Evolution of regulatory complexes: a many-body system

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Armita Nour Mohammad

aus Teheran, Iran

Köln, 2012

Berichterstatter: Prof. Dr. Michael Lässig

Prof. Dr. Joachim Krug

Prof. Dr. Martin Kreitman

Tag der letzten mündlichen Prüfung: 12.04.2012

Abstract

The recent advent of large-scale genomic sequence data and improvement of sequencing technologies has enabled population genetics to advance from a mostly abstract theoretical basis to a quantitative molecular description. However, functional units in DNA are typically combinations of interacting nucleotide segments, and evolutionary forces acting on these segments can result in very complicated population dynamics. The goal is to formulate these interactions in such a way that the macroscopic features are independent of the microscopic details, as in statistical mechanics.

In this thesis, I discuss the evolutionary dynamics of regulatory sequences, which control the production of protein in cells. One of the primary forms of regulation occurs through interactions of proteins called *Transcription factors*, with *binding sites* in the DNA sequence, and the strength of these interactions influence the individual's fitness in the population. What makes this an ideal model system for quantitative analysis of genomic evolution, is the possibility of inferring this relationship.

Compared to prokaryotes and yeast, gene regulation is much more complex in higher eukaryotes. Regulatory information is organized in modules with multiple binding sites that are linked to a common function. In Chapter. 2, we show that binding site complexes are commonly formed by local sequence duplications, as opposed to forming from scratch by single point mutations. We also show that the underlying regulatory grammar is in tune with this mechanism such that the duplication events confer an adaptive advantage.

Regulatory complexes resemble a many-particle system whose function emerges from the collective dynamics of its elements. In Chapter. 3, we develop a thermodynamic framework to characterize the effective affinity of site complexes to multiple transcription factors with cooperative binding. These affinities are the phenotype, or trait of binding complexes on which selection acts, and we characterize their evolution. From the yeast genome polymorphism data, we infer a fitness landscape as a function of binding affinity by using the novel method developed in Chapter. 4. This method of quantitative trait analysis can deal with long-range correlations between

sites which arise in asexual populations. Our fitness landscape quantitatively predicts the amount of conservation of the phenotype, as well as the amount of compensatory changes between sites.

Our results open a new avenue to understand the regulatory "grammar" of eukaryotic genomes based on quantitative evolution models. They prove that a combination of theoretical models, high-throughput experimental measurements, and analysis of genomic variation is necessary for a proper quantitative understanding of biological systems.

Acknowledgements

First of all, I would like to express my gratitude to my advisor Michael Lässig. Thanks Michael for your dedicated guidance, for the continuous support and advice with all the (un)expectables and for giving me the courage to "just give it a try"! I have learned a lot from you and would always be thankful for the enjoyable time that I had in the past couple of years. Thanks Johannes for being very considerate, helping me upon my arrival in Cologne and pointing out always some good pieces of science to me. Our occasional climbings have engraved long-lasting memories. Thanks Ville for the scientific and life advices and the occasional humor. I would like to acknowledge all the members of Lässig and Berg group, Lilia, Stéphane, Daniel, Pras, Filippos, Christa, Chau, Nico, Joachim, Jörn and specially my officemates, Marta, Stephan, Laleh and Donate. Special thanks to Daniel, Marta and Stephan for helping me with the German version of my abstract.

In these years, I had the chance to meet many great scientists. I would specially like to acknowledge Luca Peliti, Leonid Mirny, David Hughes, Kevin Foster, Aleksandra Walczak and Shamil Sunyaev for making up all those enjoyable and inspiring moments.

I am thankful to all my friends with whom I traveled, hiked, climbed, watched movies, played games; you made my post-graduate years really fun! Special thanks to Jakub, Valeska and Taha for also giving me their useful comments on this manuscript.

At last, I would like to thank my family who has always supported me and has always been there for. My parents, who from the very beginning, thought me to think outside of the box and to value my dreams. Arman, Shervin and Arvin who offered me the warm feeling of being with family even when I was far from home. And Armin, for his kindness and cares.

Financial supports. My work has been supported by *SFB 680*, "Molecular Basis of Evolutionary Innovations" and by the Bonn-Cologne Graduate school of physics and Astronomy.

Collaborations. The work presented in Chapter. 4 has been done in collaboration with Stephan Schiffels.

Publications. Parts of the material covered in Chapter. 2 is published in PLoS Computational Biology (Nourmohammad and Lässig, 2011).

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

1.1 Gene regulation

Proteins, as functional units in the cell, are encoded by "genes" in DNA sequence. The decoding process from DNA to proteins involves the following steps: a molecule called *RNA-polymerase* (RNAP) transcribes the encoded gene in the DNA sequence to another polymer, *RNA*, which is later translated into amino-acid chain molecules that are the building blocks for the protein; see Fig. 1.1(a). Although, all cells in an organism carry identical DNA molecules which encode similar proteins, they perform distinct functions ranging from blood cells to brain tissue cells. The answer to this dilemma lies in the intermediate yet influential role of the control machinery (*regulatory system*) during protein production. The most pervasive form of gene regulation occurs during the first step of RNA transcription. Special types of proteins, *transcription factors* (TF), recognize and bind to specific site sequences (*binding sites*) in the regulatory region of a gene (*promoter sequence*) and affect the transcription rate and hence protein production; Fig. 1.1(b). The significance of regulatory variation as a driving force for phenotypic evolution has been suggested some time ago (Monod and Jacob, 1961). Most of the phenotypic variations (i.e., difference of functional characters) between species are not due to their protein code but rather the control machinery which determines the combination and amount of available proteins in the cell at various points in time (King and Wilson, 1975; Monod and Jacob, 1961; Ptashne and Gann, 2002).

Despite the biological significance of the regulatory system, a quantitative understanding of gene regulation has become possible only after the advent of large-

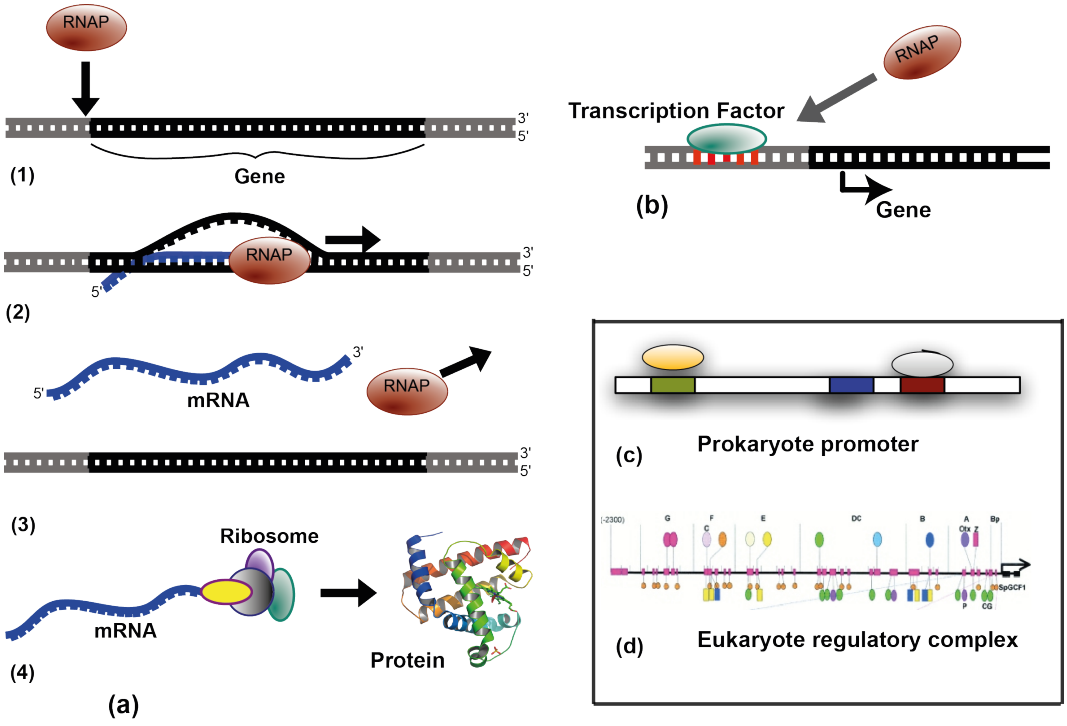

Figure 1.1: Gene expression and regulatory processes. (a) Information transfer from DNA sequence to protein molecule: (1) RNA-polymerase (RNAP) binds to the DNA sequence and (2) transcribes a single stranded polymer, RNA molecule. (3) Messenger RNA (mRNA) is the transcribed copy of the DNA sequence which encodes the information for protein production. (4) A ribosome molecule then translates the mRNA sequence to amino acids which are the building blocks of the protein. The structure will be folded and form a 3-dimensional protein molecule. (b) Transcription regulation is one of the most pervasive forms of gene regulation. Throughout this process a type of protein molecule called *transcription factors* binds to the *binding sites* in the regulatory region of the gene and form a complex that interacts with the RNA-polymerase. In this way, the level of protein production in the cell is regulated by proteins called *transcription factors*, which interact with RNA-polymerase. (c) Promoter architecture depends on the organism's complexity. Prokaryotes and unicellular eukaryotes have short regulatory regions, which encode a few binding sites. (d) Multi-cellular eukaryotes exhibit more complex regulatory architectures organized in modules with multiple binding sites that interact with different transcription factors. The example here is a regulatory region in the sea urchin (Davidson, 2006).

scale genomic sequences and regulatory interaction data. Important building blocks are genome-wide maps of protein-DNA binding, statistical inference methods (Berg and von Hippel, 1987; Stormo and Fields, 1998), high-throughput measurements of sequence-specific binding affinities of transcription factors (Badis and et. al., 2009; Fields et al., 1997; Maerkl and Quake, 2007; Mukherjee et al., 2004), and cross-species

comparisons of regulatory sequences and regulatory functions (Stark and et. al., 2007).

1.2 Complexity of regulatory structures

The evolutionary constraint of regulatory sequence and function depends on the level of complexity in promoter architecture. Prokaryotes and unicellular eukaryotes have short intergenic regions, and regulatory functions are often encoded by only a few binding sites; Fig. 1.1(c). The more complex cis-regulatory information in higher eukaryotes is organized into *regulatory modules*, which are typically a few hundred base pairs long and are spatially separated by larger segments of intergenic DNA (Bergman and et. al., 2002; Ondek et al., 1988); Fig. 1.1(d). Within modules, regulatory functions often depend on clusters of neighboring binding sites for multiple transcription factors, which are coupled by cooperative interactions (Davidson, 2006; Harbison and et. al., 2004; Lynch, 2006; Ptashne and Gann, 2002; Sinha et al., 2004). The relative order and spacing of sites within clusters follows a regulatory "grammar", which distinguishes functionally neutral site changes from rearrangements affecting promoter function (Arnosti and Kulkarni, 2005; Kulkarni and Arnosti, 2005; Lusk and Eisen, 2010; Markstein et al., 2002; Small et al., 1993; Stanojevic et al., 1991).

The combinatorial complexity of this grammar ensures the specificity of regulation in the larger genomes of multicellular eukaryotes (Buchler et al., 2003; Levine and Tjian, 2003). Unlike prokaryotes, single binding sites in eukaryotes are not specific enough to be recognized by transcription factors to alter gene expression. The required specificity however, can emerge from module structures with an agglomeration of binding sites; Fig. 1.1(d). We can address this feature in a simple quantitative fashion. Information theory dictates that finding a unique object among L alternatives requires $I_{\mathrm{min}} = \log_2 L$ bits of information. Similarly, a minimum of $I_{\mathrm{min}} = \log_2 L$ bits of information is required to specify a unique location in a genome containing L possible sites for a transcription factor to bind (i.e. L bp sequence). In Table. 1.1, we compare the required regulatory information to the actual information content of transcription factors in three classes of species: prokaryotes (represented by *E. coli*), unicellular eukaryotes (represented by *S. cerevisiae*) and multicellular eukaryotes (*D. melanogaster*).

	L (bp)	$I_{\min} = \log_2 L$ (bits)	$\langle I \rangle$(bits)
E. coli	10^6	20	$20 - 27$
S. cerevisiae	10^7	23	$2 - 17$
D. melanogaster	10^8	27	$6 - 8$

Table 1.1: Different regulatory strategies between prokaryotes and eukaryotes. Unlike prokaryotes, individual transcription factors in multi-cellular eukaryotes, such as *Drosophila*, do not encoded sufficient amount of information to identify single binding sites in the DNA sequence. The third column shows the amount of information required to identify a single binding site, $I_{\min}$, and the fourth column shows the average information content of transcription factors $\langle I \rangle$ in the organism. To overcome this inconsistency, the regulatory information in eukaryotes is organized in modules with multiple binding site which provide the required specificity in those large genomes.

The genome length and thus the minimum required information $I_{\min}$ from the transcription factors increase with the organisms complexity. The actual amount of information encoded by the transcription factors however, does not follow this pattern. We compute the information content of a typical transcription factor I, from the redundancy of the functional sequence patterns that bind to it; see the Discussion on information content of the binding motifs in Chapter. 2.The specificity of the transcription factors are sufficient in prokaryotes but far below the minimum limit in multicellular eukaryotes, such as flies. In this case, the presence of multiple sites in proximity to each other, i.e., regulatory modules, play the leading role in specifying a regulatory region; see e.g., (Wunderlich and Mirny, 2009). Characterizing regulatory modules, their formation, function and evolutionary conservation is the central focus of this thesis.

1.3 Regulatory interaction: a biophysical approach

As mentioned above, regulation is grounded in the biophysics of TF-DNA interactions which has been approached both by statistical inference methods and high-throughput experimental measurements of sequence-specific binding affinities. The strength of these interactions influence the functional output of the cell (i.e., gene expression) and hence, it is a molecular phenotype on which natural selection acts. The possibility of inferring (albeit simplistic) genotype-phenotype maps, makes the regulatory sequence an ideal model system for quantitative analysis of genomic evolution. Thus, characterizing the

regulatory code is a significant step in an emerging interaction-based picture of the genome.

Binding of a transcription factor to a DNA sequence is a probabilistic process with an affinity that depends on the nucleotide content of the binding sequence, with a length of about 10-15 base pairs. The binding energy of this protein-DNA interaction determines the probability of the transcription factor to be bound at the binding site. In a reversible binding interaction,

$$\mathrm{TF} + \mathrm{DNA} \underset{k_{diss}}{\overset{k_{bind}}{\rightleftharpoons}} \mathrm{TF\text{-}DNA}$$

the rate constants, i.e., binding constant k_{bind} and dissociation constant k_{diss}, are related to the binding energy of the interaction E through the Boltzmann weight,

$$\frac{k_{bind}}{k_{diss}} \propto \exp(-\beta\ E) \tag{1.1}$$

where β is proportional to the inverse temperature, $\beta = 1/k_B T$. We can compute the occupancy probability $\rho(E)$ for a sequence with binding energy E in a solution that contains the corresponding transcription factor,

$$\rho(E) = \frac{k_{bind}\ n_{tf}}{k_{bind}\ n_{tf} + k_{diss}} = \frac{1}{1 + \exp[\beta(E - \nu)]} \tag{1.2}$$

ν is the so-called chemical potential in statistical physics and is related to the transcription factor density n_{th}: $\nu = k_B T \log \kappa n_{tf}$ (Buchler et al., 2003; Djordjevic et al., 2003; Lässig, 2007). κ is the proportionality factor in eq. (1.1). The binding probability in eq. (1.2) is a nonlinear function (Fermi function) of the sequence dependent binding energy between the protein and the DNA sequence. The chemical potential in eq. (1.2) acts as a threshold for the sequence specific binding energy E, below which the transcription factor is more likely to be bound to the sequence. We will see that the nonlinear dependency of the binding probability on the interaction energy will also be reflected in the fitness value of the binding site.

Binding energy is of course a biophysical quantity that can be measured experimentally. High-throughput techniques, such as microfluidic experiments (Maerkl and Quake, 2007), have been developed to measure the sequence specific binding affinities

of the transcription factors. Statistical inference methods also play an important role in this direction (Berg and von Hippel, 1987; Stormo and Fields, 1998). Identifying functional sites in a genome is not as strenuous as measuring the binding affinity of a protein to all nucleotide combinations. Experimental methods such as ChIP-chip (Lee, 2002; Ren, 2000) or protein binding microarrays (PBMs) (Mukherjee et al., 2004) can identify binding sites for different transcription factors on a genome-wide scale. We can then use information theoretical based methods to infer the energy contribution of single nucleotides from the ensemble of the functional binding sequences (Berg and von Hippel, 1987; Lässig, 2007; Stormo and Fields, 1998). The key idea is to weigh the over-representation of the sequence compositions associated with the class of functional sites, in relation to their binding affinity to the transcription factor.

The probability distribution of functional binding sites $Q(\mathbf{a})$ is significantly different from the genomic background distribution $P_0(\mathbf{a})$ for sites of length ℓ, $\mathbf{a} = (a_1, \ldots, a_\ell)$. If functional sites are drawn from a Boltzmann distribution, where their occurrence is in proportion to their affinity, then the best energy model to distinguish this site ensemble from the background ensemble is, $E(\mathbf{a}) = \log Q(\mathbf{a})/P_0(\mathbf{a})$. This is an immediate consequence of the Maximum Entropy Principle (Jaynes, 1957). Two simplifying assumptions make this inference a straightforward process: (i) the energy contributions are additive across the positions of the site and (ii) the correlations between nucleotides are negligible. This allows us to express the site distributions as the product of single-nucleotide frequencies,

$$Q(\mathbf{a}) = \prod_{i=1}^{\ell} q_i(a_i) \tag{1.3}$$

and $P_0(\mathbf{a}) = \prod_{i=1}^{\ell} p_0(a_i)$. The $4 \times \ell$ matrix of single-nucleotide frequencies (2.3) is called the position weight matrix (PWM) of the transcription factor. Therefore,

$$E(\mathbf{a}) = \sum \varepsilon_i(a_i) \quad \text{with,} \quad \varepsilon_i(a_i) = \log \frac{q_i(a_i)}{p_0(a_i)} \tag{1.4}$$

We will use these simplifying approximations for our analysis in the following chapters. However, more refined statistical inference methods have been developed that include higher orders of nucleotide correlations and are suitable for analysis of larger sampling sets (Gershenzon and Stormo, 2005; Siddharthan, 2010).

1.4 Population genetics of binding site evolution

Quantifying phenotype, in this case binding phenotype, is a significant and necessary step in characterizing the evolutionary dynamics of a population. The primary forces of evolutionary dynamics are (i) mutations that generate genomic variation during reproduction, (ii) genetic drift related to the stochastic sampling of a discrete population and (iii) natural selection which determines the reproduction success of a subpopulation with a certain phenotype. Mutations and genetic drift are the evolutionary forces which directly interact with genotypes, whereas natural selection is the response to the phenotypic manifestation of the genetic code in the population. A description of the evolutionary dynamics of a population should involve a map between the genetic code and its phenotypic outcome. Other molecular mechanisms, such as recombination in sexual organisms and horizontal gene transfer in bacteria, also play important parts during evolution. However, we do not include them in our following analysis.

We assume a binding locus as a sequence of nucleotides $\mathbf{a} = (a_1, \ldots, a_\ell)$ which can potentially bind to a transcription factor. In this case, the genotype is the binding sequence "$\mathbf{a}$" and the phenotype is its binding affinity to the transcription factor "$E(\mathbf{a})$" which regulates gene expression in the cell. First, we describe the neutral evolutionary dynamics in a finite population of size N. The neutral forces, mutations and genetic drift, modify the population composition of the genotype $\mathbf{a}$ during evolution. Most biological systems, except for viruses and types of mutator bacteria, evolve in the low mutation regime, $\mu N \ll 1$ where μ is the mutation rate per nucleotide per generation. The short length of binding sites in eukaryotes, about 10-15 bp, assures that $\mu N \ell < 1$, and therefore the binding sites are mostly monomorphic in the population. In this *weak mutation regime*, it is reasonable to picture the subsequent substitutions, which are the fixed mutations in the population, as independent jump events which are well separated in time. This is not the case for larger genomic loci, such as a whole promoter sequence that provides a larger mutational target during evolution; see Chapter. 3 for discussion. Since the substitutions are separated in time, we can assume that populations at most contain two genotypes, $\mathbf{a}$ and $\mathbf{b}$. At the level of individuals, mutations are the stochastic processes that change $\mathbf{a} \to \mathbf{b}$ with a rate, $\mu_{a\to b}$ or vice versa. We denote the size of the subpopulation that carry genotype $\mathbf{a}$ with $N_\mathbf{a}$ and hence the rest with genotype

b are of the size $N_{\mathbf{b}} = N - N_{\mathbf{a}}$. The change in number of **a**-carriers in each generation is,

$$\frac{\mathrm{d}}{\mathrm{d}t}N_{\mathbf{a}}(t) = \mu_{b\to a}N_{\mathbf{b}}(t) - \mu_{a\to b}N_{\mathbf{a}}(t) + \xi_{\mathbf{a}}(t) \tag{1.5}$$

where $\xi_a(t)$ is the Gaussian random variable due to the sampling from a finite discrete subpopulation of size $N_{\mathbf{a}}$ with properties,

$$\langle \xi_{\mathbf{a}}(t)\rangle = 0 \quad \text{and} \quad \langle \xi_{\mathbf{a}}(t)\xi_{\mathbf{b}}(t')\rangle = N_{\mathbf{a}}(t)\delta(t-t')\delta_{\mathbf{a},\mathbf{b}} \tag{1.6}$$

$\langle . \rangle$ denotes the ensemble average. For $N \gg 1$, we can map the discrete variable $N_{\mathbf{a}}$ and $N_{\mathbf{b}}$ onto the continuous frequency variables $y_{\mathbf{a}} = N_{\mathbf{a}}/N$ and $y_{\mathbf{b}} = 1 - y_{\mathbf{a}}$. The stochastic term in the continuous coordinate will be related to the discrete noise by,

$$\chi(t) = \frac{\partial y}{\partial N_{\mathbf{a}}}\xi_{\mathbf{a}}(t) + \frac{\partial y}{\partial N_{\mathbf{b}}}\xi_b(t) \tag{1.7}$$

In this way, we can express the dynamics in eq. (1.5) as,

$$\frac{\mathrm{d}}{\mathrm{d}t}y_{\mathbf{a}}(t) = \mu_{\mathbf{b}\to\mathbf{a}}\,[1 - y_{\mathbf{a}}(t)] - \mu_{\mathbf{a}\to\mathbf{b}}\; y_{\mathbf{a}} + \chi(t) \tag{1.8}$$

with the Gaussian noise term $\chi(t)$,

$$\langle \chi(t)\rangle = 0 \quad \text{and} \quad \langle \chi(t)\chi(t')\rangle = \frac{1}{N}y_{\mathbf{a}}(t)[1 - y_{\mathbf{a}}(t)] \tag{1.9}$$

Fig. 1.2(a) shows such stochastic evolutionary dynamics in the population. The Langevin picture in eq. (1.8) corresponds to a Fokker-Planck equation for the probability density $P_0(y_{\mathbf{a}}, t)$; see e.g., the discussion on stochastic processes in (Gardiner, 2004).

$$\frac{\mathrm{d}}{\mathrm{d}t}\; P_0(y,t) = \frac{1}{2N}\frac{\partial^2}{\partial y^2}y(1-y)\; P_0(y,t) - \mu_{\mathbf{b}\to\mathbf{a}}\frac{\partial}{\partial y}(1-y)\; P_0(y,t) + \mu_{\mathbf{a}\to\mathbf{b}}\frac{\partial}{\partial y}y\; P_0(y,t) \tag{1.10}$$

Which results in Kimura's U-shape equilibrium solution shown in Fig. 1.2(b), (Kimura,

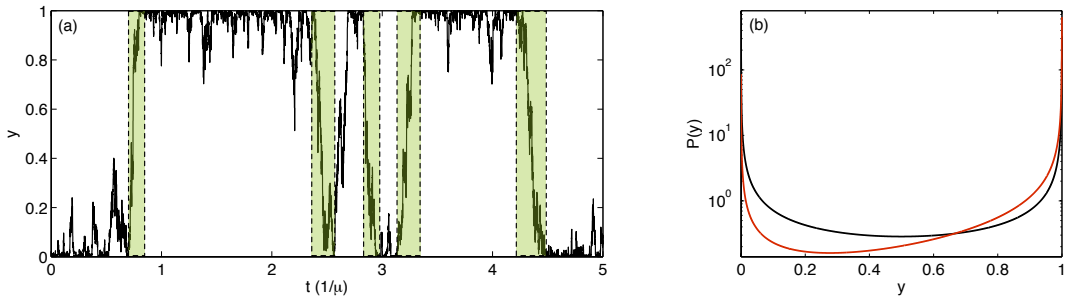

Figure 1.2: Stochastic evolutionary dynamics of a population. (a) Stochastic evolutionary dynamics of single locus allele frequency under selection in the low mutation regime $\mu N \ll 1$. Substitution events, starting from the initial mutations and ending by fixation, are highlighted in green. The time is measured in units of $1/\mu$.The parameters are chosen as, $\mu N = 0.05$ and $N\Delta F = 1$. (b) Stationary distribution of the allele frequencies y for evolutionary dynamics in neutral conditions (black), eq. (1.11) and under selection (red), eq. (1.12). The parameters are chosen as, $\mu N = 0.05$ and $N\Delta F = 0.2$.

