna I. Podgornaia and Michael T. Laub, "Pervasive degeneracy and epistasis in a protein-protein interface," *Science*, 347(6222):673–677, February 2015.

[Poelwijk *et al.*, 2006] Frank J. Poelwijk, Daniel J. Kiviet, and Sander J. Tans, "Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data," *PLoS computational biology*, 2(5):e58, 2006.

[Pougach *et al.*, 2014] Ksenia Pougach, Arnout Voet, Fyodor A. Kondrashov, Karin Voordeckers, Joaquin F. Christiaens, Bianka Baying, Vladimir Benes, Ryo Sakai, Jan Aerts, Bo Zhu, Patrick Van Dijck, and Kevin J. Verstrepen, "Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network," *Nature Communications*, 5:4868, September 2014.

[Proulx, 2012] Stephen R. Proulx, "Multiple Routes to Subfunctionalization and Gene Duplicate Specialization," *Genetics*, 190(2):737–751, February 2012.

[Rieckh and Tkačik, 2014] Georg Rieckh and Gašper Tkačik, "Noise and information transmission in promoters with multiple internal states," *Biophysical Journal*, 106(5):1194–1204, March 2014.

[Rockel *et al.*, 2012] Sylvie Rockel, Marcel Geertz, Korneel Hens, Bart Deplancke, and Sebastian J. Maerkl, "iSLIM: a comprehensive approach to mapping and characterizing gene regulatory networks," *Nucleic acids research*, page gks1323, 2012.

[Rowland and Deeds, 2014] Michael A. Rowland and Eric J. Deeds, "Crosstalk and the evolution of specificity in two-component signaling," *Proceedings of the National Academy of Sciences*, 111(15):5550–5555, April 2014.

[Sauka-Spengler and Bronner-Fraser, 2008] Tatjana Sauka-Spengler and Marianne Bronner-Fraser, "A gene regulatory network orchestrates neural crest formation," *Nature Reviews Molecular Cell Biology*, 9(7):557–568, July 2008.

[Sayou *et al.*, 2014] Camille Sayou, Marie Monniaux, Max H. Nanao, Edwige Moyroud, Samuel F. Brockington, Emmanuel Thévenon, Hicham Chahtane, Norman Warthmann, Michael Melkonian, Yong Zhang, Gane Ka-Shu Wong, Detlef Weigel, FranÃ§ois Parcy, and Renaud Dumas, "A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity," *Science*, 343(6171):645–648, February 2014.

[Schneider *et al.*, 1986] Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andrzej Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *Journal of Molecular Biology*, 188(3):415–431, April 1986.

[Schultz *et al.*, 2000] Jörg Schultz, Richard R. Copley, Tobias Doerks, Chris P. Ponting, and Peer Bork, "SMART: a web-based tool for the study of genetically mobile domains," *Nucleic Acids Research*, 28(1):231–234, January 2000.

[Schuster, 2000] Peter Schuster, "Taming combinatorial explosion," *Proceedings of the National Academy of Sciences*, 97(14):7678–7680, July 2000.

[Schuster *et al.*, 1994] Peter Schuster, Walter Fontana, Peter F. Stadler, and Ivo L. Hofacker, "From sequences to shapes and back: a case study in RNA secondary structures," *Proceedings of the Royal Society of London B: Biological Sciences*, 255(1344):279–284, March 1994.

[Sear, 2004a] Richard P. Sear, "Highly specific protein-protein interactions, evolution and negative design," *Physical Biology*, 1(3):166, 2004.

[Sear, 2004b] Richard P. Sear, "Specific protein-protein binding in many-component mixtures of proteins," *Physical Biology*, 1(2):53, 2004.

[Sella and Hirsh, 2005] Guy Sella and Aaron E. Hirsh, "The application of statistical physics to evolutionary biology," *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9541–9546, July 2005.

[Sengupta *et al.*, 2002] Anirvan M. Sengupta, Marko Djordjevic, and Boris I. Shraiman,

"Specificity and robustness in transcription control networks," *Proceedings of the National Academy of Sciences*, 99(4):2072–2077, February 2002.

[Shea and Ackers, 1985] Madeline A. Shea and Gary K. Ackers, "The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation," *Journal of Molecular Biology*, 181(2):211–230, January 1985.

[Sheinman and Kafri, 2012] M. Sheinman and Y. Kafri, "How does the DNA sequence affect the Hill curve of transcriptional response?," *Physical Biology*, 9(5):056006, October 2012.

[Sherman and Cohen, 2012] Marc S. Sherman and Barak A. Cohen, "Thermodynamic state ensemble models of cis-regulation," *PLoS Comput Biol*, 8(3):e1002407, March 2012.