1962).

$$P_0(y_{\mathbf{a}}) = \frac{1}{Z} y_{\mathbf{a}}^{-1+N\mu_{\mathbf{b}\to\mathbf{a}}} \, (1-y_{\mathbf{a}})^{-1+N\mu_{\mathbf{a}\to\mathbf{b}}} \tag{1.11}$$

Given the neutral equilibrium with mutation rates $\mu_{\mathbf{a}\to\mathbf{b}}$ and $\mu_{\mathbf{b}\to\mathbf{a}}$, the underlying distribution of an evolutionary process in a time-independent fitness landscape $Q(\mathbf{a})$ will also be in an equilibrium and have a simple relation to its neutral counterpart by a Boltzmann factor; see Fig. 1.2(b).

$$Q(\mathbf{a}) = \frac{1}{Z} P_0(\mathbf{a}) \exp[2N\overline{F}]. \tag{1.12}$$

where Z is the appropriate normalization factor and $\overline{F}$ is the mean population fitness which is the average population growth in the absence of mutations and genetic drift,

$$\frac{\mathrm{d}}{\mathrm{d}t} N(t) = \overline{F}(t) N(t) \tag{1.13}$$

We denote the malthusian fitness of the existing alleles by $F_{\mathbf{a}}$ for allele $\mathbf{a}$ and $F_{\mathbf{b}}$ for allele $\mathbf{b}$. The mean population fitness is therefore,

$$\overline{F}(t) = y_{\mathbf{a}}(t) F_{\mathbf{a}} + y_{\mathbf{b}}(t) F_{\mathbf{b}} = F_{\mathbf{b}} + y_{\mathbf{a}}(t) \Delta F_{ab} \tag{1.14}$$

with $\Delta F_{\mathbf{ab}} = F_{\mathbf{a}} - F_{\mathbf{b}}$. In this way, we can also compute the substitution rates $u_{\mathbf{a}\to\mathbf{b}}$ i.e., the rate of fixation of an allele in the population under selection (Kimura, 1968),

$$u_{\mathbf{a}\to\mathbf{b}} = N\mu_{\mathbf{a}\to\mathbf{b}}\frac{1-\exp(2\Delta F_{ab})}{1-\exp(2N\Delta F_{ab})} \tag{1.15}$$

For a binding site, we have argued that the relevant phenotype is the binding affinity $E(\mathbf{a})$ of the site sequence $\mathbf{a}$ to the transcription factor. Using a biophysical genotype-phenotype map, be it experimental measurements or statistical inference, we can project the evolutionary dynamics and the resulting distribution onto the binding phenotype,

$$P(E) = \sum_{\mathbf{a}} P(\mathbf{a})\delta(E(\mathbf{a}) - E) \tag{1.16}$$

where $\delta(x) = 1$ for $x = 0$ and 0, otherwise. The binding energy distribution in the whole genome $W(E)$ is a composition of the functional part $Q(E)$ and the background distribution $P_0(E)$, $W(E) = \lambda Q(E)+(1-\lambda)P_0(E)$. λ is a hidden Markov variable which determines the fraction of the functional sites in the genome. For a one-dimensional biophysical map from sequence to binding energy, the equilibrium state of the genotypes dictates an equilibrium state for the stationary phenotype distribution. Therefore, the relation between the neutral and the selective dynamics is,

$$Q(E) = P_0(E)\ \exp[2N\Delta F(E)] \tag{1.17}$$

We can infer the shape of the fitness landscape by comparing the phenotype (binding energy) distribution of the functional site sequences to that of the background genome. This inference will then enlighten us about the constraints imposed on binding site evolution. Fig. 1.3(a) shows the the comparison between the functional and neutral distributions of binding energy in the *E. coli* genome (Mustonen and Lässig, 2005). Not surprisingly, there is an over-representation of high-affinity (low binding energy) site sequences in the functional set compared to the genomic background. The fitness function, which is also the log-likelihood ratio of these two distributions, is then inferred in Fig. 1.3(b) (Mustonen and Lässig, 2005). Fitness is a highly nonlinear function of the binding energy which is related to the nonlinearity of the occupancy function in

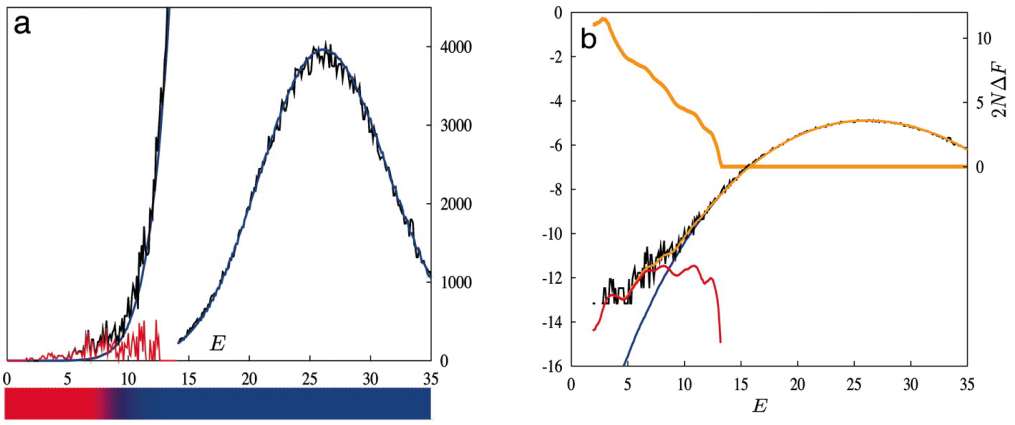

Figure 1.3: Binding energy distribution in E. coli genome. Energy statistics and fitness landscape for *CRP*-binding loci in E. coli. (a) Count histogram with energy bins of width 0.1 (black), expected background counts (blue), and excess counts above background (red), with a 30-fold zoom into the region $E < 14$. The color bar indicates the probability of functionality , ranging from 1 (red) to 0 (blue). (b) Decomposition of the counts (log-scale, left y axis) according to the hidden Markov model: background distribution $(1 - \lambda)P_0(E)$ (blue), distribution of functional loci $\lambda Q(E)$ (red), and total distribution $W(E)$. (orange). The resulting fitness landscape $\Delta F(E)$ according to eq. (1.17) is also shown in orange (thick curve, right y axis). The Figure is taken from (Mustonen and Lässig, 2005).

eq. (1.2). The immediate consequence of this nonlinearity is an asymmetry in the turnover of binding sites. Functional sites can rapidly loose their binding affinity to a transcription factor by one or two mutations. However, rapid formation of a site requires a seed sequence with marginal binding to which positive selection towards strong binding can latch on (Berg et al., 2004; Lässig, 2007). This is the topic we discuss further in Chapter. 2.

1.5 Thesis organization

The following topics are discussed in this thesis,

Chapter. 2: Binding site formation by local sequence duplication. In higher eukaryotes, genes often have complex regulatory input, which is encoded in regulatory sequence regions with multiple transcription factor binding sites. However, the modes of genome evolution generating this regulatory complexity are poorly understood. In

chapter. 2 we report a surprising finding: in fly regulatory modules, the majority of transcription factor binding sites show evidence of local sequence duplication in their evolutionary history, which relates their sequence information to that of neighboring binding sites. Our analysis suggests that local sequence duplications are a pervasive production mode of regulatory information. This mode appears to be specific to higher eukaryotes, and we have not found evidence of frequent local duplications in the yeast genome. Our results affect genomic sequence analysis, in particular, computational identification of cis-regulatory elements and alignment of regulatory DNA. At the same time, they address fundamental questions on the evolution of regulation: How much of the regulatory "grammar" observed in higher eukaryotes is due to the optimization of function, and how much reflects the underlying sequence evolution modes? What is the result and what is the substrate of natural selection? The content of this chapter is partly published by (Nourmohammad and Lässig, 2011).

Chapter. 3: Emergent selection on regulatory complexes. Individual binding sites in eukaryotic regulatory complexes have highly flexible binding affinities and relative positioning. Therefore, it is often difficult to assess which of the constituents are important for function. In particular, the functional role of low-affinity binding sites, which are found ubiquitously in eukaryotic genomes and can be produced by local duplications (Chapter. 2), has remained controversial. In Chapter. 3, we present a quantitative evolutionary analysis of such binding site complexes in yeast. These complexes consist of a strong binding site surrounded by a cloud of several low-affinity sites, whose functional importance has recently been demonstrated experimentally (Gertz et al., 2008). Based on the biophysical interactions of transcription factors and the sequence, we characterize a joint affinity phenotype for the regulatory complex which determines its functional output. We show that this collective binding phenotype is under substantial stabilizing selection and is well conserved within *Saccharomyces paradoxus* populations and between three species of *Saccharomyces*. At the same time, individual low-affinity sites evolve near-neutrally and show considerable affinity variation even within one population. We infer a fitness landscape depending on this phenotype using yeast whole-genome polymorphism data and a new method of quantitative trait analysis discussed in Chapter. 4. These quantitative studies suggest that functionality of and selection on regulatory complexes emerge from the entire cloud of sites, but

cannot be pinned down to individual sites.

Chapter. 4: Evolution of polygenic traits. In this chapter, we develop a novel theoretical framework to characterize the evolutionary dynamics of "quantitative traits" which are combinations of numerous loci all attributing to a common function. Detailed dynamics of such high-dimensional systems seems to be very complex. However, the extensive self-averaging properties of macroscopic trait observable, e.g., the average phenotype in a population, makes it possible to characterize the phenotypic composition in a population. This resembles the simple thermodynamic description of a gas despite its chaotic molecular composition. We introduce a coarse-grained description of a population by mapping its individual-based genomic components (such as allele frequencies) to population-based phenotype statistics. In this approach, characteristic parameters of the fitness landscape will be coupled to the statistics of the intra-population trait distribution and hence can be measured along with the macroscopic trait observables. As a result, we will suggest simple tests to infer the shape of the fitness landscape from phenotypic polymorphisms in a population. Our analysis covers both regimes of zero recombination (i.e., asexual genome with full linkage) and infinite recombination with perfect reassortments of genomic content. Given the current state of genomic data and advancements in high-throughput experimental techniques that produce quantitative genotype-phenotype maps, these methods can be put into practice. We present such genomic analysis in Chapter. 3.

2

Binding site formation by local duplications

2.1 Introduction

As we discussed in the previous chapter, the complex cis-regulatory information in higher eukaryotes is organized into *regulatory modules*, which are typically a few hundred base pairs long and are spatially separated by larger segments of intergenic DNA (Bergman and et. al., 2002; Ondek et al., 1988). Within modules, regulatory functions often depend on clusters of neighboring binding sites for multiple transcription factors, which are coupled by cooperative interaction (Davidson, 2006; Harbison and et. al., 2004; Lynch, 2006; Ptashne and Gann, 2002; Sinha et al., 2004). The resulted complex regulatory grammar ensures the specificity of regulation in the larger genomes of multicellular eukaryotes (Buchler et al., 2003; Levine and Tjian, 2003). At the same time, the grammar is flexible enough to allow substantial sequence evolution in a regulatory module while maintaining its overall functional output. On the other hand, the evolutionary modes of these modules are to be efficient to transports and produces cis-regulatory information.

In addition to point mutations, sequence insertions and deletions (indels) play a significant role in this dynamics. Several studies have noted the prevalence of repetitive sequence elements in promoter regions and their potential influence on regulatory function (Boeva et al., 2006; Britten, 1996; Gruen, 2006; Hancock et al., 1999; Messer and Arndt, 2007; Sinha and Siggia, 2005; Tanay and Siggia, 2008; Vinces et al., 2009).

In particular, a recent detailed analysis of the evolutionary rates of short *tandem repeats* in *Drosophila* has shown a net surplus of insertions, suggesting that these repeats may produce new regulatory sequence (Sinha and Siggia, 2005). Short tandem repeats are common sequence patterns in DNA where nucleotide segments with length of about 3-6 base paris are repeated and repetitions are directly adjacent to each other. But to what extent in this case do these sequence entities produce regulatory information? A priori, the link between repeat evolution and regulation is far from obvious: Duplications in repeats can either be part of the neutral background evolution in regulatory sequences, or increase the spacing between existing binding sites of a regulatory module, or contribute to the formation of new sites. Disentangling these roles is subtle, because detected tandem repeats in contemporary sequence overlap with only a small fraction of binding sites, motif size and total length of most repeats are shorter than length and spacing of typical binding sites in a cluster, and repeat lifetimes are much shorter than conservation times of regulatory elements (Gruen, 2006). Hence, the role of repeat dynamics for regulation is an open problem: Do local duplications actually transport and produce regulatory information?

This is the topic of the present chapter. We show that local duplications have left a striking signature in the fly genome: the majority of transcription factor binding sites in regulatory modules show evidence of a duplication event in their evolutionary history. We conclude that over long evolutionary times, local duplications are pervasive and crucial for the formation of complex regulatory modules in the fly genome. This mode of evolution sets the speed of regulatory evolution and facilitates adaptive changes of promoter function. We infer site duplications from their traces in the sequence of neighboring binding sites, but most duplication events predate the tandem repeats present in contemporary sequence. This distinguishes our study from comparative analysis of regulatory sequence between closely related species (Boeva et al., 2006; Gruen, 2006; Messer and Arndt, 2007; Sinha and Siggia, 2005; Tanay and Siggia, 2008), which can detect the insertion-deletion dynamics of contemporary repeats, but cover only a small window in the evolution of regulatory sites.

The importance of binding site evolution by duplication is grounded in the biophysics of transcription factor-DNA interactions: the sequence-dependent probability of binding between factor and site depends in a strongly nonlinear way on the binding energy (Berg and von Hippel, 1987): it takes values close to 1 in an energy range below

the maximum binding energy, then drops rapidly as the energy decreases further, and is close to 0 in the energy range of non-binding sites; see Section. 1.3. This nonlinearity generates strong epistatic effects for point mutations within binding sites (Berg et al., 2004; Mustonen et al., 2008) and, in turn, an asymmetry in the turnover of binding sites. Functional sites can rapidly lose their binding affinity to a factor by one or two point mutations. Rapid adaptive formation of a site, however, requires a *seed sequence* with marginal binding, to which positive selection for point mutations towards stronger binding can latch on. Such seeds are contained in random sequence, but at unspecific positions. Estimates of the rate of site formation based on biophysically grounded fitness models suggest that point mutations alone can explain the rapid formation of an individual site in a sufficiently large sequence interval, but not the formation of spatially confined agglomerations of sites characteristic of regulatory modules (Berg and Lässig, 2003; Berg et al., 2004; Lässig, 2007). As we show in this chapter, local sequence duplications generate seeds for new sites specifically in the neighborhood of functional sites.

Our analysis in this chapter proceeds in three steps. First, we analyze local sequence similarities in regulatory regions of the *Drosophila melanogaster* genome in a model-independent way. In regulatory modules, we find a significant autocorrelation in nucleotide content for distances up to about 70 bp. This autocorrelation includes the known contributions of tandem repeat sequences, but it extends to a much larger distance range. The signal turns out to be generated by local sequence clusters, a substantial fraction of which are functional transcription factor binding sites with similar sequence motifs. In the second part of the paper, we turn specifically to binding sites: we infer the evolutionary origin for pairs of neighboring sites, using a known set of validated sites and a probabilistic model with mutations, genetic drift, and selection. The model compares the likelihood of two alternative histories: a pair of sites evolves either independently or by duplication from a common ancestor sequence. The duplication is followed by diversification under selection for binding of two (in general different) factors. We show that the duplication pathway is the most likely history for pairs of sites with a mutual distance up to about 50 bp. Furthermore, we find evidence that this pathway is specific to regulatory modules of multicellular eukaryotes. Finally, we show that the duplication mode has adaptive potential: duplicated ancestor sites can

act as seeds for the subsequent formation of a novel binding site for the same factor and, notably, even for a different factor.

2.2 Statistics of sequence similarity in Regulatory DNA

2.2.1 Sequence autocorrelation in regulatory DNA

The most straightforward measure of local similarity in a sequence segment is the *autocorrelation function*, which is defined as the difference between the likelihood that two nucleotides at a distance of r base pairs are identical and mean identity of two random nucleotides. In a given sequence segment $a_1, \ldots, a_L$, the nucleotide frequencies are given by

$$p_0(a) = \frac{1}{L} \sum_{\nu=1}^{L} \delta(a_\nu, a), \tag{2.1}$$

where $\delta(a_\nu, a) = 1$ if $a_\nu = a$ and $\delta(a_\nu, a) = 0$ otherwise. These determine the mean similarity between two random nucleotides of the segment, $c_0 = \sum_a p_0^2(a)$. The sequence autocorrelation function is then defined by,

$$\Delta(r) = -c_0 + \frac{1}{L-r} \sum_{\nu=1}^{L-r} \delta(a_\nu, a_{\nu+r}). \tag{2.2}$$

The distance dependence of the autocorrelation signal provides information about the range, within which the nucleotides appearing in the sequence are correlated. This function is straightforward to evaluate from sequence data. We have obtained the autocorrelation function in 346 regulatory modules of the *D. melanogaster* genome with length of more than 1000 bp identified by REDfly database (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008). The results are shown in Fig. 2.1 (a). In the distance range up to about 70 bp, the function $\Delta(r)$ takes positive values that decay with r in a roughly exponential way; this signal is clearly above the noise level. The mean identity is evaluated in a local window of 500 bp (changing the window length affects the baseline of this function, but not its short-distance behavior). The autocorrelation signal is small and has several potential sources, such as multiple binding sites for similar motifs, tandem repeats at short length scales (Gruen, 2006; Messer and Arndt, 2007; Sinha and Siggia, 2005; Tanay and Siggia, 2008), homopolymeric stretches of

nucleotides characteristic of nucleosome-depleted regions (Segal and Widom, 2009), or other local inhomogeneities in sequence composition.

Information about the spatial distribution of correlated nucleotides along the genome is contained in higher orders of sequence autocorrelation (i.e., reoccurrence of doublets, triplets, etc.). As a next step, we use information theory to identify such clusters of correlated nucleotides in a sequence region. We will characterize local sequence similarity in a more specific way: we will show that mutually correlated nucleotide pairs are not evenly distributed over regulatory modules, but occur in local clusters with a characteristic length scale of around 7 bp. This signal will be analyzed from an evolutionary point of view and be linked to cis-regulatory function.

2.2.2 Sequence motifs and information

To motivate the following analysis, assume that a given sequence segment is covered by families of sites belonging to different *motifs* which are reoccurring nucleotide patterns. A motif of length ℓ is a probability distribution $Q(\mathbf{a})$ of genotypes $\mathbf{a} = (a_1, \ldots, a_\ell)$, which describes a specific set of sequence sites with ℓ consecutive base pairs and is significantly different from the background distribution $P_0(\mathbf{a})$. If we neglect correlations between nucleotides, we can write these distributions as the product of single-nucleotide frequencies,

$$Q(\mathbf{a}) = \prod_{i=1}^{\ell} q_i(a_i) \tag{2.3}$$

and $P_0(\mathbf{a}) = \prod_{i=1}^{\ell} p_0(a_i)$. The $4 \times \ell$ matrix of single-nucleotide frequencies (2.3) is called the position weight matrix of the motif (similar to that shown for transcription factor motifs in Chapter. 1). The *sequence information* of the motif is measured by the relative entropy (Kullback-Leibler distance) between these distributions (Kullback and Leibler, 1951),

$$H(Q|P_0) = \sum_{i=1}^{\ell} \sum_{a} q_i(a) \log \frac{q_i(a)}{p_0(a)}. \tag{2.4}$$

This quantity measures the statistical deviation of the motif pattern from the background and determines the average *sequence information* per site, which is often quoted in units of bits (Stormo and Fields, 1998). Multiplying $H(Q|P_0)$ with the number of

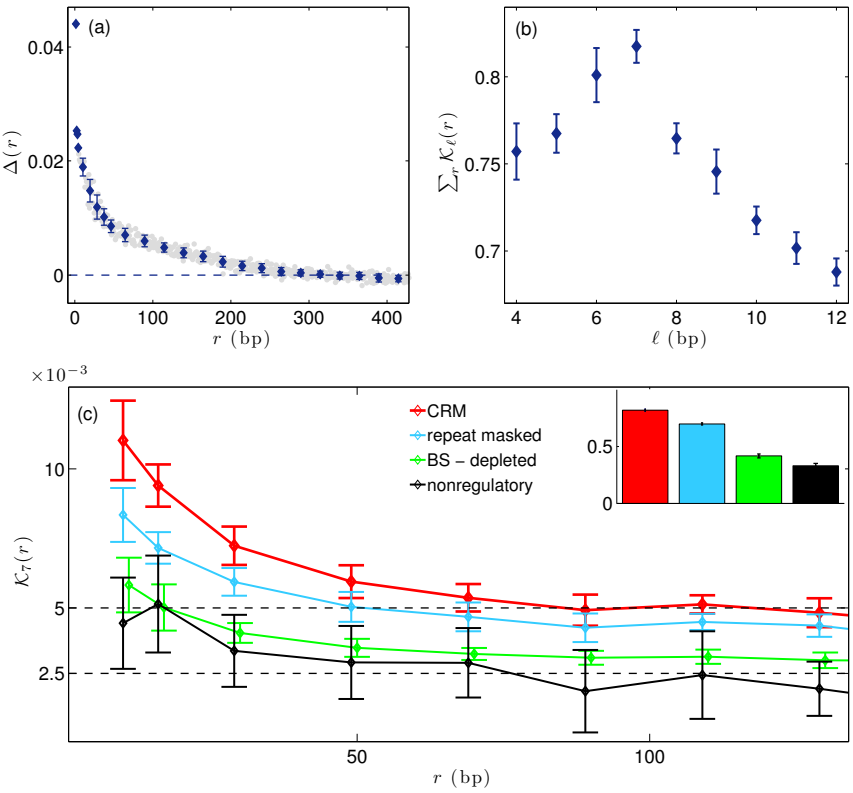

Figure 2.1: Sequence similarity in regulatory modules of the fly genome. (a) Sequence autocorrelation $\Delta(r)$ as a function of distance r, obtained from 346 regulatory modules in *D. melanogaster* (gray: unbinned data, blue: binned in intervals of variable length). The autocorrelation values are positive and depend on r in a roughly exponential way up to about 70 bp. (b) Total similarity information $\mathcal{K}_{\rm tot}(\ell) = \sum_{r=1}^{100} \mathcal{K}_\ell(r)$ as a function of motif length ℓ for all pairs of strongly correlated sites with mutual distance $r < 100$ bp in the same set of regulatory modules. This function takes its maximum at a characteristic motif length of $\ell = 7$ bp. (c) Distance-dependent similarity information $\mathcal{K}_7(r)$ for motif length $\ell = 7$ evaluated in all sequence (red), binding site-masked sequence (green), repeat-masked sequence (blue) in regulatory modules, and in generic intergenic sequence (black). Repeat-masked sequence is generated using the Tandem Repeat Finder (Benson, 1999) with match-mismatch-indel penalty parameters (2,3,5). Insert: Total similarity information $\mathcal{K}_{\rm tot}(\ell = 7)$ for the same sequence categories. Binding sites, but not tandem repeats, account for a substantial fraction of the similarity information.

sites for each motif and summing over all motifs produces a measure of the total sequence information contained in a genomic region.

Well-known motifs in regulatory DNA are the families of binding sites for a given

transcription factor. In eukaryotic systems, these sites have a typical length of about 5-10 bp and frequency distributions Q (called position weight matrices) with a typical information content $H \approx 6 - 8$ bits per site; see the recent discussion by(Wunderlich and Mirny, 2009) and the introductory discussion in Section. 1.2. Other motifs can be defined, for example, in nucleosome-depleted sequences in eukaryotes and for repeat units in tandem repeats. If all motifs occurring in a given sequence segment were known, we could try to predict their sites and evaluate the information content directly. In the present part of the analysis, we proceed differently. We only assume that sequence motifs carry a certain information content over sites of a given length ℓ, but we make no further assumptions on position weight matrices, sequence coverage, or evolutionary origin. Hence, even without any prior knowledge on frequency distributions, we can recover part of the sequence information for those motifs that occur more than once in the sequence segment. A pair of sites of length ℓ belonging to the same motif has an average *similarity information* given by the relative entropy $K(c, \ell|c_0)$, which measures the enhanced similarity c of aligned nucleotides of the site sequences compared to the background similarity c_0,

$$K(c, \ell|c_0) = \ell \left[c \log \frac{c}{c_0} + (1 - c) \log \frac{1 - c}{1 - c_0} \right]. \tag{2.5}$$

Clearly, the similarity information between pairs of sites is a somewhat diluted measure of the full information content due to motifs. As a rule of thumb, the mutual entropy per site pair, $K(c, \ell|c_0)$, recovers about half of the sequence information per site, $H(Q|P_0)$. For example, binding sites for the same transcription factor are strongly correlated, with a typical similarity $c \approx 0.7$ and a similarity information $K \approx 3$ bits per site pair whereas, the information content of a binding motif is typically $H \approx 6 - 8$ bits per site.

With this approach, we want to identify pairs of similar sites at a given distance r and relate them to the sequence autocorrelation function $\Delta(r)$ discussed above. Thus, we estimate the total similarity information $\mathcal{K}_\ell(r)$ per unit sequence length of all strongly correlated pairs of sites with distance r and length ℓ in regulatory modules.

Figure 2.2: Motif detection in sequence segments (schematic). The figure shows a configuration of correlated sequence sites of length $\ell = 10$ bp and distance $r = 14$ bp from each other. Pairs of correlated sites have the following properties: (i) The average mutual similarity between aligned nucleotides is larger than a given threshold, $c \geq c_{\min} = 0.8$. (ii) The left sites (and, hence, also the right sites) of all pairs have no common nucleotides. This condition is necessary in order to avoid overcounting of mutual similarity in overlapping site pairs. (iii) The sum of the mutual similarities of all pairs in the set is maximal. In the example shown, there are three different motifs with reoccurring sequence patterns marked by different colors (red, blue, green). To illustrate the alignment of the site pairs, we shift the whole sequence by $r = 14$ bp in the second row. The left and right site of each motif are shown in boldface in the first and the second row, respectively. Mismatches between aligned sites of the same motif are shown in boldface gray letters. The flanking regions separating the correlated sequence pairs are shown in smaller font.

This quantity can be defined by constructing a set of site pairs for given r and ℓ,

$$\{(a_{\nu_1}, \ldots, a_{\nu_1+\ell-1}), (a_{\nu_1+r}, \ldots, a_{\nu_1+r+\ell-1})\}, \ldots, \{(a_{\nu_n}, \ldots, a_{\nu_n+\ell-1}), (a_{\nu_n+r}, \ldots, a_{\nu_n+r+\ell-1})\} \tag{2.6}$$

with the following properties:

(i) The left sites (and, hence, also the right sites) of all pairs have no mutual overlaps,

$$\nu_{\alpha+1} - \nu_\alpha \geq \ell \quad \text{for } \alpha = 1, \ldots, n-1. \tag{2.7}$$

This condition is necessary in order to avoid overcounting of mutual similarity in overlapping site pairs.

(ii) The mean mutual similarity of each site pair is greater than a threshold $c_{\min}$,

$$c_\alpha \equiv \frac{1}{\ell} \sum_{i=1}^{\ell} \delta(a_{\nu_\alpha+i}, a_{\nu_\alpha+r+i}) > c_{\min} \quad \text{for } \alpha = 1, \ldots, n. \tag{2.8}$$

(iii) The sum of mutual similarities $\sum_{\alpha=1}^{n} c_\alpha$ is maximal.

Fig. 2.2 illustrates this procedure. To identify a set of site pairs with properties (i) to (iii), we use a dynamic programming algorithm with a recursion,

$$C_t = \max[C_{t-1}, C_{t-\ell} + [\frac{1}{\ell}\sum_{i=1}^{\ell}\delta(a_{t-\ell+i-r}, a_{t-\ell+i})] - c_{\min}], \tag{2.9}$$

we obtain the sequence of partial scores $C_1, \ldots, C_L$ with the initial condition $C_1 = 0$. We then use a backtracking procedure (see, e.g., (Durbin et al., 1998)) to determine the set of positions $(\nu_1, \ldots, \nu_n)$ of the high-similarity pairs (2.6). In the maximum-similarity set, we record the average mutual similarity $\bar{c}(r, \ell)$ of aligned nucleotides in site pairs, which determines the mean information content per site pair, $K(\bar{c}(r, \ell), \ell | c_0)$ (see eq. (2.5). We also record the number $n(r, \ell)$ of site pairs and compare to the number expected by chance in background sequence, $n_0(\ell)$. To estimate the expected number of pairs in background sequence, we apply the same procedure to 1000 sequences of length L, which are generated by a first-order Markov model with the same single-nucleotide frequencies $p_0(a)$ and conditional frequencies $T(a|b)$ as in the actual sequence,

$$P(a_1, \ldots, a_L) = p_0(a_1)\prod_{\nu=2}^{L} T(a_\nu | a_{\nu-1}) \tag{2.10}$$

We then evaluate the excess $\Delta n(r, \ell, c_{\min}) = n(r, \ell, c_{\min}) - n_0(r, \ell, c_{\min})$ and obtain an estimate of the total information contained in the enhanced autocorrelation of motifs as given by eq. (2.5),

$$\mathcal{K}_\ell(r) = \ell \max_{c_{\min}} \left[\frac{\Delta n(r, \ell, c_{\min})}{L} \left(\log \frac{\bar{c}(r, \ell, c_{\min})}{c_0} + \log \frac{1 - \bar{c}(r, \ell, c_{\min})}{1 - \bar{c}_0} \right) \right]. \tag{2.11}$$

We infer $c_{\min}$ by maximum likelihood analysis of the total similarity information in the sequence. This method also allows for optimization of the motif length ℓ, similar to the procedure in the local sequence alignment algorithms (Durbin et al., 1998).

Our analysis is limited to known regulatory modules and focuses on the dependence of $\mathcal{K}_\ell(r)$ on r and ℓ. A specific part of this signal, obtained from sites with distance r below 50 bp, will be associated below with local duplications as prevalent evolutionary mode.

2.2.3 Similarity information in regulatory modules of *Drosophila*

We evaluate the similarity information in the set of 346 regulatory modules of *Drosophila melanogaster* and in surrounding background sequence. The following features of local

sequence similarity can be extracted:

—*The total information of local sequence similarity is maximal for motifs of length* $\ell = 7$. Fig. 2.1(b) shows the total similarity information of all detected site pairs in the range of up to 100 bp, $\mathcal{K}_{\rm tot}(\ell) = \sum_{r=\ell}^{100} \mathcal{K}_\ell(r)$, as a function of the site length ℓ. The function $\mathcal{K}_{\rm tot}(\ell)$ takes its maximum, that is, the similarity information is most significant, for $\ell = 7$. The signal falls off at shorter length scales, because typical motif sequences are only partially covered, and at larger length scales, because uncorrelated flanking nucleotides contribute negatively to the similarity information. In this sense, detected motifs cover a characteristic length of about 7 bp. A similar length scale has been observed in tandem repeats (Boeva et al., 2006; Messer and Arndt, 2007; Tanay and Siggia, 2008).

—*The function* $\mathcal{K}_7(r)$ *takes distance-dependent positive values in the range of up to 50 bp and saturates to a positive asymptotic value for larger distances.* Thus, its distance dependence is compatible to that of the sequence autocorrelation function $\Delta(r)$ shown in Fig. 2.1(a). This pattern is due to site pairs with high mutual similarity, $c > 0.85$.

—*Correlated binding sites explain a substantial part of the similarity information.* We estimate this contribution by masking all functional sites (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008) and re-evaluating the function $\mathcal{K}_7(r)$ in their sequence complement; see Fig. 2.1(c). Known binding sites cover about 10% of the regulatory modules, but the signal is reduced by about 50%, indicating that these sites are an important source of similarity information. The binding site-masked signal is comparable to its counterpart $\mathcal{K}_7(r)$ in non-regulatory intergenic sequence.

—*Short tandem repeats explain only a small part of the similarity information.* We identify such repeats using the Tandem Repeat Finder (Benson, 1999). If we remove about 5% of the sequence in regulatory modules as repeats, the similarity information is reduced by less than 10%; Fig. 2.1(c). This is not surprising, because our sequence similarity measure differs from that of repeat analysis. In particular, our measure is sensitive to correlated segments on larger distance scales than typical tandem repeats,

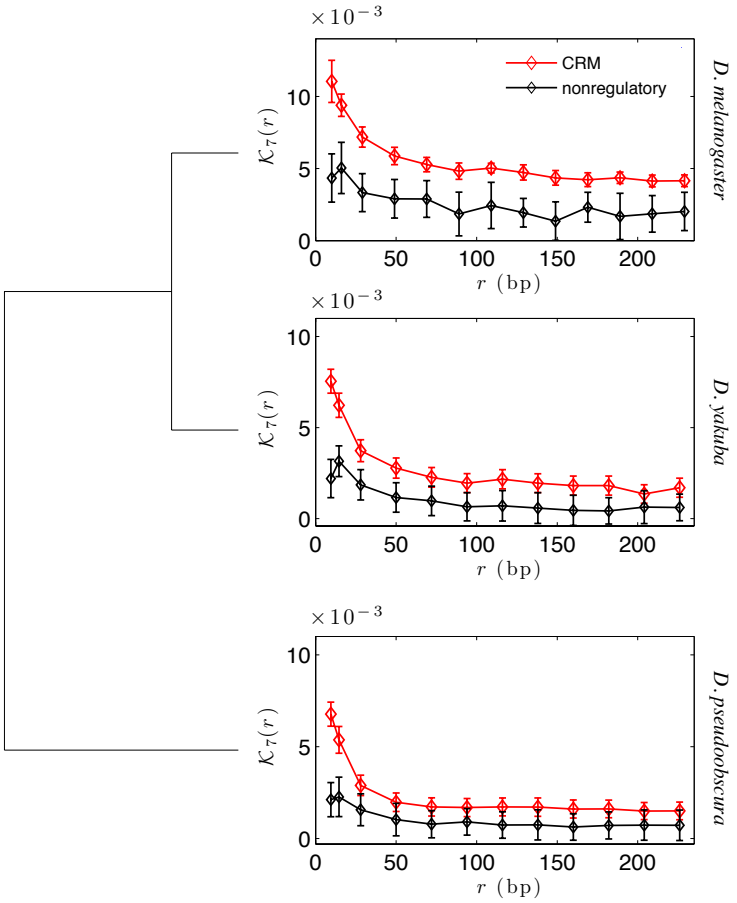

Figure 2.3: Sequence similarity in regulatory modules of 3 *Drosophila* species. Distance-dependent similarity information $K_7(r)$ for motif length $\ell = 7$ in regulatory modules (red) and in generic intergenic sequence (black), evaluated in *D. melangaster* and in the homologous regions of *D. yakuba* and *D. pseudoobscura* (see Materials and Methods in Section. 2.5). These data show a consistent pattern of overall amplitudes and of decay lengths.

because it does not require a contiguous interval of self-similar sequence in between.

—*Homologous regions in other fly genomes show a consistent form of* $\mathcal{K}_7(r)$. We analyze homologous regions of two other *Drosophila* species, *D. yakuba* and *D. pseudoobscura*; see Materials and Methods in Section. 2.5. As shown in Fig. 2.3, these putative regulatory modules have patterns $\mathcal{K}_7(r)$ of very similar overall amplitude and distance-dependence, with enhanced values in the range of up to 50 bp.

In summary, our model-independent analysis shows that motifs with a characteristic length of about 7 bp play an important part in the distance-dependent sequence autocorrelation of *Drosophila* regulatory modules. The characteristic length coincides with the typical length of binding sites, and a substantial fraction of the signal can be explained by sequence correlations involving known binding sites. Therefore, we now focus the analysis on a smaller, but experimentally validated set of sites (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008). This allows us to analyze in detail the evolutionary mechanism generating the sequence similarity between neighboring sites.

2.3 Evolutionary modes of binding sites

Binding sites are ideal objects to study the production of information by sequence evolution. The sequence motif in the form of a position weight matrix, is approximately known for about 70 transcription factors in *Drosophila*. Thus, we can analyze the full position-dependent sequence information of these motifs, not just the similarity information of motif pairs. Furthermore, there is a simple link between sequence statistics and evolution of binding sites: assuming the sequence distribution Q defines a motif at evolutionary equilibrium, its sequence information H is proportional to the average fitness effect of its binding sites,

$$N\langle F\rangle = H(Q|P_0) = \sum_{\mathbf{a}} Q(\mathbf{a}) \log \frac{Q(\mathbf{a})}{P_0(\mathbf{a})} \tag{2.12}$$

with a proportionality constant equal to the effective population size (Berg and Lässig, 2003; Berg et al., 2004; Moses et al., 2003, 2004). The fitness contribution of a particular binding sequence, $F(\mathbf{a})$, is proportional to its log-likelihood ratio in the distributions Q and P_0. The ensemble of these fitness values defines an information-based *fitness landscape* F for binding of a specific transcription factor. These relations between sequence statistics and fitness of binding sites quantify our intuition that specific sequences are overrepresented in a motif to the extent they confer a selective advantage over random sequences (Stormo and Fields, 1998). If we write the motif distribution Q in the product form of a position weight matrix, we obtain an approximate expression

for the fitness $F(\mathbf{a})$ in terms of the position-specific single-nucleotide frequencies $q_i(a)$ in the motif sequence and their counterparts $p_0(a)$ in background sequence:

$$NF(\mathbf{a}) = \sum_{i=1}^{\ell} f_i(a_i) \quad \text{with } f_i(a) = \log \frac{q_i(a)}{p_0(a)}. \tag{2.13}$$

This expression, which is in its simplest form already contained in Kimura's U-shaped equilibrium distribution for a two-allele locus (Fig.1.2(b) (Kimura, 1962)), is known as Bruno-Halpern model in the context of protein evolution (Halpern and Bruno, 1998) and has been used to infer fitness effects of mutations in binding sites (Berg and Lässig, 2003; Berg et al., 2004; Lässig, 2007; Moses et al., 2003, 2004; Mustonen and Lässig, 2009). Although this additive fitness model neglects fitness interactions between nucleotides within binding sites as well as between sites within a regulatory module, it is justified for the purpose of this study (see below).

The fitness landscape F defines the selection coefficient of any change from a state $\mathbf{a}$ to a state $\mathbf{b}$ of a binding site, $\Delta F_{\mathbf{ab}} = F(\mathbf{b}) - F(\mathbf{a})$. Here, we use the standard Kimura-Ohta formalism to infer the rates $u_{\mathbf{a}\to\mathbf{b}}$ of point substitutions $\mathbf{a} \to \mathbf{b}$ from the fitness model and the point mutation rates $\mu_{\mathbf{a}\to\mathbf{b}}$ (see Section. 1.4),

$$u_{\mathbf{a}\to\mathbf{b}} = \mu \frac{N\Delta F_{\mathbf{ab}}}{1 - \exp(-N\Delta F_{\mathbf{ab}})}, \tag{2.14}$$

For simplicity, the mutation rates are assigned a uniform value $\mu_{\mathbf{a}\to\mathbf{b}} = \mu_{\mathbf{b}\to\mathbf{a}} = \mu$. This relation is valid in the regime $\mu N \ll 1$ (in which subsequent substitution processes are unlikely to overlap in time) and $\Delta F_{\mathbf{ab}} \ll 1$ (Kimura, 1962; Kimura and Ohta, 1969). The matrix of these substitution rates then determines the transition probabilities (propagators) $G^{\tau}(\mathbf{b}|\mathbf{a})$ from an ancestor site $\mathbf{a}$ to a descendent site $\mathbf{b}$ through a series of point substitutions within an evolutionary distance τ (Durbin et al., 1998; Mustonen and Lässig, 2005).

Here, we use this quantitative sequence evolution model to infer modes of binding site evolution. For any given pair of adjacent sites $\mathbf{a}$ and $\mathbf{b}$ that bind transcription factors A and B, respectively, we want to evaluate the likelihood of two different histories of site formation.

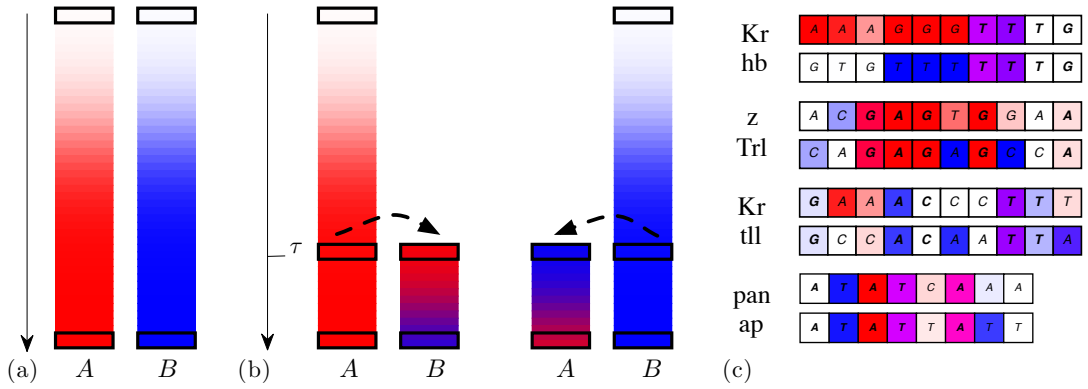

Figure 2.4: Evolutionary modes of transcription factor binding sites. The figure shows alternative formation histories for two adjacent binding sites, whose present sequences bind transcription factors A and B, respectively. The color coding indicates the evolution of binding function for factor A (red) and B (blue) with evolutionary time t. (a) Evolution from independent ancestor sequences. The sites evolve to their present states by independent evolutionary processes under stationary selection given by different fitness landscapes F_A and F_B (see text). In this mode, adjacent sites will show no enhanced average sequence similarity compared to the similarity of their motifs. (b) Evolution by duplication of a common ancestor sequence. Left panel: The original site evolves in the stationary fitness landscape F_A. At a distance τ from the present, this site undergoes a duplication. The duplicated site evolves its new function of binding B in the fitness landscape F_B. Right panel: The same process with the roles of A and B interchanged. In the duplication mode, the sites retain an enhanced sequence similarity, which reflects their common descent. (c) Examples of adjacent functional binding sites with enhanced sequence similarity in the *D. melanogaster genome*. The sites of each pair are aligned. The color background of nucleotide a at position i indicates its contributions to fitness (binding affinity) for factor A and B, i.e., $f_{i,A}(a)$ (level of red) and $f_{i,B}(a)$ (level of blue). The sequence similarity leads to hybrid binding characteristics: some nucleotides of the A-site (top row) have binding characteristics of the B-motif, and vice versa. Examples from top to bottom (factor A / factor B, genomic positions, duplication score): (i) *Kruppel* / *hunchback*, chr3L: 8639822 / 8639878, $S = 4.40$, (ii) *zeste* / *Trithorax-like*, chr3R: 12560236 / 12560218, $S = 3.97$, (iii) *Kruppel* / *tailless*, chr3L: 8639586 / 8639596, $S = 3.40$, (iv) *pangolin* / *apterous*, chr3R: 22997722 / 22997752, $S = 2.38$.

(i) Evolution from independent ancestors. In the first mode of evolution, the sites are assumed to evolve to their present sequence states by point substitutions from independent ancestor sequences and under independent selection given by the fitness landscapes F_A and F_B, as illustrated in Fig. 2.4(a). If the selection for binding is assumed to act over a sufficiently long evolutionary time, the probability of observing the present sequence states **a** and **b** in this independent mode of evolution is simply $Q_A(\mathbf{a})Q_B(\mathbf{b})$. This mode of evolution can only result in distance-dependent sequence

similarity arising from an increased coverage with pairs of adjacent sites with correlated motifs Q_A and Q_B (evidence for this effect will be discussed below). However, it does not generate increased similarity of individual pairs of adjacent sites beyond that of their motifs.

(ii) Evolution from a common sequence ancestor. In the second mode of evolution, the sites are assumed to evolve from a common ancestor sequence by a local duplication event at a distance τ from the present, followed by diversification under selection given by separate fitness landscapes F_A and F_B: either the original site is under stationary selection for binding factor A and the duplicated site has evolved the new function of binding the $B-$factor or vice versa, as illustrated in Fig. 2.4(b). In this mode, the present sequences $\mathbf{a}$ and $\mathbf{b}$ have evolved from their last common ancestor $\mathbf{c}$ by independent substitution processes with transition probabilities $G_A^\tau(\mathbf{a}|\mathbf{c})$ and $G_B^\tau(\mathbf{b}|\mathbf{c})$. The dynamics results in a joint probability of the form,

$$Q^\tau(\mathbf{a},\mathbf{b}) = \sum_{\mathbf{c}} G_A^\tau(\mathbf{a}|\mathbf{c}) G_B^\tau(\mathbf{b}|\mathbf{c}) Q(\mathbf{c}) \tag{2.15}$$

To proceed, we first assume that the ancestor site $\mathbf{c}$ is at evolutionary equilibrium under selection to bind factor A, that is, the contemporary site $\mathbf{a}$ has the ancestral function and $\mathbf{b}$ has evolved a new function after duplication. This gives the contribution,

$$\begin{aligned} Q_A^\tau(\mathbf{a},\mathbf{b}) &= \sum_{\mathbf{c}} G_A^\tau(\mathbf{a}|\mathbf{c}) G_B^\tau(\mathbf{b}|\mathbf{c}) Q_A(\mathbf{c}) \\ &= \sum_{\mathbf{c}} G_B^\tau(\mathbf{b}|\mathbf{c}) G_A^\tau(\mathbf{c}|\mathbf{a}) Q_A(\mathbf{a}), \end{aligned} \tag{2.16}$$

where we have used the detailed balance condition of the substitution dynamics, i.e., $G_A(\mathbf{a}|\mathbf{c})Q_A(\mathbf{c}) = G_A(\mathbf{c}|\mathbf{a})Q_A(\mathbf{a})$ (Mustonen and Lässig, 2005). There is a second contribution $Q_B^\tau(\mathbf{b},\mathbf{a})$ describing the case of the ancestor $\mathbf{c}$ under stationary selection to bind factor B. Weighing these cases with equal prior probabilities, $Q(\mathbf{c}) = [Q_A(\mathbf{c}) + Q_B(\mathbf{c})]/2$, we obtain

$$Q^\tau(\mathbf{a},\mathbf{b}) = \frac{1}{2}\big[Q_A^\tau(\mathbf{a},\mathbf{b}) + Q_B^\tau(\mathbf{b},\mathbf{a})\big]. \tag{2.17}$$

In this mode, distance-dependent sequence similarity arises due to common descent, causing the sequences of adjacent sites to be more similar than their motifs Q_A and Q_B. Importantly, this effect is generic and not tied to any functional properties of the transcription factors A and B. Fig. 2.4(c) shows a few examples of enhanced sequence similarity in pairs of adjacent binding sites in regulatory modules of *D. melanogaster*.

The relative likelihood of common versus independent descent for a specific pair of sites $\mathbf{a}, \mathbf{b}$ is given by the *duplication score*,

$$S(\mathbf{a}, \mathbf{b}) = \log \frac{Q^\tau(\mathbf{a}, \mathbf{b})}{Q_A(\mathbf{a}) Q_B(\mathbf{b})} \tag{2.18}$$

The information about common or independent descent comes from the similarity between the sequences $\mathbf{a}$ and $\mathbf{b}$ in a gapless alignment. The particular feature of site sequences is that they have evolved under selection for the binding motifs of the transcription factors A and B. Therefore, our score measures the similarity between the sequences $\mathbf{a}$ and $\mathbf{b}$ in a specific way: it gauges matches and mismatches depending on the weights of aligned nucleotides in their respective binding motifs Q_A and Q_B. For example, a match gets low score if it concurs with a common preferred nucleotide of the motifs, and high score if it goes against the preferred nucleotide of at least one of the motifs. A positive score value indicates that the pair $\mathbf{a}$, $\mathbf{b}$ is more likely to have evolved by duplication from a common ancestor sequence than independently.

The similarity score in eq. (2.18) depends on the evolutionary distance parameter τ. We infer the optimal value of τ by maximizing the likelihood ratio between the score distribution estimated from the set of closeby site-pairs (with mutual distance $r < 50$) and the score distribution of pairs with independent origin. This parameter describes the expected excess similarity of site pairs related by common descent, but it is not a linear clock of divergence time and should be regarded as model fit parameters for the observed sequence similarities. Energy-based fitness models (Mustonen et al., 2008; Mustonen and Lässig, 2005), which take into account the epistasis between mutations within binding sites, are required to obtain more accurate estimates of τ, which can be tested against phylogenetic data. Epistasis will increase the inferred values of τ compared to the additive (Bruno-Halpern) model (Mustonen et al., 2008; Mustonen and Lässig, 2005).

Below, we use the distribution $W(S)$ of duplication scores to infer the mode of evolution prevalent in a given class of site pairs. We evaluate the score distribution $W(S)$ of a given class of site pairs in terms of a mixture model of common and independent descent,

$$W(S) = (1-\lambda)Q_0(S) + \lambda Q(S). \tag{2.19}$$

The distribution of scores for independent descent, $Q_0(S)$, is obtained from pairs of sites in a common module with a relative distance $r > 200$ bp (Fig. 3(a), dashed line). This distribution is approximately Gaussian and has a width of order one. Because we build Q_0 from sites in a common module, its score average is above that for pairs of sites located in different modules. In this way, the overall sequence similarity within modules, which depends on the local GC-content, is assigned to the background model and does not confound the evidence for common descent. The distribution $Q(S)$ is the best fit to the the large-score excess of the distribution $W(S)$ for adjacent sites with a relative distance $r < 50$ bp (Fig. 3(a), violet-shaded).

Given a set of k site pairs $(\mathbf{a}, \mathbf{b})$ with scores $S(\mathbf{a}, \mathbf{b})$ described by the distribution $W(S)$, the log-likelihood of the mixed-descent model (2.19) relative to the independent-descent background model is given by

$$\begin{aligned} \Sigma = k\, H(W|Q_0) &= \sum_{\text{site pairs}} \log\left[\frac{W(S(\mathbf{a},\mathbf{b}))}{Q_0(S(\mathbf{a},\mathbf{b}))}\right] \\ &= \sum_{\text{site pairs}} \log\left[(1-\lambda) + \lambda\frac{Q(S(\mathbf{a},\mathbf{b}))}{Q_0(S(\mathbf{a},\mathbf{b}))}\right] \end{aligned} \tag{2.20}$$

it equals the product of the number of sites and the relative entropy $H(W|Q_0)$. The extensive quantity Σ, measures the statistical evidence for the mixture model based on the number and the score distribution of site pairs, whereas $H(W|Q_0)$ quantifies only the shape differences between the distributions $W(S)$ and $Q_0(S)$. This likelihood analysis goes beyond the inference of the sequence similarity $\mathcal{K}_\ell(r)$ introduced above in eq. (2.11). It can be seen as a decomposition of the distance-dependent similarity between sites into two parts: the similarity between their motifs, and the excess similarity of the actual site pairs beyond that of their motifs. The first part reflects functional correlations within regulatory modules and is assigned to the background model $Q_A(\mathbf{a})Q_B(\mathbf{b})$. Only the second part provides evidence for common descent,

which is gauged by the scoring function $S(\mathbf{a}, \mathbf{b})$.

2.3.1 Local sequence duplications in *Drosophila*

Using the duplication score S, we have evaluated the sequence similarity of 506 pairs of neighboring binding sites in regulatory modules of the *Drosophila melanogaster* genome. These sites are experimentally validated and recorded in the REDfly database (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008) (see Materials and Methods). We infer the prevalent mode of evolution as a function of the distance r between sites and obtain the main result of this paper:

—*In fly, binding sites with a distance of up to about* 50 *bp are more likely to share a common ancestor than to have evolved from independent origins.* Fig. 2.5(a) shows the histogram of duplication scores $S(\mathbf{a}, \mathbf{b})$ for the set of $k = 306$ binding site pairs with $r \leq 50$ bp. The score distribution $W(S)$ of these pairs is clearly distinguished from the background distribution $Q_0(S)$, which is obtained from pairs of sites located in the same module at a distance $r > 200$ bp and is associated with independent descent. We decompose the score distribution of adjacent sites in the form $W(S) = (1-\lambda)Q_0(S) + \lambda Q(S)$, attributing the excess of large scores to pairs of sites of common descent with a score distribution $Q(S)$. Our best fit of this mixed-descent model to the data distribution has a fraction $\lambda = 57\%$ of adjacent site pairs formed by duplication; see Fig. 2.5(a). The total log-likelihood of the mixed-descent model relative to the background model is given by multiplying the relative entropy of the distributions W and Q_0 with the number of site pairs, $\Sigma = k\, H(W|Q_0)$. We estimate $\Sigma > 234$, providing significant statistical evidence that the prevalent mode in adjacent sites is evolution from common descent. We note that this significance emerges for the ensemble of the adjacent site pairs, whereas the relative log-likelihood for duplication per site pair, $H(W|Q_0)$, is of order one: individual site sequences are inevitably too short to reliably discriminate between the two evolutionary modes. Our conclusion that local sequence duplications generate the observed excess similarity of adjacent sites is supported by a number of further controls and a comparison with the yeast intergenic regulatory sequences:

—*The relative log-likelihood for duplication per site pair decreases with increasing distance r between sites.* In Fig. 2.5(b), we evaluate the relative entropy $H(W_r|Q_0)$ for

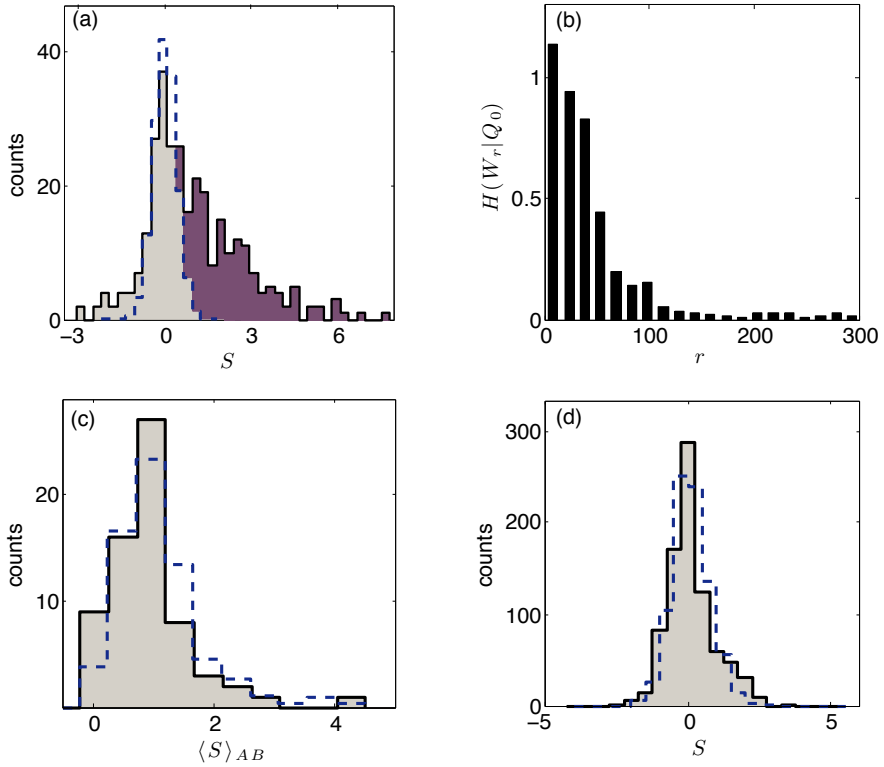

Figure 2.5: Common vs. independent descent of binding sites in fly and yeast. (a) Histogram of the duplication score S for 306 pairs of binding sites with a mutual distance of up to 50 bp in the genome of *D. melanogaster* (sum of grey-shaded and violet-shaded part). Decomposition of counts according to the mixed-descent model (see Materials and Methods): 43% of the site pairs are of independent descent and have the score distribution $Q_0(S)$ (obtained from pairs with relative distance $r > 200$ bp, dashed line), 57% of the site pairs of are of common descent and have the score distribution $Q(S)$ (violet-shaded). (b) Relative log-likelihood for duplication per site pair, i.e., relative entropy $H(W_r|Q_0)$ obtained from the score distribution $W_r(S)$ of site pairs in the relative distance range $(r, r+15)$ bp (evaluated from a total of 506 sites). The rapid decay of this function suggests a local mechanism generating excess similarity between adjacent sites. (c) Histogram of partial score averages $\langle S \rangle_{AB}$ for all factor pairs (A, B) binding the site pairs of (a) (grey-shaded) and corresponding distribution of averages obtained after scrambling the score values of site pairs (normalized to the same number of total counts, dashed line). The two distributions are statistically indistinguishable (KS-test p-value = 0.8378), which shows that positive duplication scores are not limited to a subset of factor pairs. (d) Histogram of the duplication score S for 833 pairs of binding sites with a mutual distance of up to 50 bp in the genome of *S. cerevisiae* (grey-shaded). The distribution is not significantly different from the null distribution obtained from random site pairs (normalized to the same number of total counts, dashed line), i.e., there is no evidence for common descent as prevalent evolutionary mode.

the score distributions $W_r(S)$ of site pairs with different values of mutual distance r. We find a rapid decay up to about 100 bp, that is, the score distribution W_r becomes successively more similar to the background distribution Q_0 with increasing site distance. This pronounced distance-dependance is comparable to that of the total sequence similarity shown in Fig. 2.1(c) and is consistent with local duplications as underlying mechanism.

—*Similarity of neighboring sites is broadly distributed over pairs of transcription factors.* We partition the 306 site pairs with a mutual distance of less than 50 bp by factor pairs and evaluate the partial score averages $\langle S \rangle_{AB}$. We compare the distribution of these averages with the corresponding distribution of averages evaluated after scrambling the score values of the site pairs, as shown in Fig. 2.5(c). The two distributions are statistically indistinguishable, which shows that excess sequence similarity is a broad feature of adjacent binding sites and is not limited to a subset of sites for factor pairs with specific functional relationships. This supports our conclusion that the excess sequence similarity reflects common descent and not fitness interactions (epistasis) between sites. Of course, epistasis is common for binding sites in the same regulatory module, because these sites perform a common regulatory function. However, generic interactions couple the binding energies of adjacent sites, not directly their sequences. Epistatic effects generating excess sequence similarity are conceivable for specific factor pairs, but do not appear to be a parsimonious explanation for the broad similarities of adjacent binding sites we observe.

—*In yeast, binding site duplications are not frequent.* For comparison, we have also evaluated a set of 1352 pairs of binding sites in the *Saccharomyces cerevisiae* genome. Fig. 2.5(d) shows distribution of duplication scores $S(\mathbf{a}, \mathbf{b})$ for the set of binding sites with $r \leq 50$ bp. This distribution is strongly peaked around zero (because the maximum-likelihood value of τ is large; $\tau > 1/\mu$) and indistinguishable from the distribution of the control set of random site pairs; both distributions have a negative average. As in *Drosophila*, most binding sites in the same intergenic region of *S. cerevisiae* are located within 50 bp from each other. However, we do not observe evidence for local duplications as a mode of binding site formation in yeast. Clearly, this result does not exclude that such duplications take place, but they do not appear to be

frequent enough to generate a statistically significant excess similarity of neighboring sites. This is not surprising given the differences in regulatory architecture between yeast and fly: individual sites in *S. cerevisiae* are more specific than in *Drosophila*; the average sequence information of a binding motif is $H \approx 12 - 17$ bits, compared to $H \approx 6 - 8$ bits; see Table. 1.1 and (Wunderlich and Mirny, 2009). Accordingly, a larger part of the regulatory functions in yeast relies on single sites, and there are no regulatory modules which would require frequent duplications for their formation.