[Shultzaberger *et al.*, 2012] Ryan K. Shultzaberger, Sebastian J. Maerkl, Jack F. Kirsch, and Michael B. Eisen, "Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA Binding Domain," *PLoS Genetics*, 8(3):e1002614, March 2012.

[Simionato *et al.*, 2007] Elena Simionato, Valérie Ledent, Gemma Richards, Morgane Thomas-Chollier, Pierre Kerner, David Coornaert, Bernard M. Degnan, and Michel Vervoort, "Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics," *BMC Evolutionary Biology*, 7:33, 2007.

[Skerker *et al.*, 2008] Jeffrey M. Skerker, Barrett S. Perchuk, Albert Siryaporn, Emma A. Lubin, Orr Ashenberg, Mark Goulian, and Michael T. Laub, "Rewiring the specificity of two-component signal transduction systems," *Cell*, 133(6):1043–1054, June 2008.

[Skoge *et al.*, 2013] Monica Skoge, Sahin Naqvi, Yigal Meir, and Ned S. Wingreen, "Chemical Sensing by Nonequilibrium Cooperative Receptors," *Physical Review Letters*, 110(24):248102, June 2013.

[Smith *et al.*, 2016] A. F. A. Smith, R. Hubley, and P. Green, *RepeatMasker Open-4.0.(2013-2015)*, 2016.

[Spitz and Furlong, 2012] François Spitz and Eileen E. M. Furlong, "Transcription factors: from enhancer binding to developmental control," *Nature Reviews Genetics*, 13(9):613–626, September 2012.

[Spivak and Stormo, 2012] Aaron T. Spivak and Gary D. Stormo, "ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species," *Nucleic Acids Research*, 40(D1):D162–D168, January 2012.

[Stern and Orgogozo, 2009] David L. Stern and Virginie Orgogozo, "Is genetic evolution predictable?," *Science*, 323(5915):746–751, 2009.

[Stewart *et al.*, 2012] Alexander J. Stewart, Sridhar Hannenhalli, and Joshua B. Plotkin, "Why Transcription Factor Binding Sites Are Ten Nucleotides Long," *Genetics*, 192(3):973–985, November 2012.

[Strom *et al.*, 2017] Amy R. Strom, Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen, "Phase separation drives heterochromatin domain formation," *Nature*, 547(7662):241–245, July 2017.

[Suyama *et al.*, 2006] Mikita Suyama, David Torrents, and Peer Bork, "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments," *Nucleic Acids Research*, 34(suppl_2):W609–W612, July 2006.

[Swain and Siggia, 2002] Peter S. Swain and Eric D. Siggia, "The role of proofreading in signal transduction specificity," *Biophysical Journal*, 82(6):2928–2933, June 2002.

[Thompson *et al.*, 1994] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, 22(22):4673–4680, November 1994.

[Tkačik and Bialek, 2016] Gašper Tkačik and William Bialek, "Information Processing in Living Systems," *Annual Review of Condensed Matter Physics*, 7(1):89–117, 2016.

[Tkačik *et al.*, 2010] Gašper Tkačik, Jason S. Prentice, Vijay Balasubramanian, and Elad Schneidman, "Optimal population coding by noisy spiking neurons," *Proceedings of the National Academy of Sciences*, 107(32):14419–14424, August 2010.

[Tkačik and Walczak, 2011] Gašper Tkačik and Aleksandra M. Walczak, "Information transmission in genetic regulatory networks: a review," *Journal of Physics: Condensed Matter*, 23(15):153102, April 2011.

[Todeschini *et al.*, 2014] Anne-Laure Todeschini, Adrien Georges, and Reiner A. Veitia, "Transcription factors: specific DNA binding and specific gene regulation," *Trends in Genetics*, 30(6):211–219, 2014.

[Tuğrul *et al.*, 2015] Murat Tuğrul, Tiago Paixão, Nicholas H. Barton, and Gašper Tkačik, "Dynamics of Transcription Factor Binding Site Evolution," *PLoS Genet*, 11(11):e1005639, November 2015.

[van Nimwegen, 2004] Erik van Nimwegen, "Scaling laws in the functional content of genomes: Fundamental constants of evolution?," *arXiv:q-bio/0405022*, May 2004.

[Vaquerizas *et al.*, 2009] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nature Reviews Genetics*, 10(4):252–263, April 2009.

[Villar *et al.*, 2014] Diego Villar, Paul Flicek, and Duncan T. Odom, "Evolution of transcription factor binding in metazoans - mechanisms and functional implications," *Nature Reviews Genetics*, 15(4):221–233, April 2014.