2.3.2 Adaptive potential of duplications

Do the inferred site duplications have adaptive potential for the formation of novel binding sites? Here, we use the term adaptive potential to indicate that the duplication itself may be a neutral process, and selection for factor binding may latch on later to duplicated sites. The duplication of a site for a given transcription factor has obvious adaptive potential towards formation of an adjacent site for the same factor. But local duplications also have adaptive potential if the duplicated site is to evolve the new function of binding a different factor, because the binding motifs of transcription factors with adjacent sites are correlated. This correlation quantifies the ability of one factor to recognize the binding sites of another factor, including seed sites generated by sequence duplications. Specifically, we define the binding correlation H_{AB} of a transcription factor A with another factor B as the average information-based fitness to bind factor B in the ensemble of A-sites.

$$H_{AB} = \langle F_B \rangle_A = \sum_{i,a} q_{A,i}(a) f_{B,i}(a) \quad \text{with} \quad f_{B,i}(a) = \log \frac{q_{B,i}(a)}{p_0(a)}. \tag{2.21}$$

This value is an estimate for the compatibility of the A-sites with the transcription factor B and equals, up to a constant, the information-theoretic *cross entropy* between the distributions Q_A and Q_B.

In Fig. 2.6, this quantity is evaluated for all factor pairs (A, B) with adjacent binding sites and is compared to (i) the sequence information H_B of the motif Q_B, which equals the average fitness of B-sites for the B-factor by eq. (2.4),

$$H_B \equiv H(Q_B|P_0) = \sum_{i,a} q_{B,i}(a) f_{B,i}(a), \tag{2.22}$$

and (ii) to the average fitness of background sequence for the B-factor,

$$H_{0B} = \sum_{i,b} p_0(b) f_{B,i}(b). \tag{2.23}$$

For most such factor pairs, the fitness of a typical A-site is seen to be similar to that of weak B-sites and significantly larger than the average fitness of background sequence. This binding correlation between motifs is sufficient so that an A-site duplicate can act as a seed for a B-site, which can subsequently adapt its strength by point mutations. The binding correlation is specific to factors which have adjacent binding sites; we have found no such effect in the control ensemble of all factor pairs (A, B) (most of which do not have adjacent sites). Furthermore, some highly specific motifs, such as *hunchback*, *twi* and *z* do not show binding correlations with other factors.

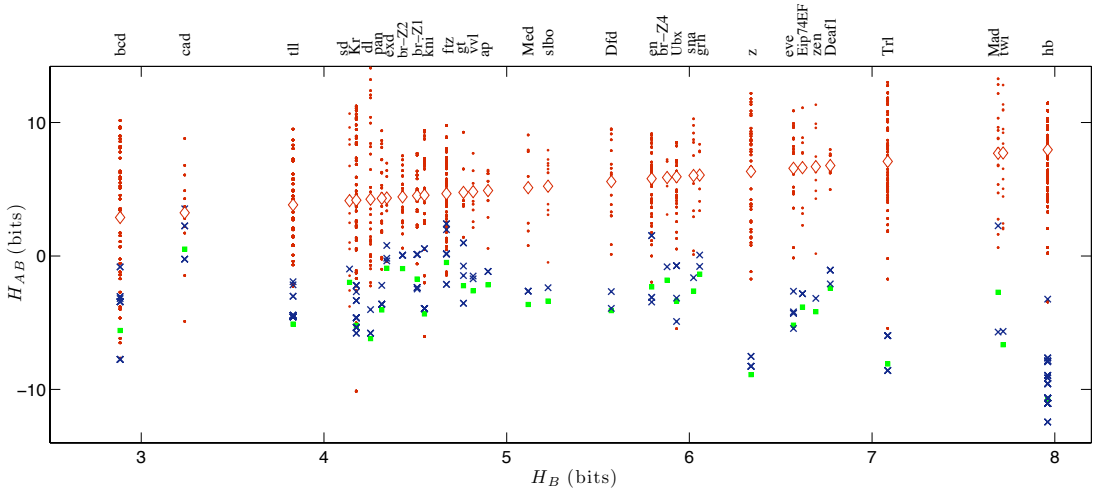

Figure 2.6: Adaptive potential of binding site duplications. The binding correlation H_{AB} of all pairs of *Drosophila* transcription factors (A, B) which have adjacent binding sites in a common regulatory module is evaluated as the average information-based fitness of A-sites for factor B and plotted against the sequence information H_B of the binding motif of factor B (blue crosses); see eqs. (2.21) and (2.22). The binding correlation is compared to the distribution of fitness values F_B of the B-sites (red dots, the average fitness for each factor is shown as diamond and equals the abscissa H_B) and to the average fitness F_B in background sequence (green dots); see eq. (2.23). The binding correlation H_{AB} is significantly larger than the background average of F_B and is comparable to the fitness F_B of weak B-sites in a substantial fraction of cases. Some highly specific motifs, such as *hunchback*, *twi* and *z* do not show binding correlations with other factors.

2.4 Discussion & Outlook

Local sequence duplication as a mechanism of regulatory evolution

Local sequence duplications (and deletions) are a generic evolutionary characteristic of intergenic DNA and, in particular, of regulatory sequence (Boeva et al., 2006; Gruen, 2006; Messer and Arndt, 2007; Sinha and Siggia, 2005; Tanay and Siggia, 2008; Vinces et al., 2009). In this chapter, we have established evidence for local sequence duplications as a mechanism that transports and produces cis-regulatory information. These duplications generate specific, distance-dependent sequence similarity in strongly correlated pairs of sites with a relative distance of up to about 50 bp, which account for a substantial part of the sequence autocorrelation in fly regulatory modules. In particular, they provide a parsimonious explanation for the excess sequence similarity of transcription factor binding sites, which is broadly observed in this range of relative distance. We conclude that the majority of these adjacent site pairs have evolved from a common ancestor sequence. The large amplitude of the duplication signal may be the most surprising result of this study. It far exceeds the level expected from the repeats in contemporary sequence, which cover only about 5 percent of binding sites and are typically shorter than the distance between correlated sites. Common-descent site pairs are the cumulative effect of past duplications over macro-evolutionary intervals, whose trace is conserved by selection on site functionality.

This result establishes local duplication as a pervasive formation mode of regulatory sequence, which generates, for example, the known local variations in site numbers between *Drosophila* species. Of course, our evidence for this mode is statistical and, at this point, is confined to a limited dataset of binding sites with confirmed functionality (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008). The duplication mode appears to be specific to multicellular eukaryotes; we have not found comparable evidence in the yeast genome. Our findings are relevant for genome analysis in two ways: including local duplications should inform inference methods for binding sites as well as alignments of regulatory sequence with improved scoring of indels (Gruen, 2006; Messer and Arndt, 2007; Sinha and Siggia, 2005; Tanay and Siggia, 2008). With such methods, it may become possible to follow the evolutionary history of binding site duplications across species.

Life cycle of a binding site

We have found evidence that local duplications can confer adaptive potential for the formation of novel binding sites, because they generate seed sequences with marginal binding specifically in the vicinity of existing sites. This mechanism is necessary, because point mutations alone can only lead to rapid loss but not to gain of new sites with positional specificity. Thus, duplications and point mutations complement each other, suggesting that typical binding sites within multicellular eukaryotes have an asymmetric life cycle: formation within a functional cluster by local duplication, adaptation of binding energy by point mutations, evolution of relative distance to neighboring sites by insertions and deletions in flanking sequence, conservation by stabilizing selection on binding energy, and loss by point mutations.

The life cycle of individual binding sites interacts with other levels of genome evolution. Gene duplications with subsequent sub-functionalization have been identified as an important evolutionary mode specifically in higher eukaryotes (Lynch and Conery, 2003). If subfunctionalization is initialized at the level of gene regulation, it amounts to a loss of regulatory input for both gene duplicates and provides a mechanism for adaptive loss of binding sites. This process alone would lead to genomes with many genes, but few functions per gene. Maintaining regulatory complexity with multi-functional genes as observed in eukaryotic genomes (Davidson, 2006; Ptashne and Gann, 2002) requires a converse evolutionary mode: gain of new functions by existing genes. At the regulatory level, this amounts to gain of regulatory input, i.e., adaptive formation of new binding sites.

Sequence evolution and regulatory grammar

Previous studies have identified regulatory modules as important units of transcriptional control, in which clusters of binding sites bind multiple transcription factors with cooperative interactions. The sites in a cluster follow a regulatory grammar resulting from natural selection acting on site order, strength, and relative distances (Kulkarni and Arnosti, 2005; Markstein et al., 2002; Small et al., 1993). If sequence duplications play a major role in the formation of such clusters, we may ask how much of their observed structure reflects this mode of sequence evolution, rather than optimization

of regulatory function by natural selection. Local duplications generically produce descendant sites, which are weak binding sites for another factor at best, as shown in Fig. 2.6. (Significant heterogeneity in binding strength between adjacent sites is indeed observed in our sample.) The resulting binding sequences are hardly optimal in terms of specificity and discrimination between different factors. Cooperative binding between transcription factors may have evolved as a secondary mechanism to confer regulatory function to these sequence structures. This is the topic discussed in Chapter. 3.

In this chapter, we have argued that local sequence duplications facilitate the adaptive evolution of gene regulatory interactions. However, the adaptive potential of duplications does not imply that the duplication process itself has to be adaptive or even confined to regulatory sites. Similar to gene duplications (Lynch and Conery, 2003), many site duplications may be neutral and provide a repertoire of marginal regulatory links. Adaptive diversification can build subsequently on this repertoire, conserving and tuning those links that confer a fitness advantage and discarding others.

2.5 Materials and Methods

The sequence analysis of *D. melanogaster* is based on the *cis*-regulatory modules and experimentally validated binding sites collected in the REDfly v.2.2 database (Bergman et al., 2005; Gallo et al., 2006; Halfon et al., 2008), and on the position weight matrices of Dan Pollard's dataset, (*http://www.danielpollard.com/matrices.html*).

To measure the distance-dependent sequence similarity $\mathcal{K}_\ell(r)$, we use the 346 known regulatory modules with length of more than 1000 bp in *D. melanogaster*. The analysis in *D. yakuba* and *D. pseudoobscura* is based on the 249 well-aligned homologous regions obtained from multiple alignments of 12 Drosophila species (dm3, BDGP release5); see Fig. 2.3. For the evolutionary inference in the second part of the paper, we use only the experimentally validated binding sites contained in these modules which are not necessarily selected for high similarity to motifs or for high mutual similarity. To avoid biases in our analysis, the set of sites is truncated in three ways: (i) We only use binding sites for transcription factors that occur in at least two different regulatory modules, so that the position weight matrix is not biased by the sequence context of a single module. (ii) We use only sites that have no sequence overlap with other sites in the dataset, because our inferred fitness landscapes describe the selection for

a single regulatory function (Mustonen et al., 2008). (iii) We exclude sites in the X chromosome, which could bias the results by its high rate of recent gene duplications and the abundance of repeat sequences (Katti et al., 2001; Thornton and Long, 2002). These conditions produce a cleaned set of 506 transcription factor binding site pairs located in 74 *cis*-regulatory modules.

For the analysis in *S. cerevisiae*, we use sites and position weight matrices from the SwissRegulon database (Pachkov et al., 2007). These footprints do not always match the length ℓ of their position weight matrices. To produce a set of site sequences of common length ℓ, longer footprints are cut and shorter ones joined with flanking nucleotides, such that the binding affinity is maximized.

3

Emergent selection on regulatory complexes

3.1 Introduction

In this chapter we will discuss another aspect of the regulatory evolution, that is the evolution of the binding complexes as single biological units. Gene expression in higher eukaryotes is regulated by combinations of transcription factors which in cooperation influence the rate of transcription. The combinatorial complexity ensures the specificity of regulation in the larger genomes of multicellular eukaryotes (Bintu et al., 2005; Buchler et al., 2003; Gertz and Cohen, 2009; Kuhlman et al., 2007; Levine and Tjian, 2003; Shea, 1985). At the same time, the resulting flexibility in the regulatory modules allows for a substantial sequence changes while maintaining its overall functional output. Among the well-characterized systems is the *eve*-stripe enhancer region in the fruitfly (*Drosophila*) which plays a significant role during the development of the organism. Studies on four Drosophila species by (Ludwig et al., 2000; Ludwig and Kreitman, 1995; Ludwig et al., 1998) show that the wild sequence divergence has minor effects on expression regulation of the *even-skipped* gene and binding site turnover has been mostly tolerated. This property is even present on larger evolutionary distances (Hare et al., 2008). This form of evolution can be understood as a compensatory gain and loss of binding sites through which the functional output is maintained. Expectantly, there are also numerous case-studies to show modified expression by variation in binding module structures. Nonetheless, the maintenance and robustness of the function is

rather a surprising outcome.

In Section. 1.3, we discussed the biophysics of protein-sequence interactions for single binding sites. We characterize a fitness landscape for evolution of the single site sequences based on their interactions with a transcription factor (eq. 1.17). The fitness of a binding site, which is associated with its ability to regulate gene expression in the cell, is a highly nonlinear function of the site affinity and yields epistatic interactions between nucleotide changes in the site sequence; see Fig. 1.3(b). Considering a regulatory region with multiple binding sites, the picture becomes even more complex: The tolerance for site turnover and functional exchange results in an epistatic evolutionary effect not only at the level of nucleotides in a single binding site but also between the binding sites in a regulatory region. The complexity is indeed two fold: (i) The biophysical interactions are more complex for a system with multiple interacting binding sites, and thus a collective phenotype has to be associated with such many-particle (i.e., multiple site) complexes. (ii) The evolutionary dynamics of these regulatory sequences are also more involved: these regions with a length of several hundred base pairs provide larger mutational targets compared to single binding sites with a length of about 10 base pairs. Hence, different individuals in a population carry polymorphic binding complexes whereas single binding sites have mostly monomorphic population compositions known for the weak-mutation regime. This is simply a manifestation of a quantitative trait with multiple contributing loci. The three underlying evolutionary forces, mutation, selection and genetic drift affect the trait on comparable timescales and the simplifying assumptions previously used for separation of these timescales are not anymore applicable; see Section. 1.4. In this way, the stationary population composition is set by the mutation-selection-drift balance. Furthermore, a limited amount of genomic reassortment by mechanisms such as recombination or horizontal gene transfer, cause mutations to be physically linked to each other on the chromosome. This condition known as linkage-disequilibrium, adds further complications in characterizing the observed variations in the population.

These are the issues that we encounter in this chapter. We study the evolution of a type of regulatory complexes in yeast which consists of a central high-affinity binding site surrounded by a cloud of weak site sequences; see Fig. 3.1 (b). The biological importance of such structures has been recently highlighted through a series of synthetic experiments in yeast (Gertz and Cohen, 2009; Gertz et al., 2008). Transcription factors

in yeast typically exhibit cooperative interactions with their adjacent factors on DNA and thus, form dimers and not longer oligomers. The resulting protein-DNA complexes with cooperatively bound transcription factors are more stable than the structures formed by single site interactions. In this picture, low affinity binding sites influence the regulatory output of a gene through cooperation with the strong site and hence confer a significant biological role. Besides the mechanistic importance of these structures, we have also identified the evolutionary role of the weak binding sites as seed sequences that facilitate the formation of the regulatory modules in eukaryotes (Nourmohammad and Lässig, 2011); see Chapter. 2.

Despite the functional significance of the low-affinity site sequences, inference of selection pressure during their evolution has proven to be a challenge. Individual weak sites do not experience a strong purifying selection pressure and thus have a high turnover rate. The resulted divergence across species has illuded to the understanding that these flanking regions are evolving near-neutrally and hence are not of any phenotypic significance. As we pointed out however, their effects should be studied in the context of their neighboring binding sites which together establish a collective binding phenotype for the regulatory complex .

Here, we develop a thermodynamic framework that quantifies a binding phenotype for a regulatory complex with multiple cooperatively interacting binding sites. This will serve as a molecular phenotype which is related to the rate of protein production in the cell and hence is the functional target for natural selection during evolution. We develop a population genetics test to infer the shape of the fitness landscape from the phenotypic polymorphism in a population. We apply this method to infer the selection pressure on the binding complexes associated with the transcription factor *Rap1* in the *S. paradoxus* population. Using the inferred evolutionary rates, we will then predict the long-term sequence divergence between different yeast species. The resulting fitness landscape explains the compensatory evolution and the binding site turnover observed in these regulatory complexes.

3.2 Binding phenotype of regulatory complexes

A regulatory complex contains several binding sites for multiple transcription factors. In chapter. 1, we discussed both experimental and statistical methods to infer the

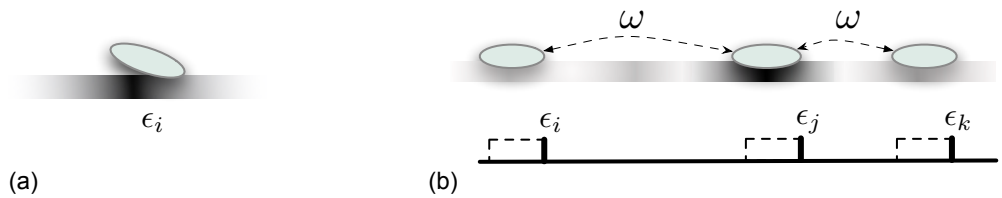

Figure 3.1: Transcription factor-binding site interactions. (a) Single binding site interaction with a transcription factor. ϵ_i is the sequence specific binding energy of this interaction. (b) Cooperative interaction of transcription factors in a regulatory complex. Upper row: The direct binding energy of the factors to DNA varies along the sequence depending on the nucleotide compositions of a sequence segment. The darkness shows binding affinity at each position ϵ_i. A strong binding site in the middle (dark) is surrounded by a cloud of low-affinity site sequences (light). The cooperative strength of the bound factors is ω. Lower row: The binding configuration can be specified by a binary model: $\sigma_i = 1$ for occupied positions and $\sigma_i = 0$, otherwise.
The dashed lines show the extension of the occupied position to exclude the overlapping bound factors.

binding affinity of a transcription factor to a specific sequence pattern which we note as the single site binding phenotype. Here, we want to extend such phenotypic picture to larger stretches of sequence regions that interact with multiple transcription factors. This requires a transition from a single-particle to many-particle statistics; see Fig. 3.1.

In the following analysis, we will characterize the binding phenotype of a sequence stretch in contact with one type of a cooperatively binding transcription factor. This level of complexity is sufficient to describe the regulatory system of the *Rap1* transcription factor that we study in this chapter. The cooperative binding between two transcription factors can be achieved through various mechanisms such as protein-protein interactions (direct or via DNA looping) (Arnosti et al., 1996; Bintu et al., 2005; Ptashne and Gann, 2002) and nucleosome-mediated interactions (Lam et al., 2008; Mirny, 2010). In our analysis, we do not distinguish between these different biological mechanisms and use a minimal model to characterize these cooperative interactions. The more general models which incorporate several types of transcription factors (i.e., different binding motifs) and even their interactions with chromatin structures are not much harder; see e.g., discussions by (Bintu et al., 2005; He et al., 2010; Raveh-Sadka and Levo, 2009; Shea, 1985).

Here, similar to Section. 1.3, we use an additive energy model based on the position weight matrix (PWM) of the transcription factor under study, to infer the binding

affinity of its interaction with a single binding site of length d; see eq. (1.4). This provides an estimate for binding energy of direct interactions between the transcription factor and site sequence segments at each position of the regulatory complex. Our minimal interaction model assigns only an effective cooperativity, ω, between the transcription factors bound next to each other without any interfering factors in between (see Fig. 3.1(b)). Assuming such short-range interactions is justified in prokaryotes and in yeast (Bintu et al., 2005; Gertz et al., 2008; Ptashne and Gann, 2002). However, in higher eukaryotes long-range interactions are essential parts of the regulatory grammar and should be taken into account (Arnosti and Kulkarni, 2005; Kulkarni and Arnost, 2005). These interaction parameters, together with the transcription factor density in the cell, characterize the collective binding phenotype of a regulatory complex.

To proceed, we will use the analogy between this system and well known spin interaction models in physics. We associate a state vector, $\sigma_i \quad (i = 1 \dots \ell)$ to a binding configuration of transcription factors on a sequence of length ℓ. The state vector takes values 0 and 1 to denote the occupancy at different genomic positions: $\sigma_i = 1$ if the site sequence $(a_{i-d+1}, \dots, a_i)$ is occupied and $\sigma_i = 0$, otherwise; Binding position "i" is by construct the location of the right-most tip of the bound transcription factor on the sequence. Therefore, the interaction affinity assigned to a genomic position "i" is the binding energy for direct interactions between the transcription factor and the connected sequence on its left, $\varepsilon_i = E(a_{i-d+1}, \dots, a_i)$; see Fig. 3.1(b). The cooperative interaction for a binding configuration σ can be described by a Hamiltonian,

$$H(\sigma) = \sum_i \sigma_i \varepsilon_i + \omega\theta \left(\sum_{j<i} \sigma_i \sigma_j - 1 \right) \tag{3.1}$$

$\theta(x)$ is a step function: $\theta(x) = 1$ for positive values of $x > 0$ and $\theta(x) = 0$, otherwise. The second term on the right hand side of eq. (3.1) accounts for the cooperative interaction of the immediate neighboring bound factors on the sequence. Later, we will add the constraint that the binding interactions are exclusive and factors do not overlap with each other.

Since binding interactions are stochastic, there are numerous possible binding configurations, σ for each regulatory complex. Expression regulation as the ultimate phenotype, is a response to all of these configurations. Therefore, a sound biophysical phenotype should take into account all of these binding configurations and weigh them

according to their regulatory contributions. This is a well-known procedure in statistical physics and the corresponding function is termed the partition sum, Z which depends on the interaction parameters of the system. The likelihood that a specific binding configuration σ occurs in equilibrium is given by the Boltzmann factor, $\exp(-\beta H(\sigma))$. where $H(\sigma)$ is the interaction Hamiltonian in eq. (3.1) and β is inversely related to the temperature times Boltzmann's constant, $\beta = 1/k_B T$. In our analysis, we will rescale the energy values by β. The partition sum for the binding interactions between the transcription factors and a regulatory complex of length ℓ is,

$$\begin{aligned} Z &= \sum_{\sigma_\ell=0,1} \cdots \sum_{\sigma_1=0,1} \mathrm{e}^{-H(\sigma)+\sum_i \left(\nu\sigma_i + V_\infty \sigma_i \sum_{j=i+1}^{i+d-1} \sigma_j\right)} \\ &= \sum_{\sigma_\ell=0,1} \cdots \sum_{\sigma_1=0,1} \mathrm{e}^{-\sum_i \left[(\varepsilon_i-\nu)\sigma_i + \omega\theta\left(\sum_{j<i} \sigma_i\sigma_j - 1\right) + V_\infty \sigma_i \sum_{j=i+1}^{i+d-1} \sigma_j\right]} \end{aligned} \tag{3.2}$$

ν is the chemical potential related to the finite density of transcription factors in the cell: $\nu \propto \log n_{TF}$; see eq. (1.2). The last term in the exponent assures the exclusive binding criteria: $V_\infty \gg 1$ is a large potential barrier that dismisses the configurations with overlapping bound factors.

Enumerating all possible configurations that grow exponentially with the sequence length $\approx 2^\ell$ is a strenuous task. Since the interactions assumed in this model are local (i.e., only adjacent transcription factors cooperate), we can compute the partition sum in eq. (3.2) in a recursive form. This technique is known as transfer matrix method in physics and dynamical programming in computer science. The recursive form of the partition function is,

$$Z_r = Z_{r-1} + \mathrm{e}^{-(\varepsilon_r-\nu)}[Z_0 + \mathrm{e}^{-\omega}(Z_{r-d} - Z_0)] \tag{3.3}$$

with the initial condition, $Z_0 = 1$. The first term on the right hand side of eq. (3.3) is the contribution of the unbound state at position r ($\sigma_r = 0$), to the partition sum. The second term is the contribution of the bound state ($\sigma_r = 1$) with two possibilities: (i) there is no other factor bound on the sequence prior to position r and hence no cooperative interaction, ($\mathrm{e}^{-(\varepsilon_r-\nu)} Z_0$) and (ii) there is at least one other factor bound prior to the position r, ($\mathrm{e}^{-(\varepsilon_r-\nu+\omega)}[Z_{r-d} - Z_0]$). The recursive form in eq. (3.3) proves to be very useful in computing the large partition sum in eq. (3.2).

From the partition function, we can then compute different macroscopic observables that characterize the molecular phenotype of a regulatory complex. One of the extensive thermodynamic parameters is the free-energy of the system, $G = \log Z$ which accounts for all the contributing configurations with an appropriate weight. The traditional definition of Helmholtz free energy in statistical mechanics is $A = -k_B T \log Z$. Here, we define G in units of $\beta = 1/k_B T$ and adopt a positive sign to relate it to the effective binding affinity of a regulatory complex: higher values G are associated with larger effective bindings. In contrast to the binding energy, which is approximately additive in its single nucleotide contributions, the free-energy of a regulatory complex G, is a highly nonlinear function of the individual site energies.

Another thermodynamics observable that we will use for our analysis is the marginal occupancy of each nucleotide position,

$$\mathcal{O}_i = \frac{Z(\ell|\sigma_i = 1)}{Z_\ell} \tag{3.4}$$

$Z(\ell|\sigma_i = 1)$ is the conditional partition sum over all configurations in which the sequence position "i", $(a_{i-d+1}, \ldots, a_i)$ is bound. This measure confers positional information for all nucleotides and account for their contribution to the effective binding of a regulatory complex. These two observables, free-energy and marginal occupancy, will be the central phenotypic observables in our evolutionary analysis. Using the partition function in eq. (3.2), we can easily estimate other statistical observables such as the average number of bound factors or the binding fluctuations in a regulatory complex, which provide some more intuition about the binding statistics of the region.

3.3 Evolutionary dynamics of regulatory complexes

Biological traits as combinations of multiple loci, known as "quantitative traits", have been studied in the context of classical quantitative genetics. We will treat the free-energy of the regulatory complexes as a quantitative trait. In this case, a sequence of 100 base pairs encodes the phenotypic characteristics of this trait, and mutations at any of these positions can influence the trait value. The number of contributing loci makes the regulatory complex a large mutational target, such that its binding phenotype remains polymorphic between the individuals of a population. This is the main difference compared to the single binding site with a length of about 10 bp,

which is practically monomorphic in the population. The relevant population based picture for regulatory complexes is a cloud of individuals with a broad spectrum of trait values whereas the picture for a single binding site resembles that of a point-like population; see Fig. 3.2(a). Inevitably, the evolutionary dynamics of these sequence regions becomes more complex. Mutations, selection and genetic drift all influence the trait on comparable time scales and hence distinguishing their contributions to the evolutionary dynamics is cumbersome.

One of the main difficulties in addressing the evolution of a binding complex, is the genomic linkage that physically connects the fate of all its constituent loci. This condition, known as linkage disequilibrium, adds a substantial amount of complications to the theoretical descriptions of this system. There is a lot going on in the population at the same time: the fate of a single mutation is determined not only by its own selective advantage but by the fitness contribution of all the other genomic loci which coexist in its background. This makes the natural selection to be less effective and the population to appear more neutral (Desai and Fisher, 2007; Gerrish and Lenski, 1998; Park and Krug, 2007; Rouzine et al., 2008; Schiffels et al., 2011). Single allele statistics given e.g., by Kimura's U-shape distribution, (eq. (1.11) and Fig. 1.2) cannot anymore characterize the genomic composition of a population (Kimura, 1962). Higher orders of allele correlations should be taken into account. Classical quantitative genetics however, mostly considers the traits as perfect reassortments of the genomic information for which single allele statistics are applicable (Barton and Coe, 2009; de Vladar and Barton, 2011b; Falconer, 1989; Kirkpatrick et al., 2002; Lande, 1976; Lynch and Walsh, 1998). Such assumption is justified for traits with loci that are encoded far away from each other on the genome or on different chromosomes, but certainly not for the binding complexes that we study here.

In Chapter. 4, we developed a novel theoretical framework to characterize the evolutionary dynamics of the polygenic traits. By drawing analogies between thermodynamics and quantitative genetics, we characterize the phenotypic composition of a population despite its complicated dynamics at the level of individual loci. This picture is valid for traits with a large number of contributing loci (Barton and Coe, 2009; de Vladar and Barton, 2011a; Kirkpatrick et al., 2002; Neher and Shraiman, 2011). In this section, we briefly discuss this mathematical framework to the extent that it is applicable to the analysis of the regulatory complex evolution. This analysis will yield

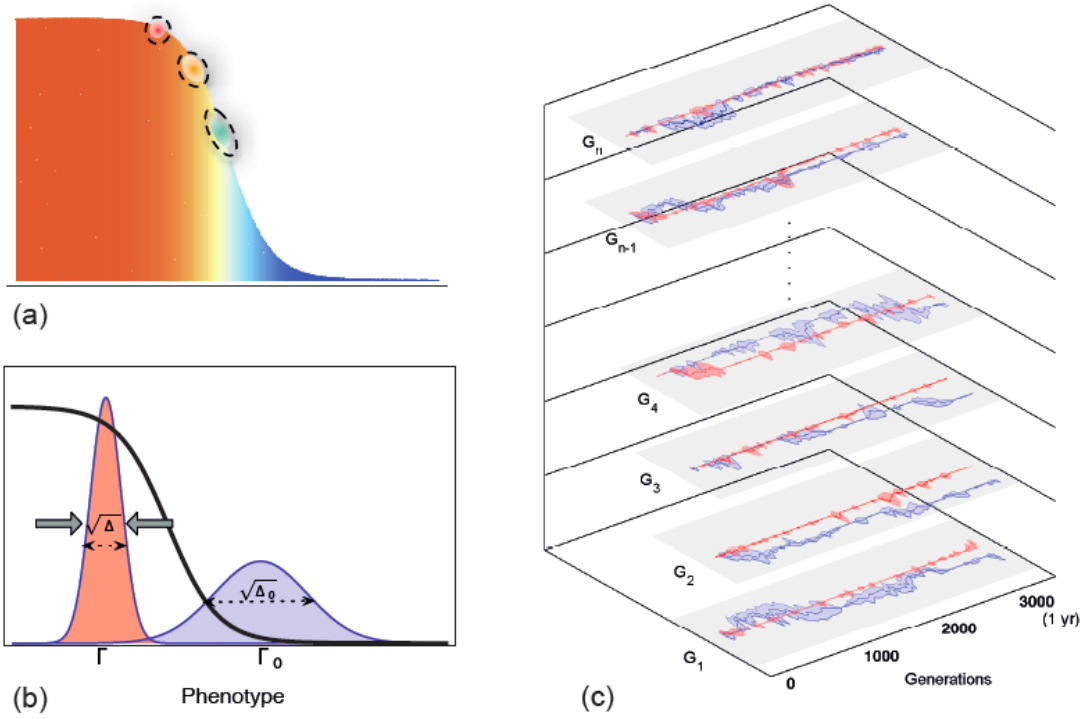

Figure 3.2: Evolution of quantitative traits under stabilizing selection. (a) Quantitative traits with multiple loci are polymorphic in the population.Their constitutive loci provide a large mutational target which in turn is reflected in phenotypic composition of the population. Therefore, populations should be viewed as clouds of individuals (enclosed by dashed lines in the figure) which span a range of phenotypes and evolve in a fitness landscape (here, a non-linear step-like landscape). The color indicates the probability of functionality , ranging from 1 (red) to 0 (blue). The plot shows fitness (y-axis) as a function of the population phenotype (e.g., a regulatory phenotype) on the $x-$axis. (b) The phenotypic diversity of a population that evolves under stabilizing selection pressure (in red) is reduced compared to its neutral counterpart (in blue). (c) Population realizations for simulated evolution of a regulatory complex binding phenotype under a minimal stabilizing selection (i.e., quadratic fitness landscape $Nf(G) = -0.38(G - G^{\star})^2$) (shown in red) and under neutral conditions (shown in blue). Each realization is depicted on separate z-plane. The evolutionary trajectories span over 3000 generations of yeast populations (approximately 1 year). The solid line in each of the trajectories show the population mean phenotype and the shadowed region around represents the phenotypic spread in the population. The phenotypic spread is reduced under stabilizing selection. The optimal phenotype for each realization $G^{\star}$ is set separately according to the equilibrated value of the original binding complex in yeast. In our analysis, we view different regulatory complexes as independent population realizations and use the phenotypic statistics across these loci to infer the shape of the minimal epistatic fitness landscape.

a population genetics test to infer the strength selection from the phenotypic polymorphism in a population. More details are discussed in Chapter. 4 and summarized in Section. 4.4.

Quantitative traits, here the binding free-energy, are combinations of a large number of loci and thus vary even between the individuals in a population. If the number of constituent loci is very large, the phenotype distribution in a population approaches a Gaussian form (central limit theorem) and can be characterized only by its first two moments. Here, we follow the dynamics of the intra-population mean and variance of the binding phenotype, respectively denoted by Γ_G and Δ_G. This is an approximation to the full phenotype distribution which is not a perfect Gaussian in the parameter regime of $\mu N\ell \sim 1$ relevant to the regulatory complexes. Still, our following results will show the power of this formalism in describing the genomic composition of a population.

A population can be thought as a single realization of an evolutionary process. Current high-throughput experimental techniques, mostly applicable to microbial populations, made it possible to follow not only one but many population realizations which evolve in similar fitness landscapes (Lenski and Travisano, 1994). Of course, the details of these populations differ from one to another, but the distributions of their macroscopic observables may inform us about the common evolutionary constraints that they have experienced.

This is the central idea behind our approach. We will view the different regulatory complexes as parallel evolutionary realizations that are to fulfill certain functional criteria; see simulated results in Fig. 3.2(c) for demonstration. In this approach, we naturally disregard the possibility that the evolutionary histories of these regulatory complexes might have met in the past e.g, through duplication events. This is a reasonable approximation if the number of regulatory complexes considered in the population ensemble is large. We also assume that all these loci have evolved in similar fitness landscapes. In the following section, the advantages and disadvantages of this analogy will be discussed. Nonetheless, this will offer a unique opportunity to study the phenotypic statistics in a genomic context.

The neutral dynamics, i.e., mutations and genetic drift, affect the mean and the variance of the intra-population phenotype distribution. Our goal is to characterize the stationary state of the population phenotype statistics, Γ_G and Δ_G, so that we can quantify their deviation due to selection. The subscript stands for the phenotype of

interest e.g., G for the binding free-energy. If the loci were independent, as in a freely recombining genome, both phenotype statistics would be additive in the comprising loci and thus have Gaussian distributions across populations due to the central limit theorem. In the presence of linkage, evolutionary dynamics of loci are correlated and hence follow more complex statistics. In Chapter. 4 we characterize the dynamics of Γ_ϕ and Δ_ϕ for an additive phenotype $\phi = \sum_i^\ell \phi_i$. We consider a binary genome where each locus has two states. This is a justified approximation in a low-mutation rate regime $\mu N \ll 1$ which is valid for most of biological systems; the polymorphic sites in the genome do not carry more than two types of nucleotides in the population. In this way, we can reduce the 4-state nucleotide composition at each locus $a_i = A, C, G, T$ to a 2-state system, $\phi_i = 0, 1$ where the allele "1" contributes to the phenotype $\phi = \sum_i^\ell \phi_i$. The intra-population phenotype statistics can be expressed in terms of the marginal allele frequencies, $y_i = \overline{\phi_i}$, and haplotype frequencies of pairs of loci, $y_{ij} = \overline{\phi_i \phi_j}$. The overlines denote the intra-population averages. The phenotype statistics follow,

$$\Gamma_\phi = \overline{\phi} = \sum y_i \tag{3.5}$$

$$\Delta_\phi = \mathbf{var}\phi = \overline{\phi^2} - \overline{\phi}^2 = \sum_i y_i(1-y_i) + \sum_{i\neq j}(y_{ij} - y_i y_j) \tag{3.6}$$

We can characterize the stochastic dynamics of these trait statistics under neutrality. Due to the stochastic nature of this process, we characterize the population state in neutrality by the joint probability density, $P_0(\Gamma_G, \Delta_G, t)$. The temporal changes of the cross-population marginal probability distributions, $P_0(\Gamma_\phi; \Delta_\phi = \langle\Delta_\phi\rangle, t) = P_0(\Gamma_\phi; \langle\Delta\rangle)$ and $P_0(\Delta_\phi; \langle\Gamma\rangle)$ follow,

$$\partial_t P_0(\Gamma; \langle\Delta\rangle, t) = \frac{1}{2N}\langle\Delta\rangle \frac{\partial^2}{\partial\Gamma^2} P_0(\Gamma; \langle\Delta\rangle, t) + 2\mu \frac{\partial}{\partial\Gamma}(\Gamma - \ell/2) P_0(\Gamma; \langle\Delta\rangle, t) \tag{3.7}$$

$$\partial_t P_0(\Delta; \langle\Gamma\rangle, t) = \frac{1}{N}\frac{\partial^2}{\partial\Delta^2}\Delta^2 P_0(\Delta; \langle\Gamma\rangle, t) + 4\mu \frac{\partial}{\partial\Delta^2}(\Delta - \ell/4 + \Delta/4\mu N) P_0(\Delta; \langle\Gamma\rangle, t) \tag{3.8}$$

The difference between this dynamics to that of the free-recombining genome, is the large fluctuations in the trait diversity, Δ, which is caused by the long-range correlations

between loci. The equilibrium state of eq. (3.7) and eq. (3.8) for phenotype statistics in a binary genome follows,

$$\begin{aligned} P_0(\Gamma_\phi;\langle\Delta\rangle) &= \frac{1}{Z_\Gamma}\ \exp\left[\frac{-2(\Gamma-\ell/2)^2}{\ell}\right] && (3.9)\\ P_0(\Delta_\phi;\langle\Gamma\rangle) &= \frac{1}{Z_\Delta}\Delta_\phi^{-3-4\mu N}\ \exp\left[\frac{-\mu N\ell}{\Delta_\phi}\right] && (3.10)\end{aligned}$$

where Z_Γ and Z_Δ are the appropriate normalization factors; see eq. (4.80) eq. (4.82). This way, we characterize the neutral dynamics of the population phenotype distribution. Given the neutral equilibrium, the distribution of the phenotype statistics for an evolutionary process in a time-independent fitness landscape $f(\phi)$ of a gradient form, will have a simple relation to its neutral counterpart,

$$\begin{aligned} Q(\Gamma_\phi;\langle\Delta_\phi\rangle) &= \frac{1}{Z}\ P_0(\Gamma_\phi;\langle\Delta\rangle)\ \exp[2NF(\Gamma_\phi;\Delta_\phi=\langle\Delta_\phi\rangle)]\\ Q(\Delta_\phi;\langle\Gamma\rangle) &= \frac{1}{Z}\ P_0(\Delta_\phi;\langle\Gamma_\phi\rangle)\ \exp[2NF(\Delta_\phi;\Gamma_\phi=\langle\Gamma_\phi\rangle)] && (3.11)\end{aligned}$$

The Boltzmann factor that relates each of the two distributions is the rescaled mean population fitness $2NF(\Gamma_\phi,\Delta_\phi)$ projected on the corresponding plane in the $\Gamma-\Delta$ coordinate. In a population with a trait distribution $\rho(\phi)$,

$$F(\Gamma_\phi,\Delta_\phi)=\overline{f(\phi)}=\int d\phi\rho(\phi)f(\phi) \qquad (3.12)$$

and the corresponding projections are obtained, $F(\Gamma_\phi;\Delta_\phi)=\int d\Delta_\phi F(\Gamma_\phi,\Delta_\phi)Q(\Gamma_\phi,\Delta_\phi)\approx F(\Gamma_\phi;\Delta=\langle\Delta_\phi\rangle)$, and similarly for $F(\Delta_\phi;\langle\Gamma_\phi\rangle)$. In a well-behaved fitness landscape $f(\phi)$, we can expand the fitness values around an arbitrary trait of interest $\phi^\star$,

$$f(\phi)-f(\phi^\star)=f'(\phi^\star)\ (\phi-\phi^\star)+\frac{1}{2}f''(\phi^\star)\ (\phi-\phi^\star)^2+\ldots \qquad (3.13)$$

In this way, we can compute the mean population fitness $F(\Gamma_\phi,\Delta_\phi)$ as a function of the intra-population trait statistics Γ, Δ and in relation to the derivatives of the fitness landscape. This will prove to be useful to infer the shape of the fitness landscape. The derivatives of the fitness function are coupled to the intra-population trait statistics, Γ

and Δ and hence, we can infer its shape from changes in the phenotype statistics under selection.

In Chapter. 4, we discuss in length the changes in phenotype statics as a response to different types of analytical gradient-form fitness functions. The results are also summarized in Section. 4.4. Here, our goal here is to characterize the evolutionary dynamics of regulatory complexes with ongoing compensatory binding site turnover. Thus, we only confine ourselves to a minimal fitness landscape that can accommodate epistatic evolutionary interactions i.e., a quadratic fitness landscape with a negative curvature, $f(\phi) = -\omega(\phi - \phi^\star)^2$. $\phi^\star$ is the location of the fitness peak. We will see that this fitness function can already explain the divergence patterns of regulatory complexes in *Saccharomyces*. The mean population fitness in a quadratic landscape has the form,

$$F(\Gamma_\phi; \langle\Delta\rangle) = -\omega\ (\Gamma - \phi^\star)^2 \quad , \quad F(\Delta_\phi; \langle\Gamma\rangle) = -\omega\ \Delta \tag{3.14}$$

We can now determine the modified marginal distributions for the trait statistics under selection, $Q(\Gamma_\phi; \langle\Delta\rangle)$ and $Q(\Delta_\phi; \langle\Gamma\rangle)$,

$$Q(\Gamma_\phi; \langle\Delta\rangle) = \frac{1}{Z}\ P_0(\Gamma_\phi; \langle\Delta\rangle) \exp[-2N\omega(\Gamma_\phi - \phi^\star)^2] \tag{3.15}$$

$$Q(\Delta_\phi; \langle\Gamma\rangle) = \frac{1}{Z}\ P_0(\Delta_\phi; \langle\Gamma\rangle) \exp[-2N\omega\Delta_\phi] \tag{3.16}$$

Binding free-energy G, is a non-linear function of the binding contributions from the constituent site sequences in a regulatory complex (see section. 3.2). Nonetheless, we apply the theoretical techniques that we have developed for linear traits to describe the evolutionary dynamics of the binding phenotype. This approximate scheme is applicable because of the sparse regulatory interactions in the binding complexes of interest where transcription factors are not likely to compete for overlapping binding sites. We will see the strength of this method in the following sections.

Trait statistics in a quadratic fitness landscape. Evolution in a a quadratic fitness lasndscape with a negative curvature reduces the phenotypic diversity and hence, sharpens the Γ_G distribution, $P_0(\Gamma_G; \langle\Delta\rangle)$ and shortens the long-tail of the Δ_G dis-

tribution, $P_0(\Delta_G; \langle\Gamma\rangle)$); see Fig. 4.7. We can simply evaluate the effect of quadratic selection, $f(G) = -\omega(G - G^\star)^2$ on the cross-population (here, cross-loci) trait statistics; see Section. 4.2.5.2 for more details.

- Trait average ($\Gamma_G^{(0)} \to \Gamma_G^{(s)}$)

$$\langle\Gamma_G\rangle_s = G^\star - \frac{G^\star - \langle\Gamma_G\rangle_0}{1 + 4\omega N(\langle\Gamma_G^2\rangle_0 - \langle\Gamma_G\rangle_0^2)} \tag{3.17}$$

$$(\mathbf{var}\Gamma)_s = \langle\Gamma_G^2\rangle_s - \langle\Gamma_G\rangle_s^2 \quad = \quad \frac{\langle\Gamma_G^2\rangle_0 - \langle\Gamma_G\rangle_0^2}{1 + 4\omega N\left(\langle\Gamma_G^2\rangle_0 - \langle\Gamma_G\rangle_0^2\right)} \tag{3.18}$$

- Trait variance ($\Delta_G^{(0)} \to \Delta_G^{(s)}$)

$$\langle\Delta_G\rangle_s = \langle\Delta_G\rangle_0 - 2\omega N\langle\Delta_G\rangle_0^2 + \mathcal{O}[(\omega N)^2] \tag{3.19}$$

$\langle.\rangle_s$ refers to averages over the population ensemble that evolved under selection pressure and $\langle.\rangle_0$ denotes the averages over the neutral ensemble.

Using the results in eq. (3.17) and eq. (3.19), we can infer the parameters of the fitness function, ω and $G^\star$.

$$2\omega N = \frac{\langle\Delta_G\rangle_0 - \langle\Delta_G\rangle_s}{\langle\Delta_G\rangle_0^2} \tag{3.20}$$

and

$$\frac{G^\star - \langle\Gamma_G\rangle_s}{G^\star - \langle\Gamma_G\rangle_0} = 1 + 4\omega N\left(\langle\Gamma_G^2\rangle_0 - \langle\Gamma_G\rangle_0^2\right) \tag{3.21}$$

Stabilizing selection reduces the spread of the phenotype in the population and as a result narrows the distribution of the average phenotype across populations,

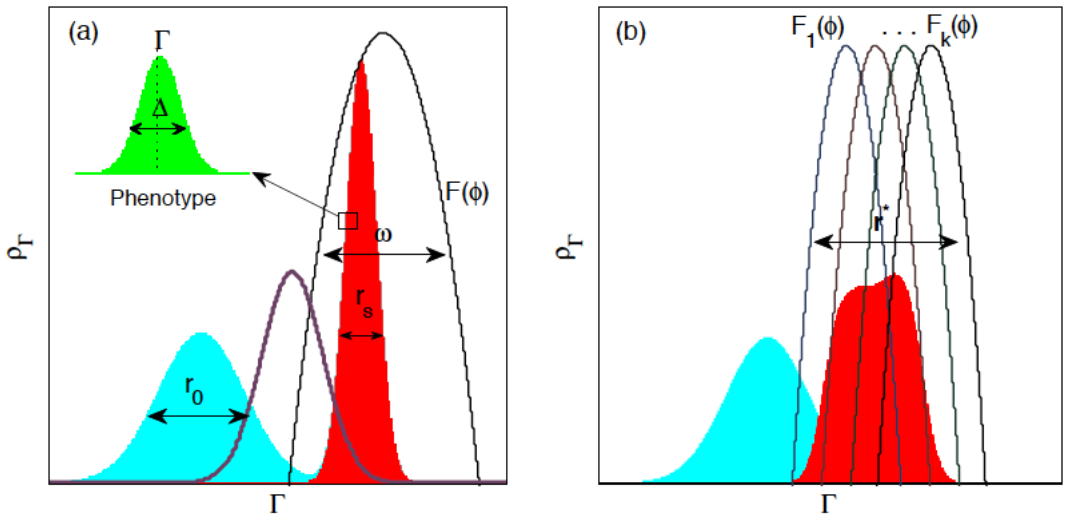

Figure 3.3: Trait statistics in quadratic fitness landscapes. (a) Distribution of intra-population trait average Γ in neutrality (light blue) and under stabilizing selection (red). Each value Γ is computed from the phenotype distribution of a single population (shown as an insert). The fitness landscape has a quadratic form with a width ω (solid black line). The phenotype range under selection r_s is smaller than in the neutral case, r_0. The stationary state of a finite population is determined by both selection and stochastic fluctuations. In very small populations selection is less efficient and entropic forces make the population look more neutral; a Γ distribution for a smaller population size is shown in purple solid line. Free fitness $\tilde{F} = F - \mathcal{H}$ provides an expression for the balance between natural selection (mean population fitness F) and stochastic drift (population entropy $\mathcal{H}$). (b) Γ distribution (in red) across loci which evolve in quadratic fitness landscapes with different optimums but similar width ω (solid dark lines). The resulting probability distribution $Q(\Gamma)$, as the sum of several different distributions, resembles the neutral distribution (in blue) and does not indicate an evolutionary dynamics under stabilizing selection. In this case Δ-statistics are useful.

$(\mathbf{var}\Gamma)_s < (\mathbf{var}\Gamma)_0$; eq. (3.18). The ratio between the phenotype range under selection $r_s^2 \equiv (\mathbf{var}\Gamma)_s$ and in neutrality $r_0^2 \equiv (\mathbf{var}\Gamma)_0$ is then a dimension-less quantity that indicates the strength and efficacy of stabilizing selection on the trait; see Fig. 3.3(a).

$$r_s^2 = \frac{1}{4\omega N} - \frac{1}{(4\omega N r_0)^2} + \mathcal{O}(1/(\omega N r_0)^3) \tag{3.22}$$

Since Γ has a Gaussian distribution (eq. (3.9) and eq. (3.15)), the phenotypic range as defined above is related to the entropy, $\mathcal{H}_\Gamma = -\langle \log P(\Gamma) \rangle$ and hence the information content and redundancy of the distribution.

Free fitness: a measure of selection strength on the trait. In finite populations, it is not only the maximization of the mean population fitness $F(\Gamma_G, \Delta_G)$ that determines the stationary phenotype composition, but rather it is the free fitness that is maximized at the stationary state (Barton and Coe, 2009; Berg et al., 2004; Iwasa, 1988; Mustonen and Lässig, 2010; Sella and Hirsch, 2005). A free fitness function provides an expression for the balance between natural selection, mutations and stochastic drift, which is related to the entropy of the population phenotype distribution $\mathcal{H}$. In analogy to statistical physics, free fitness in finite populations plays the role of free energy in finite temperature, and is maximal at equilibrium; this is the evolutionary analog of Boltzmann's H theorem. It is the free fitness $\tilde{F} = NF - \mathcal{H}$ that gauges the stability of a population state by balancing between the evolutionary tendencies in finite populations to increase both fitness and entropy. Therefore, the excess of the free fitness for the evolutionary state under selection over the neutral counterpart, $\delta\tilde{F} = \tilde{F}_s - \tilde{F}_0$ measures the strength of selection that constrains the phenotype composition in the population; see Fig. 3.3. We compute this quantity from the phenotype distributions across loci and compare it to the neutral expectation evaluated in the simulated evolution of regulatory complexes.

Genomic analysis across different loci. As mentioned above, we will study the evolutionary dynamics of *Rap1* regulatory complexes in yeast. These complexes are responsible for regulation of different genes, but through interactions with common transcription factor. The underlying biochemistry of these interactions is their shared characteristic. However each is tuned to regulate the output expression level of its own gene. In other words, the fitness optimum for each of these trait loci is located at the value that best matches its own regulatory output. If fitness functions for all loci are equal, they can be thought of as different realizations of the same evolutionary process, and cross-loci statistics will match the cross-population statistics of the trait observables as discussed above. If only the fitness optimum $G^\star$ differs between the loci, then the cross-loci Γ-statistics differ from that of the cross-population description, but the higher order statistics (Δ-statistics for our analysis) will remain compatible. These relations can be generalized to higher moments.

The spread of the cross-loci fitness optima, $r^\star = (\mathbf{var}G^\star)^{1/2}$ is then a measure for deviation of the cross-loci Γ-statistics from the cross-population descriptions; see Fig. 3.3(b). If $r^\star \sim r_0$, the Γ-statistics are diluted and we cannot infer the fitness parameter ω from the reduction in the spread of the intra-population trait average Γ. On the other hand, if loci-specific fitness functions have similar width ω, the reduction in trait diversity Δ is comparable across loci and is related to the fitness width by eq. (3.19).

These results are applicable to actual genomic data. The biophysical map between genotype and binding phenotype is a peculiar feature of regulatory complexes. Mutations are random events which introduce an unbiased change to the nucleotide content of the trait loci. Selection however, acts on the binding phenotype which is related to gene expression and protein production. Clearly, the genotype-phenotype map is an essential element for pursuing evolutionary analysis of this type. Comparing cross-population phenotype statistics of the actual genomic data to its neutral counterpart then provides information about the shape of the underlying fitness landscape. We will see that a minimal epistatic fitness landscape (quadratic fitness) presented here, can already explain the conservation of the binding phenotype as well as ubiquitous compensatory changes between sites in a regulatory complex.

3.4 *Rap1* binding complexes in *S. paradoxus*

We have evaluated the free-energy of 411 promoter regions in 37 individuals of the *S. paradoxus* population. We treat the different promoters as different realizations of a regulatory complex that evolved to interact with the *Rap1* transcription factor. In this way, we distinguish between the intra-population statistics evaluated from homologous regions in 37 individuals and cross-loci statistics evaluated from the 411 regulatory complexes. As explained above, the cross-loci statistics are not necessarily compatible with the cross-population formalism that we introduced in Section. 3.3. We will see the limitations of these analogies in our analysis. We compare these statistics with the phenotypic statistics evaluated from the population realizations that evolved neutrally. The neutral expectation is measured from simulated evolution of the regulatory complexes for which the central strong binding site is maintained, but the flanking

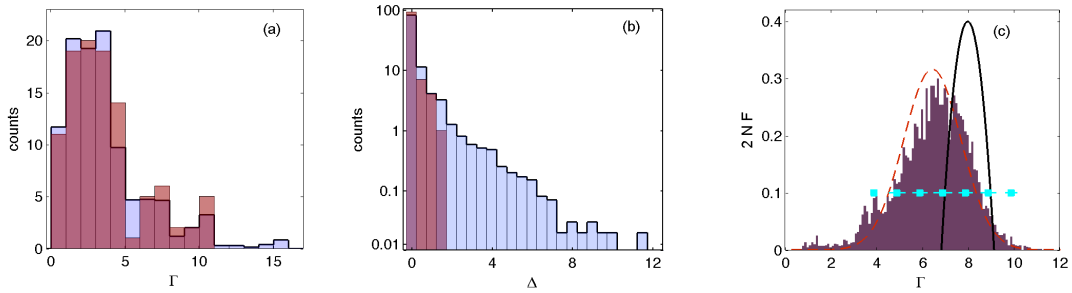

Figure 3.4: Inference of selection from phenotypic polymorphism. (a) Cross-loci distribution of the intra-population phenotype (free-energy) average Γ_G over 411 binding complexes for *Rap1* transcription factor in S. paradoxus (in red). The blue distribution in the back is the neutral expectation of the intra-population phenotype average. The two distributions are not significantly different. This is due to the different preference of the binding phenotype across regulatory complexes which results in an inhomogenous ensemble of loci. (b) Cross-loci distribution of the intra-population phenotype variance Δ over 411 binding complexes. The actual distribution is less spread out and has a smaller mean value (by a factor of two) compared to its neutral counterpart. This suggests that the underlying fitness landscape allows for epistatic interactions between the trait loci. The minimal fitness landscape of such type has a quadratic form, $f(G) = -\omega\,(G-G^\star)^2$ with the curvature parameter $2\omega N = 0.38$; see eq. (3.20). The location of the fitness peak $G^\star$ is set independently for individual loci. (c) Simulated evolution in a quadratic fitness landscape, $2Nf(G) = -0.38\,(G-G^\star)^2$ with a peak set at the value $G^\star = 8$ (black curve). The cross-loci histogram of intra-population phenotype average (average free-energy) Γ_G is shown in purple. The intra-population width of the phenotype distributions Δ_G is indicated for a number of cases (cyan lines). The phenotypic equilibrium does not coincide with the fitness landscape due to the mutational load effect.

low-affinity binding regions have evolved neutrally; see *Materials and Methods* in Section. 3.7. The following features can be extracted:

—*The cross-loci distribution of the mean phenotype* Γ_G *is not significantly different from the neutral expectation.* We evaluate the intra-population average binding free-energy (phenotype) for each of the 411 regulatory complexes in the *S. paradoxus* genome. We compare the cross-loci distribution of Γ_G to that of the neutrally evolved populations which still carry the strong binding sites, but have not experienced constraints on their low-affinity sequence regions. The two distributions are not significantly different; see Fig. 3.4(a). The intra-population mean phenotype in *S. paradoxus* is as widely spread as its neutral counterpart which suggests that the phenotypic preferences $G^\star$ vary across loci. Since the intra-population phenotype diversity is small, Γ can be a good approximation for phentype preference $G^\star$ at each loci. In this way, we can argue that the

spread of the $G^\star$ distribution is comparable to the width of the Γ distribution in neutrality, $r^\star \sim r_0$. Therefore, Γ-statistics cannot inform us about the shape of the fitness landscape for these regulatory complexes.