[Vinckenbosch *et al.*, 2006] Nicolas Vinckenbosch, Isabelle Dupanloup, and Henrik Kaessmann, "Evolutionary fate of retroposed gene copies in the human genome," *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3220–3225, February 2006.

[von Dassow *et al.*, 2000] George von Dassow, Eli Meir, Edwin M. Munro, and Garrett M. Odell, "The segment polarity network is a robust developmental module," *Nature*, 406(6792):188–192, July 2000.

[Von Hippel and Berg, 1986] P. H. Von Hippel and O. G. Berg, "On the specificity of DNA-protein interactions," *Proceedings of the National Academy of Sciences*, 83(6):1608, 1986.

[Von Hippel *et al.*, 1974] Peter H. Von Hippel, Arnold Revzin, Carol A. Gross, and Amy C. Wang, "Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects," *Proceedings of the National Academy of Sciences*, 71(12):4808–4812, 1974.

[Wang *et al.*, 2016] Jiyong Wang, Sharon T. Jia, and Songtao Jia, "New Insights into the Regulation of Heterochromatin," *Trends in Genetics*, 32(5):284–294, May 2016.

[Weinert *et al.*, 2014] Franz M. Weinert, Robert C. Brewster, Mattias Rydenfelt, Rob Phillips, and Willem K. Kegel, "Scaling of gene expression with transcription-factor fugacity," *Physical Review Letters*, 113(25):258101, December 2014.

[Wittkopp and Kalay, 2012] Patricia J. Wittkopp and Gizem Kalay, "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence," *Nature Reviews Genetics*, 13(1):59–69, January 2012.

[Wray, 2007] Gregory A. Wray, "The evolutionary significance of cis-regulatory mutations," *Nature Reviews Genetics*, 8(3):206–216, March 2007.

[Wray *et al.*, 2003] Gregory A. Wray, Matthew W. Hahn, Ehab Abouheif, James P. Balhoff, Margaret Pizer, Matthew V. Rockman, and Laura A. Romano, "The Evolution of Transcriptional Regulation in Eukaryotes," *Molecular Biology and Evolution*, 20(9):1377–1419, September 2003.

[Wright, 1931] Sewall Wright, "Evolution in Mendelian populations," *Genetics*, 16(2):97–159, 1931.

[Wunderlich and Mirny, 2009] Zeba Wunderlich and Leonid A. Mirny, "Different gene regulation strategies revealed by analysis of binding motifs," *Trends in Genetics*, 25(10):434–440, October 2009.

[Yamane and Hopfield, 1977] T. Yamane and J. J. Hopfield, "Experimental evidence for kinetic proofreading in the aminoacylation of tRNA by synthetase," *Proceedings of the National Academy of Sciences*, 74(6):2246–2250, June 1977.

[Yang *et al.*, 2017] Peng Yang, Yixuan Wang, and Todd S. Macfarlan, "The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution," *Trends in Genetics*, 33(11):871–881, November 2017.

[Yang, 2007] Ziheng Yang, "PAML 4: Phylogenetic Analysis by Maximum Likelihood," *Molecular Biology and Evolution*, 24(8):1586–1591, August 2007.

[Yang *et al.*, 2005] Ziheng Yang, Wendy S. W. Wong, and Rasmus Nielsen, "Bayes empirical bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, 22(4):1107–1118, April 2005.

[Yona *et al.*, 2015] Avihu H. Yona, Idan Frumkin, and Yitzhak Pilpel, "A relay race on the evolutionary adaptation spectrum," *Cell*, 163(3):549–559, October 2015.

[Yona *et al.*, 2012] Avihu H. Yona, Yair S. Manor, Rebecca H. Herbst, Gal H. Romano, Amir Mitchell, Martin Kupiec, Yitzhak Pilpel, and Orna Dahan, "Chromosomal duplication is a transient evolutionary solution to stress," *Proceedings of the National Academy of Sciences*, 109(51):21010–21015, December 2012.

[Zhang *et al.*, 2008] Jingshan Zhang, Sergei Maslov, and Eugene I. Shakhnovich, "Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size," *Molecular systems biology*, 4(1), 2008.

[Zhang *et al.*, 2006] Zhang Zhang, Jun Li, Xiao-Qian Zhao, Jun Wang, Gane Ka-Shu Wong, and Jun Yu, "KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging," *Genomics, Proteomics & Bioinformatics*, 4(4):259–263, January 2006.

[Zhou *et al.*, 2007] Qing Zhou, Hiram Chipperfield, Douglas A. Melton, and Wing Hung Wong, "A gene regulatory network in mouse embryonic stem cells," *Proceedings of the National Academy of Sciences*, 104(42):16438–16443, October 2007.