—*The intra-population phenotype diversity Δ_G is significantly reduced compared to the neutral expectation.* As shown in Fig. 3.4(b), the distribution of Δ_G is less spread than the neutral expectation. The cross-loci average of the population phenotype diversity is $\langle\Delta_G\rangle_s = 0.27$, which is significantly smaller that its neutral counterpart, $\langle\Delta_G\rangle_0 = 0.5$. This is an indication of evolution in a stabilizing fitness landscape. Phenotype diversity statistics is not affected by the preferred binding value $G^\star$ which differs between regulatory complexes. In the following, we proceed with a minimal model that assumes a similar width ω for fitness landscapes of all loci and use Δ-statistics to infer this fitness parameter; see Section. 4.3. We will then show that this minimal model can adequately explain the data.

—*Regulatory complexes have evolved in a stabilizing fitness landscape.* We compute the curvature of the fitness landscape by comparing the distribution of the cross-loci trait diversity in the population, $Q(\Delta_G; \langle\Gamma_G\rangle)$ to its neutral counterpart, $P_0(\Delta_G; \langle\Gamma_G\rangle)$; see eq. (3.20). The two-fold reduction of the mean trait diversity suggests evolutionary dynamics in a fitness landscape with a negative curvature, $f(G) = -\omega\ (G - G^\star)^2$. We estimate the width of the rescaled fitness landscape from eq. (3.20), $2\omega N = 0.38$. This type of fitness landscape imposes a stabilizing selection on the trait value: loss/gain of binding sites are compensated by other gain/loss events. Evolutionary dynamics in this fitness landscape reduces the phenotypic spread in a population r_s by eq. (3.22). We estimate the phenotype spread in the neutral evolution, $r_0 = 2.7$ and under stabilizing selection, $r_s = 1.04$. The 2.5-fold reduction of the trait range for typical loci suggests a substantial fitness effect on regulatory complexes. We will quantify this in the following steps.

In Fig. 3.4(c), we show the distribution of the average binding phenotype Γ_G estimated from the simulated evolution of regulatory complexes in a single quadratic fitness landscape with a negative curvature, $2Nf(G) = -0.38(G - 8)^2$. Each population realization is a cloud of binding phenotypes. The ensemble of populations equilibrates

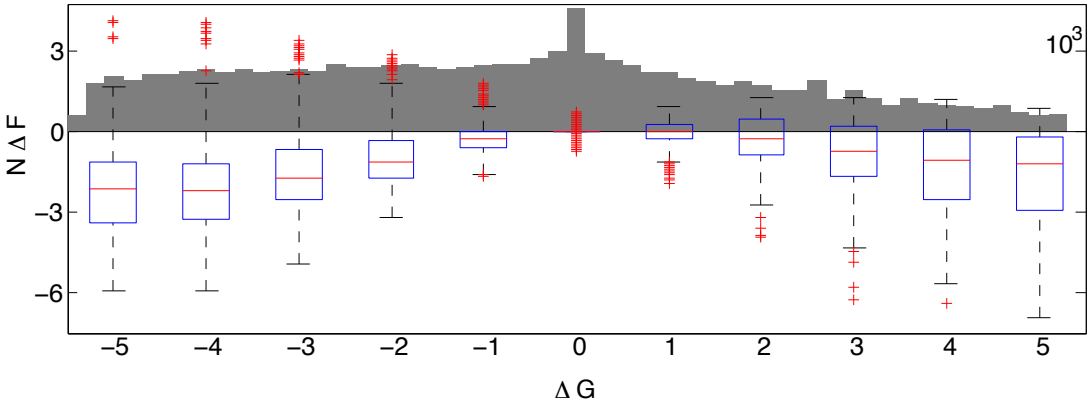

Figure 3.5: Fitness effects of single mutations in regulatory complexes. The gray histogram shows the phenotypic change ΔG for single mutations that we introduce to the low-affinity binding region of 411 *Rap1* regulatory complexes in the *S. paradoxus* genome. The bar plot shows the rescaled fitness effects of these mutations $N\Delta f$ as a function of their phenotypic effect. The fitness function is the minimal epistatic landscape that we infer from the phenotypic polymorphism in Section. 3.4. Single mutations have fitness effects of order one and thus expected to evolve near-neutrally. The substantial stabilizing selection on the trait emerges only from their collective effect; see Fig. 3.4.

around a value which is lower than the peak of the fitness landscape, $G^\star = 8$. This effect is related to the accumulation of deleterious mutations (mutational load) during asexual reproduction and is termed Muller's Ratchet in evolutionary genetics (Muller, 1964).

—*Single mutations experience negligible selection pressure but their collective effect is under strong stabilizing selection.* We compute the fitness effect of single mutations in the weak binding regions of the 411 regulatory complexes in *S. paradoxus*. Most of these mutations reduce the binding phenotype of the regulatory complex, $\overline{\Delta G} < 0$ (Fig. 3.5). We compute the fitness effect of these mutations Δf in the quadratic fitness landscape, $f(G) = -\omega\ (G - G^\star)^2$ with $2N\omega = 0.38$. The location of the peak $G^\star$ is set for each binding complex according to the mean phenotype in the homologous regulatory regions of the 37 *S. paradoxus* samples. The fitness effect of single nucleotide mutations are of order, $N\Delta F \sim 1$ which suggests their effect to be nearly neutral; see Fig. 3.5. The fitness effect of a regulatory complex can be computed by difference between the free-fitness $\tilde{F} = NF - \mathcal{H}$ of the population under selection and in neutrality. $\mathcal{H}$ is the entropy of the phenotype distribution in the population. We estimate

$\delta\tilde{F} = \tilde{F}_s - \tilde{F}_0 = 15.7$, for a typical regulatory complex. There is a substantial fitness effect $\delta\tilde{F} \gg 1$ per regulatory complex to maintain its phenotype during evolution. The stabilizing selection pressure on the regulatory complex emerges from the collective effect of its comprising low-affinity sites in a population.

In summary, our population genetics test can infer the shape of the fitness landscape from the reduction of the phenotype diversity compared to the neutral expectation. This method is particularly useful for systems with heterogenous phenotype preferences. We show that the binding phenotype (free-energy) of the regulatory complexes is under substantial stabilizing selection and is well conserved within *S. paradoxus* populations. At the same time, individual low-affinity sites evolve near-neutrally and show considerable variation in affinity even within one population. Thus, functionality of and selection on regulatory complexes emerge from the entire cloud of sites, but cannot be pinned down to individual sites.

3.5 Divergence of *Rap1* binding complexes across species

In this section, we analyze the evolution of the binding site complexes on a longer timescales i.e., divergence time between three yeast species, *S. paradoxus, S. cerevisiae* and *S. bayanus*. Binding site turnover is more pronounced on longer timescales than among the individuals of a single population. Therefore, we can quantify the ongoing compensatory evolution in the regulatory complexes. Furthermore, we compare the cross-species phenotypic divergence of the regulatory complexes to the expected divergence from the evolutionary dynamics in the quadratic fitness landscape inferred in Section. 3.4. In this way, we evaluate the power of our fitness inference method, which is based on short-term evolutionary dynamics, in predicting the long-term evolution i.e., cross-species divergence.

3.5.1 Conservation of regulatory phenotype between species

Here, we study the phenotypic (free-energy) divergence of the 411 regulatory complexes across homologous regions in three yeast species *S. paradoxus, S. cerevisiae* and *S. bayanus*; see *Material and Methods* in Section. 3.7. We compare the divergence

pattern to the neutral expectation evaluated from the forward simulations of the evolutionary dynamics; see *Materials and Methods* in Section. 3.7. In these simulations, the strong binding sites are kept as observed in the actual genomes and only the flanking weak binding regions have evolved neutrally. Similarly, we simulate the evolution of regulatory complexes under the influence of stabilizing selection that constrains the binding phenotype (free-energy). The fitness landscape is the inferred quadratic fitness in Section. 3.4, $f(G) = -\omega\ (G - G^{\star})^2$. The fitness peak $G^{\star}$ is set separately for each regulatory complex depending on the average binding affinity of that region in the *S. paradoxus* population. The simulations cover the divergence time relevant to species pairs: *S. cerevisiae-S. paradoxus* and *S. cerevisiae-S. bayanus*. The following features can be extracted:

—*The weak-site contribution to the collective binding phenotype is significantly conserved across yeast species compared to the neutral expectation.* We evaluate the binding phenotype (free-energy) of the 411 homologous *Rap1* regulatory complexes in three yeast species, *S. paradoxus, S. cerevisiae* and *S. bayanus*. The difference in the free-energy of the homologous regions is a measure of phenotypic divergence. We show that the collective binding phenotype is highly conserved between the homologous regions of the species pairs and neutral evolution cannot explain such reduction in phenotypic variations; see Fig. 3.6. Neutrally evolved sequences have lower binding affinity (free-energy) and an enhanced phenotypic divergence compared to the actual regulatory complexes. These results are consistent with evolution under stabilizing selection in Section. 3.4.

—*Fitness landscape inferred from the phenotypic variations in the* S. paradoxus *population explains the phenotypic divergence across yeast species.* The fitness landscape that we inferred from the phenotypic polymorphism of the *S. paradoxus* population in Section. 3.4 imposes stabilizing selection on the trait value. Here, we examined the consistency of this minimal epistatic fitness landscape with the long-term divergence data. Can the reduced phenotypic divergence observed across species be explained by the strength of the stabilizing selection, $2Nf(G) = -3.8\ (G - G^{\star})^2$? The phenotypic divergence estimated from the *in-silico* evolutionary dynamics under stabilizing selection matches the phenotypic difference between the homologous region of the species

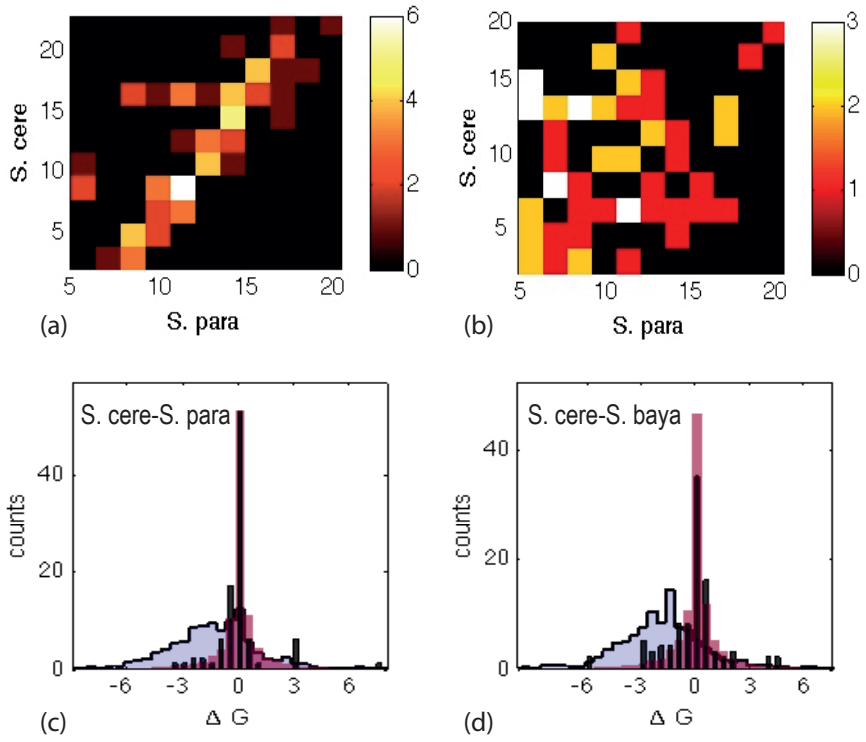

Figure 3.6: Conservation of the collective binding phenotype between yeast species. (a) 2D histogram of the binding free-energy estimates for 411 orthologous regulatory complexes of *Rap1* transcription factor in two species S. cerevisia and S. paradoxus. The color code denotes the number of orthologous binding complexes with the corresponding free-energy values in the two species. The diagonal pattern observed in this histogram shows the significant conservation of the binding phenotype between *S. cerevisiae* and *S. paradoxus*. (b) Shows the same type of 2D histogram for binding free-energy but estimated between *S. cerevisiae* and its neutrally evolved counterpart for a divergence time equal to the evolutionary distance between *S. cerevisiae* and *S. paradoxus*. The fuzzy pattern in comparison to (a) suggests that the neutral dynamics of the low-affinity flanking regions cannot explain the conservation of the binding phenotype across species. (c) The black bars show the distribution of free energy differences between the homologous regulatory regions of *S. cerevisiae* and *S. paradoxus*. The blue histogram on the back shows the expected difference from neutral dynamics of the low-affinity regions. The red histogram shows the free-energy difference between *S. cerevisiae* and the simulated evolution under minimal stabilizing fitness landscape $f(G) = -\omega\ (G - G^{\star})^2$ inferred from phenotypic polymorphism in Section. 3.4. The *in-silico* divergence time corresponds to the evolutionary distance between *S. cerevisiae* and *S. paradoxus* . The inferred quadratic fitness landscape can explain the conservation of the binding phenotype across species. (d) similar to (c) but for a longer evolutionary distance between S. cerevisiae and S. bayanus.

pairs, *S. cerevisiae -S. paradoxus* and *S. cerevisiae-S. bayanus*; see Fig. 3.6(c) and Fig. 3.6(d). In both cases, the divergence is far less than the neutral expectations.

3.5.2 Compensatory evolution in regulatory complexes

The high rate of site turnover across species gives us a chance to study the compensatory evolution of the weak site sequences quantitatively. As the divergence time between organisms increases, their sequence similarity decays. Phenotypic conservation between homologous regulatory complexes in diverged species can be explained in two ways: (i) the constituent loci are all conserved during evolution, i.e. no binding site turnover, or (ii) gain/loss of site sequences is compensated by another loss/gain in the regulatory complex. This analysis requires more knowledge of the functional contributions of individual loci to the overall binding phenotype. For this purpose, we use the marginal occupancy of each locus "o_i" introduced in eq. 3.4. If the constituent loci of the trait are all conserved, their marginal occupancy also remains unchanged between the diverged species, $\delta o_i \sim 0$ (scenario (i)). On the other hand, if the phenotype is conserved only by compensation, a change in the occupancy at one position δo_i is fixed by another change δo_j with opposite sign, $\delta o_i \delta o_j < 0$ (scenario (ii)). This way the overall occupancy profile can still remain conserved. Phenotypic evolution under stabilizing selection can feature both of these scenarios depending on the evolutionary divergence time between organisms. Analysis of short-time divergence (e.g., intra-population comparisons) mostly resemble the first description whereas, the long-term divergence shows evidence for compensatory functional turnovers i.e., the second scenario. On the other hand, the neutral counterpart of the long-term divergence results in independent positional changes with no compensatory effect between mutations to retain the overall function.

With this background, we introduce the following statistic to quantify compensatory evolutionary changes in a regulatory region. The extensive occupancy variable $O = \sum_{i=1}^{\ell} o_i$ is the sum of individual position occupancies in a regulatory complex of length ℓ. The overall occupancy difference between two orthologous sequence regions of two species is a sum of contributions from individual positional differences, $\Delta O = \sum_i o_i^a - o_i^b$ where the upper index denotes the species "a" or "b" . The choice of the reference genome is arbitrary. We compare the overall occupancy divergence $\langle \Delta O^2 \rangle$ for the

homologous regulatory complexes with the additive divergence, $\sum_i \langle \delta o_i^2 \rangle$. The overall averaged divergence reads,

$$\langle (\Delta O)^2 \rangle = \langle \sum_{i=1}^{\ell} \delta o_i^2 \rangle + \sum_{i \neq j} \langle \delta o_i \ \delta o_j \rangle \tag{3.23}$$

$\langle . \rangle$ refers to cross-loci averages as used in previous sections. It is clear that if positional changes in a regulatory complex are independent, the two divergence measures would be equal, $\langle (\Delta O)^2 \rangle = \sum_i \langle \delta o_i^2 \rangle$. We primarily fix $\langle \delta o_i \rangle = 0$ for all positions by subtracting the single species positional average occupancy $\langle o_i^a \rangle$ from individual occupancy values. The difference between the additive and the overall occupancy divergence shows the level of epistasis and compensatory evolution that occurs between the two species.

—*Divergence of regulatory complexes show evidence for compensatory evolution.* We computed the additive and overall occupancy divergence between pairs of species *S. cerevisiae-S. paradoxus* and *S. cerevisiae-S. bayanus*. Fig. 3.7(a) shows that the overall occupancy divergence $\langle \Delta O^2 \rangle$ is significantly smaller than the additive divergence $\sum_i \langle \delta o_i^2 \rangle$ between the two species pairs. This implies that mutations are often compensatory and despite the substitutions between the two species, stabilizing selection maintains the overall binding affinity of the regulatory complex. The difference between these two divergence measures increases with the evolutionary distance between the two species and is more pronounced in the large-distance limit. We also compared the occupancy difference across loci and applied the same type of statistics on those estimates (triangles in Fig. 3.7(a)).This can be thought as an approximate measure for an infinite-time divergence of the regulatory traits that have to maintain their functionality during evolution.

—*Bindings site turnover and stabilizing evolution of the binding phenotype can be reproduced by the minimal epistatic fitness landscape in Section. 3.4.* The observed compensatory evolution in Fig. 3.7(a) cannot be explain by means of linear fitness landscape. The fitness landscape that we inferred from the phenotypic polymorphism in Section. 3.4 however, allows epistatic dynamics during trait evolution. We evaluate the occupancy statistics for the set of simulated sequence pairs that have evolved in

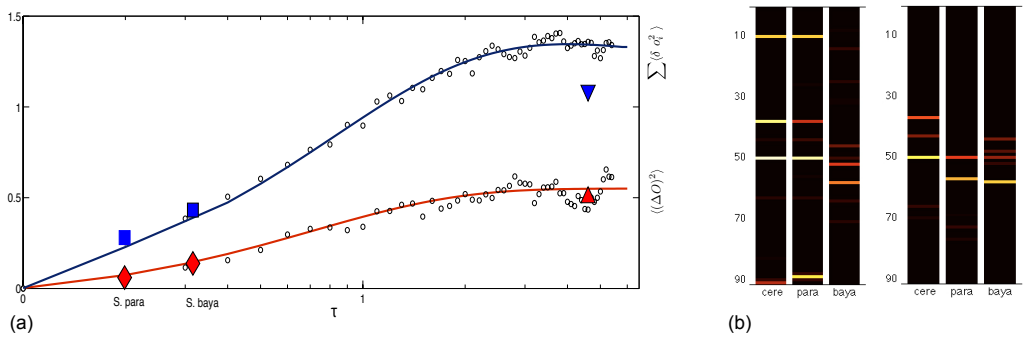

Figure 3.7: Compensatory evolution of the *Rap 1* regulatory complexes. (a) Overall occupancy divergence $\langle(\Delta O)^2\rangle$ (red filled diamonds) versus the additive divergence $\sum_i \langle \delta o_i^2 \rangle$ (blue filled squares) estimated over 411 homologous *Rap1* regulatory complexes between pairs of yeast species, *S. cerevisiae- S. paradoxus* and *S. cerevisia-S. bayanus*. The x-axis shows the divergence time between *S. cerevisiae* and each of the species estimated from nucleotide substitutions in each pair. The long-time data shown by filled triangles are estimated from occupancy divergence between different regulatory complexes. The open circles show the same statistics for simulation data of evolution in a minimal epistatic fitness landscape $f(G) = -\omega(G-G^\star)^2$ with $2\omega N = -0.38$ inferred from phenotypic polymorphism in the *S. paradoxus* population; see Section. 3.4. The full curves are exponential fits to these data points. The significant agreement between the simulations and the actual divergence data proves the strength of our fitness inference method discussed in Section. 3.4. (b) Binding occupancy of two sets of homologous regulatory complexes with length of 100 bp in 3 yeast species, *S. cerevisiae, S. paradoxus* and *S. bayanus*. Each column spans the genomic sequence stretch in which the middle site at position 50 is a strong binding site. The color code shows the relative occupancy (light to dark as high to low). The low affinity sites around the strong site are not all conserved across the species. We can see patterns of site turnover with compensatory effects.

an epistatic fitness landscape with a rescaled width, $2\omega N = 0.38$. Fig. 3.7(a) shows perfect agreement between the time-dependence of the occupancy-statistics predicted by the quadratic fitness model in Section. 3.4 and the actual population divergence. This result further confirms the power of our fitness inference method to predict the long-term evolutionary divergence.

In summary, we show that the joint binding phenotype of the regulatory complexes in yeast is under substantial stabilizing selection and is well conserved within *S. paradoxus* populations and between three species of Saccharomyces. Binding affinity of individual sites vary considerably across yeast species however, their collective function is conserved. We infer a fitness landscape depending on this phenotype based on the

polymorphism data in *S. paradoxus*. Evolution in this minimal epistatic fitness landscape explains the compensatory changes between sites of the regulatory complexes.

3.6 Discussion & Outlook

Evolution of regulatory site complexes

Site clusters in eukaryotes exhibit a different kind of selection pressure during evolution compared to single functional binding sites in prokaryotes. Our thermodynamic framework characterizes the binding affinity of a regulatory region with multiple interacting transcription factor binding sites. We extend the biophysical fitness estimates for single binding site evolution towards the evolution of multiple interacting sites, which involves the transition from single particle to many particle statistics. We infer a fitness landscape depending on this phenotype using yeast whole-genome polymorphism data of *Rap1* regulatory complexes and a new method of quantitative trait analysis. The stationary state of the population is set by mutation-selection-drift balance. These three evolutionary forces influence the binding phenotype on similar time-scales. This analysis goes beyond previous work, which often did not consider regimes where all of three forces are prominent. For example, the quasi-species model disregards genetic drift, and hence is a mutation-selection approximation, or the conventional analysis of the weak selection regime in which fitness differences are small. Incorporating all three evolutionary forces in our analysis, we showed that the joint binding phenotype of these regulatory structures is under substantial stabilizing selection, and is well conserved within *Saccharomyces paradoxus* populations and between three species of *Saccharomyces*. At the same time, individual low-affinity sites evolve near-neutrally and show considerable affinity variation even within one population. Therefore, functionality of and selection on regulatory complexes emerge from the entire cloud of sites, but cannot be pinned down to individual sites.

Modularity and pleiotropy in regulatory complexes

Modularity of a trait, which refers to the connectedness of its constituent elements, is an important aspect of biological organization. In this chapter we studied one type of these functional modules i.e., groups of binding sites which are connected by a common functional regulatory output. We viewed this modularity from two angles. First, we

characterized the modular binding complexes in light of the biophysical properties of cooperative interactions between their constituent binding sites. We then discussed modularity from an evolutionary perspective. Genetic load and accumulation of deleterious mutations make the evolutionary dynamics of linked genomes less efficient. It is essential for organizations such as regulatory regions to maintain their functional efficacy during evolution. The modular structure of these regions brings an opportunity for compensatory evolution by applying only a small selective pressure on individual elements of the module and yet maintaining the overall function. We showed strong evidence for this type of evolutionary dynamics in regulatory complexes.

The fitness effects of genomic mutations in modular regulatory complexes, at least for binding to the transcription factor *Rap1*, is very small. *Rap1* is a highly pleiotropic transcription factor, and regulates a broad range of essential genes in the yeast genome. The regulatory complexes of these genes are all tuned to interact efficiently with the same transcription factor. On the other hand, they confer a high level of sequence flexibility as a result of their modular structures, so that they can remain functional even in challenging environmental conditions.

Combining these two views, we can ask whether genomic modularity and functional pleiotropy are two features which can stably coexist through interactions. For one, complex organisms have developed modular structures, and at the same time, recruit more restricted pleiotropic agents, such as transcription factors, which interact with a subset of the organism's gene network. This will bring further understanding to the emergence of complexity during evolution.

Adaptive dynamics of regulatory complexes

In this chapter, we characterized the evolution of regulatory complexes under equilibrium conditions. We infer a time-independent fitness landscape, which can best describe the phenotypic composition of a population at regulatory loci. However, the possibility of a genotype-phenotype map allows us to address the adaptive dynamics of these complex structures during evolution. We can characterize the cost/benefit of these many-body structures in response to environmental changes. On one hand, genomic linkage in these regions reduces the efficacy of natural selection and limits the speed of adaptation in the population. On the other hand, the standing phenotypic variation in the population, which is set by the mutation-selection-drift balance, has more potential

for adaptive dynamics. Quantifying these relations is necessary for understanding the regulatory "grammar" on larger scales.

3.7 Materials and methods

***Rap1* binding complexes.** *Rap1* transcription factor is the regulator of Telomer, Glycolysis and Ribosomal protein genes in yeast species. The transcription factor *Rap1* binds as dimer (König et al., 1996) and thus makes a suitable case for our study on cooperative regulatory interactions. The frequent occurrence of *Rap1* cognate sites in the genome is also an advantage for our statistical analysis. There is direct evidence that *Rap1* is involved in the regulation of 957 genes (Abdulrehman and Monteiro, 2011; Monteiro et al., 2008; Teixeira and et. al., 2006).

The functional sequence regions are chosen based on the conservation of their strong *Rap1* binding site: the strong site should be present in the homologous regions of the three yeast species . The species genomic alignments are extracted from the SGD genome project (Liti and et. al., 2009). We further verified the predicted functional regions in the subset of ribosomal protein genes for which experimental measurements are available (Lavoie et al., 2010). The *Rap1* regulatory complexes that we study here are regions with one strong binding site and a flanking sequence with a length of 100 base pairs (50 bp on each side of the strong site). The flanking sequence does not contain any strong binding site for *Rap1* or any other transcription factor. The low affinity sites in this region can be functional in cooperation with the strong *Rap1* binding sequence; see Fig. 3.1(b). These criteria retain 411 of the *Rap1* regulatory complexes for our analysis.

***Rap1* binding affinity.** We infer the sequence-dependent binding characteristics of the *Rap1* transcription factor, from *in vivo* CHIP-seq measurements reported by SwissRegulon database (Pachkov et al., 2007). The corresponding binding profile is shown as a frequency logo in Fig. 3.8(a).

Inference of the energy parameters for *Rap1* transcription factor. Scanning the intergenic sequence of *S. cerevisiae* with the *Rap1* position weight matrix produces the histogram of energy counts shown in Fig. 3.8(b). For $E > -5$ (non-specific binding), this distribution is close to Gaussian and can be fit by the energy distribution of

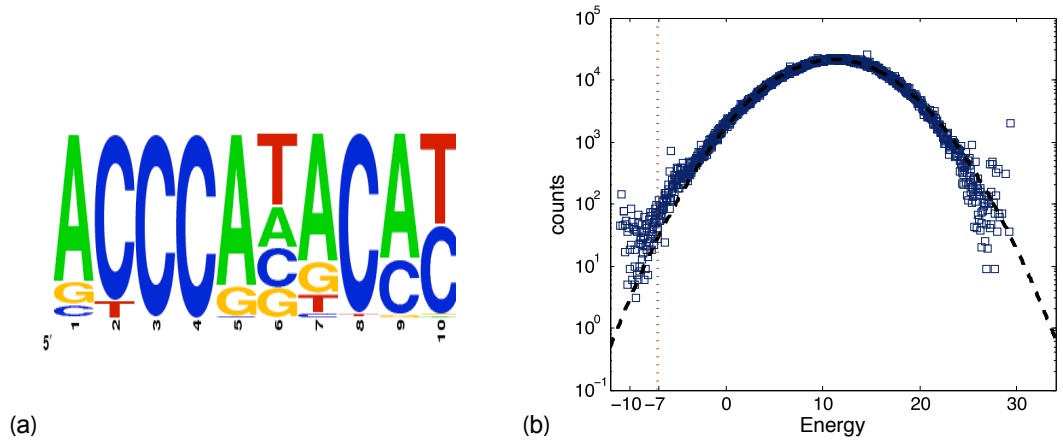

Figure 3.8: *Rap1* binding site characteristics. (a) Sequence logo of the transcription factor *Rap1* evaluated from the *in vivo* CHIP-seq measurements reported in the SwissRegulon database (Pachkov et al., 2007). (b) Histogram of the site energies as predicted by the *Rap1* energy matrix across the intergenic regions of the *S. cerevisiae* genome (dots). The expected energy distribution for the neutral sequence estimated from the background Markov model has a Gaussian form shown by dashed line. The over-representation of the low energy site sequences in e, S. cerevisiae, $E < -5$ is related to the functional *Rap1* binding sites in the genome. The strong functional binding sites are chosen from the tail of the distribution, $E < -7$ which accounts for 0.05% of the sites in the genome.

uncorrelated random site sequences with single-nucleotide frequencies $P_0(a)$ that are chosen from the intergenic sequence of *S. cerevisiae*. The chemical potential parameter in eq. (3.2) is inferred to be equal to the energy threshold for specific bindings, $\nu = -7$. This is a log-likelihood estimate for ν which best fits the single-site energy model that we use in our analysis; see e.g.,(Mustonen et al., 2008; Mustonen and Lässig, 2005). The strong sites have energy values $E < -7$ which accounts for 0.05% of the sites in the genome. The energy values are measured in arbitrary units set by the information content of individual positions in the PWM; biophysical measurements are required to determine the absolute binding energy values. In our analysis, we set the strength of cooperativity to $\omega = -4$ which is in the range for the specific binding energy of *Rap1* transcription factor. Our conclusions are not sensitive to the exact value of ω. This is a minimal biophysical model for interaction of multiple binding sites.

Forward simulations for the evolutionary dynamics of regulatory complexes. The numerical methods used to simulate the evolutionary dynamics is based on the Wright-Fisher model of asexually reproducing population with non-overlapping gener-

ations and fixed population size N. Individuals are represented by nucleotide sequences $\mathbf{a}(a_1 \ldots a_\ell)$ of length $\ell = 100$ bp. The genotype-phenotype map $G(\mathbf{a})$ is based on the binding affinity model described in Section. 3.2. Each generation is stochastically sampled from the previous generation, using a multinomial sampling process, in which the probability p_i of picking individual i is given by, $p_i = \mathrm{e}^{F(G_i)-\overline{F}}$ with mean population fitness, $\overline{F} = \frac{1}{N}\sum_i F(G_i)$. We initialize the program with a monomorphic population where all genotypes are identical to the original promoter. In each run of the simulation, we let the population equilibrate before initiating any measurement to ensure that the population is in the stationary state. In each generation, mutations change the genotypes with rate μ per nucleotide.

4

Evolution of polygenic traits

4.1 Introduction

Studies of polygenic traits such as organism's height, eye color, drug resistance, etc. have been traditionally pursued in the context of quantitative genetics (Falconer, 1989; Lynch and Walsh, 1998). While quantitative genetics mainly focuses on a phenomenological description of phenotypic distributions (Fisher, 1930; Hartl and Taubes, 1996; Rice, 1990), population genetics is concerned with the genetic composition of a population in response to the primary evolutionary forces: natural selection, mutation, genetic drift, recombination, etc. (Kimura, 1968; McDonald and Kreitman, 1991; Smith, 1970); see discussion in Chapter. 1. However, recent advancements in identifying molecular components of phenotypic traits, make the need to combine these two fields more evident; some examples include, (Bedford and Hartl, 2009; Berg et al., 2004; Fernández and Lynch, 2011; Ludwig et al., 1998; Mustonen et al., 2008; Poelwijk et al., 2007; Weinreich et al., 2006).

As the term "polygenic" suggests, a large number of loci contribute to these continuous traits, and it is their collective phenotype that effectively responds to any selection pressure during evolution. Efforts have been devoted to draw analogies between quantitative genetics and the concepts of statistical physics. For one, both fields are concerned with analysis of many-particle systems to quantify their collective behavior without following all the microscopic details. Most of the studies in this direction characterize traits that are comprised of independently evolving loci: either recombination is very rapid in the organism that assures an efficient loci reassortment, or the

loci are encoded far apart in the genome or may be on different chromosomes and are not physically linked together (de Vladar and Barton, 2011b; Kirkpatrick et al., 2002). This condition, known as "linkage-equilibrium", certainly is of interest from a theoretical point of view because of its consequent simplifications in analytical descriptions. The framework to study the evolutionary dynamics of each of these independent loci resembles that of single-particle statistics with relatively tractable microscopic details. Analogies to statistical physics then provide a quantitative description of the collective dynamics of these independent particles and characterize the role of natural selection on their organization (Barton and Turelli, 1989; Barton and Coe, 2009; de Vladar and Barton, 2011a; Fisher, 1930; Kirkpatrick et al., 2002; Neher and Shraiman, 2011; Sella and Hirsch, 2005).

However, biological systems are not necessarily tuned to simplify our theoretical understanding. As a matter of fact, genomic traits do not often satisfy the required conditions for linkage-equilibrium. The evolutionary dynamics of linked genomes, e.g., in asexual populations of microbes and viruses, involves a lot of genomic complications: many mutations appear together just because of this physical constraint and natural selection cannot treat them separately and distinguish between their individual fates (Desai, 2007; Desai and Fisher, 2007; Gerrish and Lenski, 1998; Park and Krug, 2007; Rouzine et al., 2008; Schiffels et al., 2011). In sexual populations, finite recombination rate sets the correlation length between the genomic positions. This way, linkage is not only the property of asexual organisms but clusters of nucleotides in sexual genomes with low rate of recombination, experience its evolutionary consequences (Comeron and M, 2002).

The study of these quantitative traits, which are encoded by multiple linked loci, would correspond to the field of many-particle statistics that follows complicated and chaotic dynamics at the microscopic level (Baxter and Blythe, 2007; Mustonen and Lässig, 2010). Thermodynamics has thought us that despite such microscopic complications, the macroscopic properties of a system can still follow simple rules. In this chapter, we will approach this problem from a macroscopic point of view. We introduce a coarse-grained description of a population by mapping its individual-based genomic components to population-based phenotype statistics. This macroscopic description proves to be very essential in eliminating the small-scale microscopic complications, yet it is informative enough to explain standing variation in populations. The large number

of contributing loci in a quantitative trait allows us to explain linked genomic variations by means of thermodynamics analogies. We show that despite the existence of linkage-disquilibrium on a microscopic level, intra-population macroscopic observables can still be described within the framework of an equilibrium statistical physics. We can infer the shape of the fitness landscape on which the evolutionary dynamics has been directed by quantifying the deviation of phenotype statistics from the neutral expectations. Our theory is most applicable at the current state of sequencing techniques with an ease and feasibility of acquiring large scale genomic samples from populations; see Chapter. 3 for application of this theory. The macroscopic description of the linked genome is summarized in Section. 4.4.

4.2 Evolution of genotype, allele and phenotype distribution

We start our analysis with a complete picture of the population evolution which tracks the frequency changes of all the genotypes in the population through time. Although this picture captures the full population information, it proves to be rather detailed and thus an adequate number of samples from the population can hardly be provided for this model. One of the most common approaches to overcome this problem is to follow the marginal allele frequencies in the population instead of the complete genotype compositions. We will then use the allelic description of the genomic system to build a phenotype-based evolutionary theory that characterizes the macroscopic trait observables of a population both for free-recombining and for linked genomes.

4.2.1 Stochastic evolution of genotypes

We consider the evolution of a population with k possible genotypes $\mathbf{a}_\alpha$ ($\alpha = 1, \ldots, k$), in which each genotype is a sequence $\mathbf{a} = (a_1, \ldots, a_\ell)$ of length ℓ with letters $a_i = A, C, G, T$. This dynamics is defined on the space of genotype frequencies x^α with the constraints $x^\alpha \geq 0$ ($\alpha = 1, \ldots, k$) and $\sum_{\alpha=1}^{k} x^\alpha = 1$, which is a $(k-1)$dimensional simplex denoted by Σ_{k-1}. Here we use the set of $k-1$ linearly independent frequencies $x = (x^1, \ldots, x^{k-1})$ as coordinate system on Σ_{k-1} (however, most of the equations below do not depend on the choice of any particular coordinate system). Unless otherwise

specified, coordinate indices $\alpha, \beta, \ldots$ take the values $1, \ldots, k-1$, and we use the convention that if the same coordinate appears in a product as upper and lower index, it is summed over, e.g., $s_\alpha x^\alpha \equiv \sum_{\alpha=1}^{k-1} s_\alpha x^\alpha$. The shorthand $\partial_\alpha \equiv \partial/\partial x^\alpha$ denotes partial derivatives with respect to these coordinates.

Stochastic evolution of finite populations is described by a time-dependent probability distribution of genotype frequencies, $P(x,t)$. The evolution of $P(x,t)$ can be described by a Kimura-Ohta evolution equation (generalized diffusion equation) of the form (Kimura and Ohta, 1969),

$$\partial_t P(x,t) = \partial_\alpha \left[\frac{1}{2N} \partial_\beta g^{\alpha\beta}(x) - v^\alpha(x,t) \right] P(x,t) \tag{4.1}$$

Here, N is the effective population size, $g^{\alpha\beta}$ are response coefficients, and $v^\alpha(x,t)$ are the total rates of frequency change due to selection $s^\alpha(x,t)$, mutations $m^\alpha(x,t)$, and recombination $\rho^\alpha(x,t)$,

$$v^\alpha(x,t) = s_\alpha(x,t) + m^\alpha(x,t) + \rho^\alpha(x,t) \tag{4.2}$$

The diffusion equation eq. (4.1) expresses the temporal change of $P(x,t)$ as the divergence of a probability current, $\partial P(x,t) = -\nabla \cdot \mathbf{J} P(x,t)$. It captures both changes due to the stochastic sampling noise during the reproduction of a finite size discrete population (i.e., genetic drift) and the deterministic changes cased by the forces that act on larger time scales than a generation, (i.e., mutation, selection and recombination).

Selection and fitness landscape for genotypes. Selection is given by genotype fitness values $f_\alpha(x,t)$ (reproductive success), which determine the deterministic change of genotype frequencies in the absence of mutations and genetic drift,

$$\frac{1}{x^\alpha} \frac{dx^\alpha}{dt} = f_\alpha(x,t) - \sum_{\alpha=1}^{k} x^\alpha f_\alpha(x,t), \tag{4.3}$$

for $\alpha = 1, \ldots, k$. The second term on the right hand side ensures conservation of the constraint $\sum_{\alpha=1}^{k} x^\alpha(t) = 1$. In terms of the linearly independent frequencies $x =$

$(x^1, \ldots, x^{k-1})$, these evolution equations take the form (Mustonen and Lässig, 2010),

$$\frac{dx^\alpha}{dt} = s^\alpha(x,t) \equiv g^{\alpha\beta}(x) s_\beta(x,t) \tag{4.4}$$

with selection coefficients,

$$s_\beta(x,t) = f_\beta(x,t) - f_k(x,t) \tag{4.5}$$

and response coefficients,

$$g^{\alpha\beta}(x) = \begin{cases} -x^\alpha x^\beta & \text{if } \alpha \neq \beta \\ x^\alpha(1-x^\alpha) & \text{if } \alpha = \beta. \end{cases} \tag{4.6}$$

The inverse of this matrix, $g_{\alpha\beta} = (g^{\alpha\beta})^{-1}$, plays the role of a metric on Σ_{k-1}. In writing the continuum evolution equations (4.4), it is assumed that the selection coefficients are small on the time scale of a generation and have temporal correlations over much larger times than a generation. By eq. (4.5), the selection coefficient s_α can be expressed as partial change of the mean population fitness in response to a change in the frequency x^α at constant genotype fitness values,

$$s_\alpha(x,t) = \left[\frac{\partial}{\partial x^\alpha} \sum_{\beta=1}^{k} x^\beta f_\beta(y,t) \right]_{y=x} . \tag{4.7}$$

Evolutionary equilibrium can be reached if the selection coefficients are time-independent and can be expressed as the gradient of a scalar fitness landscape,

$$s_\alpha(x) = \partial_\alpha F(x). \tag{4.8}$$

In general, we have to distinguish mean population fitness and fitness landscape: the former governs the overall growth rate of population size, the latter depends only on growth rate differences between genotypes according to eq. (4.5). Directional selection with arbitrary epistasis is given by a linear landscape, $F(x) = s_\beta x^\beta$, whereas a nonlinear landscape describes frequency-dependent selection. Epistatic interactions between loci introduces a dependency of the selection coefficient of changes at each locus on the existing alleles of other loci (background sequence). We have seen such effects in the evolution of the transcription factor binding sites discussed in Chapter. 1 (Berg

et al., 2004; Mustonen et al., 2008; Mustonen and Lässig, 2005) and the evolution of regulatory complexes in Chapter. 3.

Mutations and recombination. In the absence of selection and genetic drift, the frequency change due to mutations has the form,

$$\frac{dx^\alpha}{dt} = m^\alpha(x) = \sum_{\beta=1}^{k} \mu_\beta^\alpha x^\beta - \big(\sum_{\beta=1}^{k} \mu_\alpha^\beta\big) x^\alpha \tag{4.9}$$

given by the mutation rates $\mu_\alpha^\beta \equiv \mu(\mathbf{a}_\alpha \to \mathbf{a}_\beta)$ between genotypes $(\alpha, \beta = 1, \ldots, k)$, which we assume to be time-independent over the interval of observation. We can rewrite the rate $m^\alpha(x)$ in terms of the linearly independent frequencies $x = (x^1, \ldots, x^{k-1})$,

$$m^\alpha(x) = \hat{\mu}_\beta^\alpha x^\beta + \mu_k^\alpha \tag{4.10}$$

with

$$\hat{\mu}_\beta^\alpha = \begin{cases} \mu_\beta^\alpha - \mu_k^\alpha & (\alpha \neq \beta) \\ -\sum_{\gamma=1}^{k} \mu_\alpha^\gamma - \mu_k^\alpha & (\alpha = \beta). \end{cases} \tag{4.11}$$

The "covariant" rates $m_\alpha(x) = g_{\alpha\beta} m^\beta(x)$ are defined in analogy to (4.4).

We assume the evolutionary process is in the low mutation regime $\mu N \ll 1$ (where N is the effective population size) and the substitution rates u_α^β satisfy the detailed balance conditions,

$$\frac{u_\alpha^\beta}{u_\beta^\alpha} = \frac{p_0^\beta}{p_0^\alpha} \tag{4.12}$$

where p_0^α is the neutral probability distribution for the genotype $\mathbf{a}_\alpha$ $(\alpha = 1, \ldots, k)$. These conditions, which are fulfilled in all standard models of nucleotide mutation rates, imply that the neutral substitution dynamics in the discrete space of genotypes $\mathbf{a}_1, \ldots, \mathbf{a}_k$ has an equilibrium probability distribution p_0^α. It is then straightforward to show that the rates $m_\alpha(x)$ are asymptotically of gradient form,

$$m_\alpha(x) = \partial_\alpha M(x) + O(\mu^2 NL). \tag{4.13}$$

The gradient property is tied to the existence of an evolutionary equilibrium under

mutations and genetic drift. The equilibrium frequency distribution $P_0(x)$ turns out to be simply related to the "mutation potential" $M(x)$. Deviations from the form (4.13) arise only from multiple simultaneous mutations and are negligible for compact genomic units ($\mu NL \ll 1$) such as transcription factor binding sites.

If the detailed balance conditions (4.12) are replaced by the more restrictive conditions $\mu_\alpha^\beta = \mu^\beta$, the rates m_α are of exact gradient form for arbitrary values of μN and the mutation potential is known exactly (Baxter and Blythe, 2007)

$$M(x) = \sum_{\alpha=1}^{k} \mu^\alpha \log(x^\alpha). \tag{4.14}$$

In sexually reproducing populations, two genotypes α and β can recombine and produce a new genotype γ. The consequent reshuffling of the polymorphic loci in the population makes the exploration of the genotypic space more effective. The genotype frequency changes due to recombination are described by additional terms in eq. (4.9). The effect of finite recombination on evolutionary dynamics of multi-locus traits has been discussed by (Barton and Otto, 2005; Neher and Shraiman, 2009, 2011; Rouzine, 2010). Despite the significant adaptive role of recombination, we will only consider the two limits of zero and infinite recombination in this chapter. Further extension to include a finite recombination rate is certainly crucial (Neher and Shraiman, 2011).

4.2.2 Stochastic evolution of alleles

Genotype space for multiple genomic loci is very high-dimensional (4^ℓ-D for the system described above) and always under-sampled. However, appropriate marginal distributions and averages of $P(x,t)$ (such as allele frequencies at single loci) can be compared with observations. The genotype frequencies x can be projected onto marginal distributions at individual genomic loci, pairs of loci, etc.. These marginal frequencies can be defined by the nucleotide identity variables,

$$\varepsilon_i^a \equiv \delta(a_i, a) \tag{4.15}$$

where $\delta(a,b)$ is a discrete delta function which takes value 1 for the identity $a = b$ and 0, otherwise. The identity variable ε_i^a then takes value 1 when the letter at position

i of the sequence is a and 0, otherwise. The expectation values and correlation functions within a population (denoted by overbars) then follow,

1. *Allele frequencies at individual loci:*

$$y_i^a \equiv \overline{\varepsilon_i^a} = \sum_\alpha \delta(a_{\alpha,i}, a)\, x^\alpha. \tag{4.16}$$

As before, x^α is the frequency of genotype a_α in the population. With the normalization constraints $\sum_{a=1}^4 y_i^a = 1$ $(i = 1, \ldots, \ell)$, the space of allele frequencies has dimension 3ℓ. We use the shorthand y for the set of allele frequencies at all genomic loci,

$$y = ((y_1^1, \ldots, y_1^4), \ldots, (y_\ell^1, \ldots, y_\ell^4)). \tag{4.17}$$

2. *Allele variation at individual loci:*

$$g_i^{ab} = \overline{(\varepsilon_i^a - y_i^a)(\varepsilon_i^b - y_i^b)} = y_i^a \delta(a, b) - y_i^a y_i^b \tag{4.18}$$

The matrix g_i determines the allele diversity

$$\pi_i \equiv \mathrm{Tr}\, g_i = \sum_a y_i^a (1 - y_i^a). \tag{4.19}$$

3. *Haplotype frequencies* for pairs of loci:

$$y_{ij}^{ab} = \overline{\varepsilon_i^a \varepsilon_j^b} = \sum_\alpha \delta(a_{\alpha,i}, a)\delta(a_{\alpha,j}, b)\, x^\alpha \quad (i \neq j) \tag{4.20}$$

and connected frequency correlations (linkage disequilibrium)

$$c_{ij}^{ab} \equiv \overline{(\varepsilon_i^a - y_i^a)(\varepsilon_j^b - y_j^b)} = y_{ij}^{ab} - y_i^a y_j^b \quad (i \neq j). \tag{4.21}$$

The projection from genotype to allele frequencies involves no loss of information **if and only if** the genotypes in the population are at linkage equilibrium, i.e., if the frequency x of each sequence $\mathbf{a}$ is the product of the frequencies of its alleles,

$$x = \hat{x}(y) = \prod_{i=1}^{\ell} y_i^{a_i} \tag{4.22}$$

We refer to this case as the limit of free recombination; note that if the number of polymorphic loci is large, the factorization (4.22) can only hold approximately, because not all genotypes can be sampled in a finite population. The allele frequency distribution $P(y,t)$ then follows an autonomous evolution equation,

$$\partial_t P(y,t) = \sum_{i=1}^{\ell} \frac{\partial}{\partial y_i^a} \left[\frac{1}{2N} \frac{\partial}{\partial y_i^b} g^{ab}(y_i) - v_i^a(y,t) \right] P(y,t) \tag{4.23}$$

where $g^{ab}(y_i)$ are allele response coefficients and $v_i^a(y,t)$ are deterministic rates of the allele frequency change at locus i. This rate involves the projection of selection onto individual loci and the contribution of mutations,

$$v_i^a(y,t) = s_i^a(y,t) + \sum_{i=1}^{\ell} m_i^a(y_i) \tag{4.24}$$

4.2.3 Stochastic evolution of phenotypic traits

Polygenic traits can in general be any arbitrary nonlinear function of their comprising loci. In the following theoretical analysis, we study the simple case of additive phenotype ϕ with ℓ constitutive nucleotide loci. We are considering a low-mutation rate regime which is valid for most biological systems. Therefore, it is justified to assume that the polymorphic genomic positions carry not more than two types of nucleotides in the population. Thus, the 4-state nucleotide composition at each loci $a_i = A, C, G, T$ can be reduced to a 2-state model. Using the nucleotide identity variable in eq. (4.15), we can assign the binary states $\{0, 1\}$ to the coexisting nucleotides at position i in the population. We set our choice such that the a-allele associated with $\varepsilon_i^a = 1$ contributes by the amount γ_i and the other allele makes zero contribution to the trait. The resulted additive phenotype is,

$$\phi(\mathbf{a}) = \sum_{i=1}^{\ell} \gamma_i \varepsilon_i \tag{4.25}$$

Further nonlinearities can in general be added to this model in the form of a Fourier

expansion,

$$\phi(\mathbf{a}) = \sum_{i=1}^{\ell} \gamma_i \varepsilon_i + \sum_{i \neq j} \gamma_{ij} \varepsilon_i \varepsilon_j + \dots \tag{4.26}$$

Even in the regime of low-mutation rates $\mu N \ll 1$, the large number of comprising loci of a trait ($\ell \gg 1$) generates a broad distribution of trait values among the individuals of a population (because $\mu N \ell \sim 1$). Consequently, the populations should be pictured as clouds of haplotypes rather than point-like monomorphic objects responding to the underlying evolutionary forces. Ultimately, we would be interested in understanding the evolutionary dynamics of the trait distribution in populations. Here, we characterize the intra-population trait distribution by its moments.

The population mean phenotype $\Gamma \equiv \overline{\phi}$ depends only on the allele frequencies at individual loci,

$$\Gamma(y_i) = \sum_{i=1}^{\ell} \gamma_i \, \overline{\varepsilon_i} = \sum_{i=1}^{\ell} \gamma_i y_i \tag{4.27}$$

where y_i is chosen to be the frequency of allele "1" at position "i". The intra-population phenotype diversity, $\Delta \equiv \overline{\phi^2} - \overline{\phi}^{\,2}$ depends on the allele variation of individual loci and connected correlations for pairs of loci (see eq. (4.19) and eq. (4.21)),

$$\begin{aligned} \Delta(y_i, y_{ij}) &= \sum_{i,j=1}^{\ell} \gamma_i \gamma_j (\overline{\varepsilon_i \varepsilon_j} - \overline{\varepsilon_i}\,\overline{\varepsilon_j}) \\ &= \sum_i \gamma_i^2 \pi_i + \sum_{i \neq j} \gamma_i \gamma_j c_{ij} \end{aligned} \tag{4.28}$$

where $\pi_i = y_i(1 - y_i)$ and $c_{ij} = y_{ij} - y_i y_j$ are single locus allele variation and connected frequency correlation.

For traits with large number of contributing loci, we expect a Gaussian distribution of phenotypes in a population which is fully characterized by its first two moments (Central limit theorem). In our analysis, we will **not** explicitly use a Gaussian approximation for the trait distribution and leave our theory to be applicable beyond those limits. However, we will confine our analysis to the first two moments of the phenotype distribution Γ, Δ as independent macroscopic population observables; see *Discussion* in

Section. 4.4 for further generalization of this scheme. For now, we also assume $\gamma_i = 1$. The general case is not much different.

Distribution of trait statistics $P(\Gamma, \Delta, t)$ then follows the autonomous evolution dynamics,

$$\partial_t P(\Gamma, \Delta, t) = \partial_\alpha \left[\frac{1}{2N} \partial_\beta g^{\alpha\beta}(\Gamma, \Delta) - v^\alpha(\Gamma, \Delta, t) \right] P(\Gamma, \Delta, t) \tag{4.29}$$

We use Einstein summation convention for the distribution parameters Γ and Δ indicated by Greek indices. In the following sections, we compute the response coefficients $g^{\alpha\beta}$ and the deterministic evolutionary forces $v^\alpha = s^\alpha + m^\alpha$ for the projection from the marginal allele frequencies onto the $\Gamma - \Delta$ coordinate. We will characterize the dynamics of these macroscopic observables for the two regimes of free-recombination and asexual evolution (no recombination). The summary of these results is presented in Section. 4.4 and in Table. 4.1.

4.2.4 Phenotypic equilibrium for free-recombining loci

Classical quantitative Genetics is based on the assumption that genotypes are random re-assortments of alleles, each occurring with a certain frequency (de Vladar and Barton, 2011b; Falconer, 1989; Kirkpatrick et al., 2002; Lande, 1976; Lynch and Walsh, 1998). This absence of correlations between alleles at different loci ($c_{ij} = 0$) is termed "linkage equilibrium", implying that recombination has relaxed correlations between loci (Barton and Turelli, 1989; Barton and Coe, 2009; de Vladar and Barton, 2011a; Fisher, 1930; Kirkpatrick et al., 2002; Neher and Shraiman, 2011). The mean trait of the population remains as $\Gamma = \sum y_i$ but the trait diversity reduces to the additive part of the variance, $\Delta = \sum y_i(1 - y_i)$. In the long-run, allele frequencies at each loci reach the equilibrium state of the form of the Kimura's U-shape distribution (Kimura, 1962). The full equilibrium distribution then factorizes between loci (see eq. (4.22)).

4.2.4.1 Neutral evolution of free-recombining loci

Genetic Drift. Random sampling in a discrete population of a finite size N introduces a stochastic force during the reproductive process which is termed genetic drift. For a

two-allele system ($\varepsilon_i = 0, 1$), this stochasticity is simply a Gaussian random variable for sampling from the two subpopulations of the size N_i^1 and $N_i^0 = N - N_i^1$,

$$\langle \xi_i^a \rangle = 0 \quad , \quad \langle \xi_i^a(t) \xi_i^b(t') \rangle = N_i^a(t) \delta(t - t') \delta_{a,b} \tag{4.30}$$

where upper indices $\{a, b\}$ denote the allele types $\{0, 1\}$ and lower indices point to genomic positions. In large populations, allele frequencies $y_i = N_i^1/N$ can be treated as continuous variables. We can then simply map the noise of the discrete variables ξ_i^a in eq. (4.30) onto the continuous allele frequencies, $\chi_i(t) = (\partial y_i/\partial N_i^0)\xi_i^0 + (\partial y_i/\partial N_i^1)\xi_i^1$. Assuming linkage-equilibrium, the statistics of the allele frequency noise $\chi_i(t)$ follows,

$$\langle \chi_i(t) \rangle = 0 \quad , \quad \langle \chi_i(t) \chi_j(t') \rangle = \frac{y_i(1 - y_i)}{N} \delta(t - t') \delta_{i,j} \tag{4.31}$$

We can now map the set of allele frequencies y_i ($i = 1 \dots \ell$), to the macroscopic observables $\Gamma = \sum_i y_i$, and $\Delta = \sum_i \pi_i = \sum_i y_i(1 - y_i)$. As a result, the noise terms in the $\Gamma - \Delta$ coordinate follow,

$$\chi_\Gamma(t) = \sum_i \chi_i(t) \tag{4.32}$$

$$\chi_\Delta(t) = \sum_i (1 - 2y_i)\ \chi_i(t) = \sum_i \sqrt{1 - 4\pi_i}\ \chi_i(t) \tag{4.33}$$

with the following statistics,

$$\langle \chi_\Gamma(t) \rangle = 0 \quad , \quad \langle \chi_\Gamma(t) \chi_\Gamma(t') \rangle = \frac{1}{N} \Delta(t) \delta(t - t') \tag{4.34}$$

$$\langle \chi_\Delta(t) \rangle = 0 \quad , \quad \langle \chi_\Delta(t) \chi_\Delta(t') \rangle = \frac{1}{N} \sum_i (1 - 4\pi_i(t)) \pi_i(t)\ \delta(t - t') \tag{4.35}$$

Mutations. At the level of an individual, mutations are stochastic events often coupled to reproduction that change the alleles, $1 \to 0$ or vice versa. For simplicity, we assume that mutations occur with a uniform rate μ at all nucleotide positions (i.e., $\mu_{1\to 0} = \mu_{0\to 1} = \mu$). In order to quantify the effect of mutations on the trait statistics Γ and Δ, we first need to compute their effect on allele frequencies, y_i. The change in the number

of individuals with allele "1" (N_i^1) in the population due to mutations per generation is,

$$\delta N_i^1 = \mu(N_i^0 - N_i^1) = \mu(N - 2N_i^1) \tag{4.36}$$

We can now simply compute the temporal changes of the marginal frequency, $\delta y_i = \delta N_i^1/N$ due to mutations and genetic drift,

$$\frac{\mathrm{d}}{\mathrm{d}t} y_i = \mu(1 - 2y_i) + \chi_i \tag{4.37}$$

noise term χ_i satisfies the conditions in eq. (4.31). In populations with perfect allele re-assortment (free-recombination), marginal allele frequencies are sufficient to characterize the macroscopic trait observables, Γ and Δ.

1. *Trait average*, Γ

   Trait average of an additive phenotype in eq. (4.27) is a linear combination of single-locus allele frequencies, $\Gamma = \sum y_i$. Therefore, the temporal change of the intra-population trait average follows from eq. (4.34) and eq. (4.37),

   $$\frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t} = -2\mu(\Gamma(t) - \ell/2) + \chi_\Gamma \tag{4.38}$$

   χ_Γ satisfies the conditions in eq. (4.34). From the stochastic Langevin equation in eq. (4.38), we can derive a Fokker-Planck equation that characterizes the dynamics of the underlying probability distribution for the intra-population trait average $P_0(\Gamma, t)$; see discussions on stochastic processes e.g., in (Gardiner, 2004).

   $$\frac{\partial}{\partial t} P_0(\Gamma, t) = \frac{1}{2N} \frac{\partial^2}{\partial \Gamma^2} \Delta(t) P_0(\Gamma, t) + 2\mu \frac{\partial}{\partial \Gamma} (\Gamma - \ell/2) P_0(\Gamma, t) \tag{4.39}$$

   This is conceptually an important matter: The probability distribution $P_0(\Gamma, t)$ in eq. (4.39), characterizes the likelihood of the intra-population trait average Γ across different realizations of a population and should not be mistaken with the trait distribution in single population, $\rho(\phi)$.

We obtain the stationary solution of the marginal distribution $P_0(\Gamma; \langle\Delta\rangle)$ by using the result of the following eq. (4.49) to integrate over the intra-population trait diversity. We approximately use $\Delta \approx \langle\Delta\rangle = \mu N\ell(1-4\mu N)$; see Fig. (4.1).

$$P_0(\Gamma; \langle\Delta\rangle) = \frac{1}{\sqrt{\pi\ell/2}} \exp[\frac{-2(\Gamma-\ell/2)^2}{\ell(1-4\mu N)}] \tag{4.40}$$

Cross-population mean and variance of the trait average Γ are,

$$\langle\Gamma\rangle = \ell/2 \qquad \langle\Gamma^2\rangle - \langle\Gamma\rangle^2 = \ell(1-4\mu N)/4 \tag{4.41}$$

2. *Trait diversity*, Δ

In a free-recombining genome, trait diversity is a linear combination of single locus allele variation, $\Delta = \sum \pi_i = \sum y_i(1-y_i)$. Due to the complicated noise term in eq. (4.35), we first compute the statistic of the single locus diversity $\Delta_i = \pi_i$ and then apply the central limit theorem to characterize the equilibrium distribution of the full trait diversity, Δ. The temporal change of the single-locus allele diversity Δ_i due to mutations and genetic drift is,

$$\frac{\mathrm{d}\Delta_i}{\mathrm{d}t} = \mu(1-4\Delta_i) - \frac{\Delta_i}{N} + \chi_{\Delta_i} \tag{4.42}$$

with a Gaussian noise term χ_{Δ_i} ,

$$\langle\chi_{\Delta_i}(t)\rangle = 0 \quad , \quad \langle\chi_{\Delta_i}(t)\chi_{\Delta_i}(t')\rangle = \frac{\Delta_i(1-4\Delta_i)}{N}\ \delta(t-t') \tag{4.43}$$

The term Δ_i/N in eq. (4.43) appears due to the nonlinear dependency of the trait diversity Δ_i on marginal frequency y_i; see e.g., the discussion on Ito Calculus in Chapter. 4 of (Gardiner, 2004). The corresponding Fokker-Planck equation for the cross-population distribution of the single-locus allele variation takes the form,

$$\frac{\partial}{\partial t}p_0(\Delta_i,t) = \left[\frac{1}{2N}\frac{\partial^2}{\partial\Delta_i^2}\Delta_i(1-4\Delta_i) - \frac{\partial}{\partial\Delta_i}(\mu(1-4\Delta_i) - \Delta_i/N)\right] p_0(\Delta_i,t) \tag{4.44}$$

which yields an equilibrium distribution,

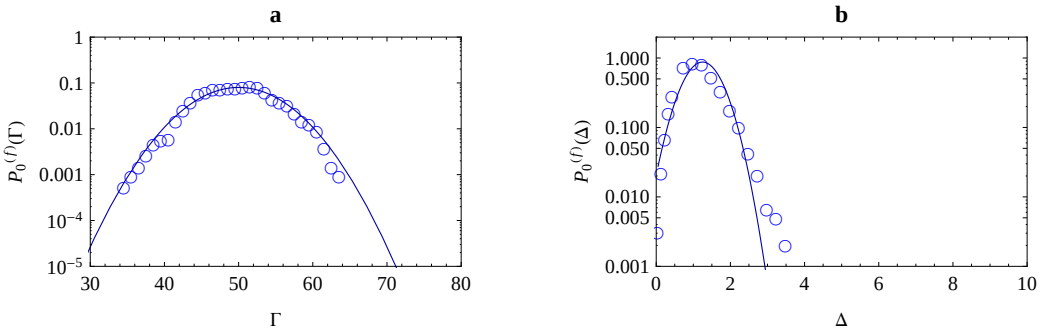

Figure 4.1: Equilibrium distribution of free-recombining traits in neutrality. (a) The analytical estimate for the marginal distribution of intra-population trait average $P_0(\Gamma;\langle\Delta\rangle)$ (full line) vs. the numerical result for simulated evolution (open circles) of the free-recombining traits; eq. (4.40). (b) Same comparison as in (a) for the trait diversity observable Δ; eq. (4.48). The simulation parameters are chosen: $N = 500$, $\ell = 100$ bp and $\mu N = 0.0125$.

$$p_0(\Delta_i) = \frac{1}{Z_{\Delta_i}} \frac{\Delta_i^{-1+2\mu N}}{\sqrt{1-4\Delta_i}} \tag{4.45}$$

Z_{Δ_i} is the appropriate normalization factor,

$$Z_{\Delta_i} = 4^{-2\mu N}\sqrt{\pi}\frac{\Gamma[2\mu N]}{\Gamma[1/2+2\mu N]} \tag{4.46}$$

Cross-population mean and variance of single-locus allele diversity are,

$$\langle\Delta_i\rangle = \mu N(1-4\mu N) + \mathcal{O}((\mu N)^3) \qquad \langle\Delta_i^2\rangle - \langle\Delta_i\rangle^2 = \mu N/6 + \mathcal{O}((\mu N)^2) \tag{4.47}$$

In free-recombining genomes, the intra-population diversity Δ of a trait with large number of loci ($\ell \gg 1$) has a Gaussian distribution (Central limit theorem); see Fig. (4.1).

$$P_0(\Delta;\langle\Gamma\rangle) = \frac{1}{Z_\Delta}\exp[\frac{-3(\Delta-\mu N\ell)^2}{\mu N\ell}] \tag{4.48}$$

with the following cross-population statistics,

$$\langle\Delta\rangle = \mu N\ell(1-4\mu N) + \mathcal{O}((\mu N)^3) \quad , \quad \langle\Delta^2\rangle - \langle\Delta\rangle^2 = \frac{\mu N\ell}{6} + \mathcal{O}((\mu N)^2) \tag{4.49}$$

4.2.4.2 Evolution of the free-recombining loci under selection

We have reduced the high-dimensional space of multi-locus allele frequencies to a 2-dimensional space of trait observables. In free-recombining genomes, allele frequencies in a low-mutation regime ($\mu N \ll 1$) reach an approximate equilibrium for an arbitrary number of loci. Clearly, the marginal stationary distributions of the macroscopic observables $P_0(\Gamma;\langle\Delta\rangle)$ and $P_0(\Delta;\langle\Gamma\rangle)$ also reach an equilibrium (de Vladar and Barton, 2011a). The evolutionary dynamics in a time-independent fitness landscape $f(\phi)$ of a gradient-form will have a simple relation to its neutral counterpart (Mustonen and Lässig, 2010),

$$Q(\Gamma;\langle\Delta\rangle) = \frac{1}{Z}\, P_0(\Gamma;\langle\Delta\rangle)\exp[2NF(\Gamma;\langle\Delta\rangle)] \tag{4.50}$$

$$Q(\Delta;\langle\Gamma\rangle) = \frac{1}{Z}\, P_0(\Delta;\langle\Gamma\rangle)\exp[2NF(\Delta;\langle\Gamma\rangle)] \tag{4.51}$$

where Z is the appropriate normalization factor. The Boltzmann factor that relates each of the two distributions is the rescaled mean population fitness $2NF(\Gamma_\phi,\Delta_\phi)$ projected on the corresponding plane in the $\Gamma-\Delta$ coordinate. The mean fitness (average growth rate) of a population with trait distribution $\rho(\phi)$ is,

$$F(\Gamma,\Delta) = \overline{f(\phi)} = \int d\phi\ \rho(\phi)f(\phi) \tag{4.52}$$

and the corresponding projections are obtained as, $F(\Gamma;\Delta) = \int d\Delta F(\Gamma,\Delta)Q(\Gamma,\Delta) \approx F(\Gamma;\Delta=\langle\Delta\rangle)$ and similarly, $F(\Delta;\Gamma=\langle\Gamma\rangle)$. This relation is a property of the Fokker-Planck equation in equilibrium; see similar relation in eq. (1.12). In a well-behaved fitness landscape $f(\phi)$, we can expand the fitness values around an arbitrary trait of interest $\phi^\star$,

$$f(\phi) - f(\phi^\star) = f'(\phi^\star)\ (\phi-\phi^\star) + \frac{1}{2}f''(\phi^\star)\ (\phi-\phi^\star)^2 + \dots \tag{4.53}$$

We will use eq. (4.53) to evaluate the mean population fitness in eq. (4.52) as a function of the intra-population trait statistics Γ and Δ. In this way, the derivatives of fitness function will be coupled to phenotype observables and hence become measurable in the population.

In this section, we will analyze two types of fitness landscapes, linear and quadratic landscapes. The first one is additive in loci, whereas the second one causes epistatic interactions between the loci.

Linear fitness (Directional selection). Directional selection on a trait ϕ applies as, $f_{lin}(\phi) = \alpha\ \phi$ with a non-zero slope α. The mean population fitness $F_{lin}(\Gamma)$ in this landscape has the form; see eq. (4.52),

$$\begin{aligned} F_{lin}(\Gamma; \langle\Delta\rangle) &= \int d\phi\ \rho(\phi) f_{lin}(\phi) \\ &= \alpha\ \Gamma \end{aligned} \tag{4.54}$$

The full distribution under selection takes the form,

$$Q_{lin}(\Gamma; \langle\Delta\rangle) = \frac{1}{Z}\ P_0(\Gamma; \langle\Delta\rangle)\ \mathrm{e}^{2N\alpha\Gamma} \quad , \quad Q_{lin}(\Delta; \langle\Gamma\rangle) = P_0(\Delta; \langle\Gamma\rangle) \tag{4.55}$$

Z is the appropriate normalization factor. Linear fitness only affects the cross-population mean of the trait average,

$$\begin{aligned} \langle\Gamma\rangle_s &= \langle\Gamma\rangle_0 + 2N\alpha\ \left[\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right] \\ &= \langle\Gamma\rangle_0 + \alpha N\ell\ (1 - 4\mu N)/2 \end{aligned} \tag{4.56}$$

$\langle\cdot\rangle_s$ refers to averages over the population ensemble that evolved under selection pressure and $\langle\cdot\rangle_0$ denotes the averages over the neutral ensemble.

Quadratic fitness. This type of fitness landscape, $f_{quad}(\phi) = \omega(\phi - \phi^\star)^2$ has non-zero slope and curvature and is the minimal fitness function that accommodates epistatic

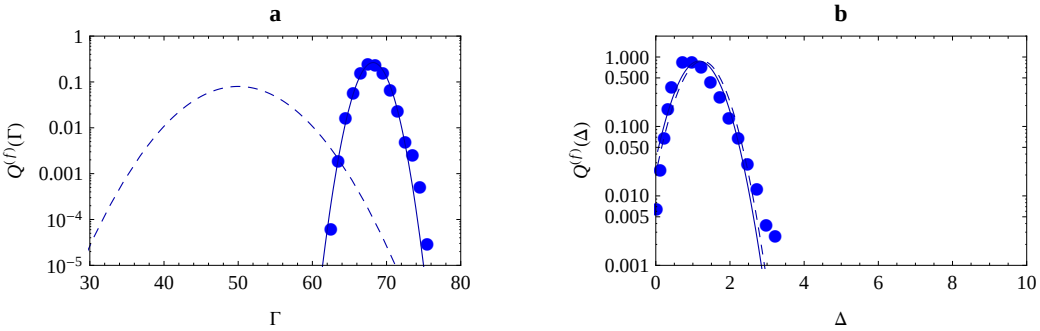

Figure 4.2: Equilibrium distribution of free-recombining traits under stabilizing selection. Quadratic selection, $f(\phi) = \omega(\phi - \phi^\star)^2$ affects both intra-population trait average Γ and trait diversity Δ. (a) The analytical estimate for the marginal distribution of the trait mean $Q^f(\Gamma)$ under stabilizing selection (full line) eq. (4.58) and the neutral expectation (dashed line) vs. the numerical result for simulated evolution in a quadratic fitness landscape (filled circles) of the free-recombining traits. (b) Same comparison as in (a) for the trait diversity $Q^f(\Delta)$ under stabilizing selection; eq. (4.59). The parameters are, $N = 200$, $\mu N = 0.0125$ and $\ell = 100$ bp with the rescaled fitness parameter, $2\omega N = -0.2$. The fitness peak $\phi^\star$ is set to 0.7ℓ. Quadratic fitness with negative curvature reduces the trait diversity and stabilizes the trait values in the population.

evolutionary interactions between loci. The mean population fitness in this landscape has the following dependencies,

$$\begin{aligned} F_{quad}(\Gamma, \Delta) &= \int d\phi\ \rho(\phi) f_{quad}(\phi) \\ &= \omega(\Gamma - \phi^\star)^2 + \omega\ \Delta \end{aligned} \tag{4.57}$$

The marginal distributions of the trait statistics under quadratic selection are,

$$Q_{quad}(\Gamma; \langle\Delta\rangle) = \frac{1}{Z}\ P_0(\Gamma; \langle\Delta\rangle) \exp[2N\omega(\Gamma - \phi^\star)^2] \tag{4.58}$$

$$Q_{quad}(\Delta; \langle\Gamma\rangle) = \frac{1}{Z}\ P_0(\Delta; \langle\Gamma\rangle) \exp[2N\omega\Delta] \tag{4.59}$$

Z is the appropriate normalization factor. Evolution in a quadratic fitness landscape modifies both distributions of trait average Γ and trait diversity Δ; see Fig. (4.2).

- *Trait average* ($\Gamma_0 \to \Gamma_s$)

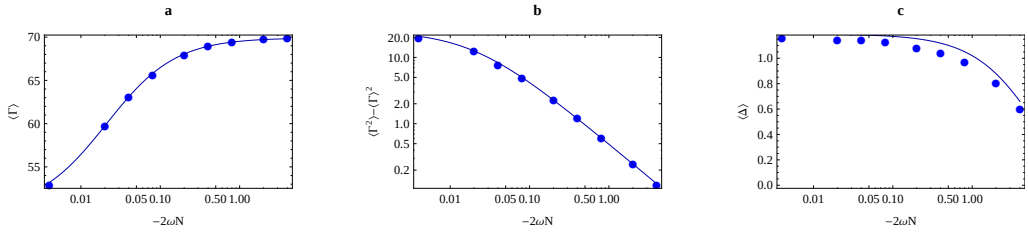

Figure 4.3: Effect of stabilizing selection on the statistics of free-recombining traits. Dependency of (a) cross-population mean of the trait average $\langle\Gamma\rangle$, (b) cross-population variance of the trait average (trait divergence) $\langle\Gamma^2\rangle - \langle\Gamma\rangle^2$ and (c) cross-population mean of the trait diversity $\langle\Delta\rangle$ on the strength of stabilizing selection $2\omega N$ in a quadratic fitness landscape. Stabilizing selection reduces the trait diversity in a population and the trait divergence across populations. Analytical estimates in eq. (4.60), eq. (4.61) and, eq. (4.62) (solid lines) are in good agreement with simulation results for evolution in a quadratic fitness landscape (full circles). Parameters are $N = 200$, $\ell = 100$ and $\mu N = 0.0125$ and the fitness peak $\phi^\star = 0.7\ell$.

$$\begin{aligned}\langle\Gamma\rangle_s &= \phi^\star - \frac{\phi^\star - \langle\Gamma\rangle_0}{1 - 4\omega N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right)} \\ &= \frac{\ell/2 - \omega N\ell\phi^\star}{1 - \omega N\ell}\end{aligned} \tag{4.60}$$

$$\begin{aligned}\langle\Gamma^2\rangle_s - \langle\Gamma\rangle_s^2 &= \frac{\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2}{1 - 4\omega N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right)} \\ &= \frac{\ell/4}{1 - \omega N\ell}\end{aligned} \tag{4.61}$$

Evolution in a quadratic fitness landscape with negative curvature $\omega < 0$ (i.e., stabilizing selection), reduces the spread of the intra-population trait average distribution.

- *Trait diversity* ($\Delta_0 \to \Delta_s$)

$$\begin{aligned}\langle\Delta\rangle_s &= \langle\Delta\rangle_0 + 2N\omega[\langle\Delta^2\rangle_0 - \langle\Delta\rangle_0^2] \\ &= \mu N\ell + \omega N\ \mu N\ell/3 + \mathcal{O}((\mu N)^2)\end{aligned} \tag{4.62}$$

$$\langle \Delta^2 \rangle_s - \langle \Delta \rangle_s^2 = \langle \Delta^2 \rangle_0 - \langle \Delta \rangle_0^2 \tag{4.63}$$

Evolution under stabilizing selection reduces the average intra-population trait diversity; see Fig. (4.3).

4.2.5 Phenotypic equilibrium of linked loci

In asexual reproduction, chromosomes resemble solid rods that carry the heritable information in a single package. The mutations occurring at different positions of a chromosome are physically linked and unlike the free-recombining case, their fates are bound to each other. This inevitable correlation which is termed "linkage-disequilibrium", then results in a less efficient role of natural selection and introduces significant constraints throughout adaptation (Desai, 2007; Desai and Fisher, 2007; Gerrish and Lenski, 1998; Park and Krug, 2007; Rouzine et al., 2008; Schiffels et al., 2011). Beneficial mutations in the genome cannot be fixed without carrying the rest of the genomic package including the coexisting deleterious mutations of other positions. Quantitative traits in linked genomes are also affected by these long-range correlations between their constituent loci. In this section, we address phenotypic evolution in asexual populations within the macroscopic framework described in the previous part.

4.2.5.1 Neutral evolution of linked loci

Genetic Drift. Genetic drift is the noise in the genotype and allele numbers due to the stochastic reproduction (or sampling) of the population at each generation. Unlike the case of the free-recombining traits, linkage introduces correlations between the sampling noise at different genomic positions. We denote the number of individuals in a population that carry the haplotype pair (α, β) at genomic positions (i, j) by $N_{i,j}^{\alpha,\beta}$. The discrete sampling noise then has the following properties,

$$\begin{aligned} \langle \xi_{i,j}^{\alpha,\beta}(t) \rangle &= 0 \\ \langle \xi_{i,j}^{\alpha,\beta}(t) \xi_{k,l}^{\nu,\gamma}(t') \rangle &= [\delta_{\alpha,\nu}\ \delta_{\beta,\gamma}\ \delta_{i,k}\ \delta_{j,l} + \delta_{\alpha,\gamma}\ \delta_{\beta,\nu}\ \delta_{i,l}\ \delta_{j,k}] \delta(t-t') N_{i,j}^{\alpha,\beta} \end{aligned} \tag{4.64}$$

The macroscopic observables Γ and Δ can be expressed as functions of marginal allele frequencies, y_i^α and haplotype frequencies of allele pairs, $y_{ij}^{\alpha\beta}$. As before, the lower Latin indices indicate the genomic positions and the upper Greek indices indicate the allele types,

$$y_{ij}^{\alpha\beta} = \frac{N_{ij}^{\alpha\beta}}{\sum_{\alpha,\beta} N_{ij}^{\alpha\beta}} \quad , \quad y_i^\alpha = \frac{\sum_{j,\beta} N_{ij}^{\alpha\beta}}{\sum_{k,\alpha,\beta} N_{ik}^{\alpha\beta}} \tag{4.65}$$

We then use the following relations to project the discrete sampling noise $\xi_{ij}^{\alpha\beta}$ onto the continuous frequency space,

$$\chi_{ij}^{\alpha\beta} = \sum_{\nu,\gamma} \frac{\partial y_{ij}^{\alpha\beta}}{\partial N_{ij}^{\nu\gamma}} \, \xi_{ij}^{\nu\gamma} \quad , \quad \chi_i^\alpha = \sum_{\nu,\gamma,k} \frac{\partial y_i^\alpha}{\partial N_{ik}^{\nu\gamma}} \, \xi_{ik}^{\nu\gamma} \tag{4.66}$$

The sampling noise for the continuous frequency variables has the following properties (use of eq. (4.64), eq. (4.66)),

$$\begin{aligned}
\langle \chi_{ij}^{\alpha\beta}(t) \rangle &= 0 \\
\langle \chi_i^\alpha(t) \rangle &= 0 \\
\langle \chi_{ij}^{\alpha\beta}(t) \chi_{i'j'}^{\alpha'\beta'}(t') \rangle &= \frac{1}{N} \, [\delta_{i,i'}\delta_{j,j'} y_{ij}^{\alpha\beta}(\delta_{\alpha\alpha'}\delta_{\beta\beta'} - y_{ij}^{\alpha'\beta'}) + \delta_{i,j'}\delta_{ji'} y_{ij}^{\alpha\beta}(\delta_{\beta,\alpha'}\delta_{\alpha,\beta'} - y_{ji}^{\alpha'\beta'})]\delta(t-t') \\
\langle \chi_i^\alpha(t) \chi_j^\beta(t') \rangle &= \frac{1}{N} \, [\delta_{i,j} y_i^\alpha(\delta_{\alpha\beta} - y_i^\beta) + (1-\delta_{i,j})(y_{ij}^{\alpha\beta} - y_i^\alpha y_j^\beta)]\delta(t-t') \\
\langle \chi_{ij}^{\alpha\beta}(t) \chi_{i'}^{\alpha'}(t') \rangle &= \frac{1}{N} \, y_{ij}^{\alpha\beta}[\delta_{i,i'}(\delta_{\alpha,\alpha'} - y_i^{\alpha'}) + \delta_{j,i'}(\delta_{\beta,\alpha'} - y_i^{\alpha'})]\delta(t-t')
\end{aligned} \tag{4.67}$$

We can now similarly compute the noise terms for the independent macroscopic trait variables, $\Gamma(y_i) = \sum_i y_i$ and, $\Delta(y_i, y_{ij}) = \sum_i y_i(1-y_i) + \sum_{i \neq j} y_{ij} - y_i y_j$. To simplify our notation, we denote the marginal frequencies of allele "1" and allele pair (1,1) by $y_i = y_i^1$ and $y_{ij} = y_{ij}^{11}$.

$$\begin{aligned} \chi_\Gamma &= \sum_i \frac{\partial \Gamma}{\partial y_i}\chi_i \\ \chi_\Delta &= \sum_i \frac{\partial \Delta}{\partial y_i} + \sum_{i\neq j} \frac{\partial \Delta}{\partial y_{ij}} \end{aligned} \tag{4.68}$$

which then results in,

$$\begin{aligned} \langle \chi_\Gamma(t)\rangle = 0, \qquad & \langle \chi_\Gamma(t)\chi_\Gamma(t')\rangle = \frac{1}{N}\Delta\ \delta(t-t') \\ \langle \chi_\Delta(t)\rangle = 0, \qquad & \langle \chi_\Delta(t)\chi_\Delta(t')\rangle \approx \frac{1}{N}2\Delta^2\ \delta(t-t') \\ & \langle \chi_\Gamma(t)\chi_\Delta(t')\rangle \approx 0 \end{aligned} \tag{4.69}$$

The approximate closed form solutions in eq. (4.69) are truncated at two-point correlations of the trait distribution $\rho(\phi)$.

Mutations. At the level of an individual, mutations are stochastic events often coupled to reproduction and change the allele $1 \to 0$ or vice versa. As before, we assume that mutations occur with a uniform rate μ at all nucleotide positions. The macroscopic trait statistics Γ, Δ in a linked genome are functions of both marginal allele frequencies y_i and haplotype frequencies of the allele pairs y_{ij}. Therefore, we need to quantify the effect of mutations on both of these marginals. As computed for free-recombining loci, the change due to mutations in the number of individuals that carry allele "1" (N_i^1) is,

$$\delta N_i^1 = \mu(N_i^0 - N_i^1) = \mu(N - 2N_i^1) \tag{4.70}$$

Similarly, the mutational effect on the number of individuals that carry allele pair "1-1" (N_{ij}^{11}) is,

$$\begin{aligned} \delta N_{ij}^{11} &= \mu[N_{ij}^{10} + N_{ij}^{01} - 2N_{ij}^{11}] + \mathcal{O}(\mu^2) \\ &= \mu[N_i^1 + N_j^1 - 4N_{ij}^{11}] \end{aligned} \tag{4.71}$$

We can then compute the corresponding marginal frequency changes, $\delta y_i = \delta N_i^1/N$ and, $\delta y_{ij} = \delta N_{ij}^{11}/N$. The stochastic temporal changes of the marginal frequencies in a neutral evolution are,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} y_i &= \mu(1-2y_i) + \chi_i & (4.72) \\ \frac{\mathrm{d}}{\mathrm{d}t} y_{ij} &= \mu(y_i + y_j - 4y_{ij}) + \chi_{ij} & (4.73) \end{aligned}$$

We project the neutral changes of the allele frequencies y_i and y_{ij} onto the macroscopic trait observables, Γ and Δ,

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}t} = \sum_i \frac{\mathrm{d}y_i}{\mathrm{d}t} = -2\mu(\Gamma - \ell/2) + \chi_\Gamma \tag{4.74}$$

and,

$$\begin{aligned} \frac{\mathrm{d}\Delta}{dt} &= \sum_i \frac{\mathrm{d}}{\mathrm{d}t}(y_i - y_i^2) + \sum_{i\neq j} \frac{\mathrm{d}}{\mathrm{d}t}(y_{ij} - y_i y_j) \\ &= -4\mu \, [\sum_i (y_i(1-y_i) - \ell/4] - \frac{\Delta}{N} - 4\mu \sum_{i\neq j} (y_{ij} - y_i y_j) + \chi_\Delta \\ &= -4\mu(\Delta - \ell/4) - \frac{\Delta}{N} + \chi_\Delta & (4.75) \end{aligned}$$

The properties of the noise terms, χ_Γ and χ_Δ are given in eq. (4.69). The term Δ/N in eq. (4.75) appears due to the nonlinear dependency of the trait variance on marginal frequency y_i; see e.g., the discussion on Ito Calculus in Chapter. 4 of (Gardiner, 2004). The temporal changes of the trait statistics in eq. (4.74) and eq. (4.75) together with noise covariance relations in eq. (4.69) can be used to derive the corresponding Fokker-Planck equation for the neutral probability distribution $P_0(\Gamma, \Delta, t)$; see eq. (4.29).

We can write the two Langevin equations, eq. (4.74), eq. (4.75), in the form of a multi-variant Ornstein-Uhlenbeck process,

$$\mathrm{d}\Omega = A(\Omega, t)\mathrm{d}t + B(\Omega, t)\mathrm{d}W(t) \tag{4.76}$$

where $\Omega^{\mathbf{T}} = (\Gamma, \Delta)$ is a vector of the state variables, A is a vector denoting the deterministic dynamics, B is a 2×2 matrix and $\mathrm{d}W(t)$ is a 2-dimensional Wiener process associated with the noise terms, χ_Γ and χ_Δ. Using eq. (4.69), eq. (4.74) and eq. (4.75),

$$A = -2\mu \begin{pmatrix} \Gamma - \frac{\ell}{2} \\ 2\Delta - \frac{\ell}{2} + \frac{\Delta}{2\mu N} \end{pmatrix}, \qquad B = \begin{pmatrix} \sqrt{\Delta/N} & 0 \\ 0 & \sqrt{2\Delta^2/N} \end{pmatrix} \tag{4.77}$$

The corresponding Fokker-Planck equation for the probability density $p(\Omega, t)$ can be derived in the following way,

$$\partial_t p(\Omega, t) = -\sum_i \partial_i A_i(\Omega, t) p(\Omega, t) + \frac{1}{2} \sum_{i,j} \partial_i \partial_j [B(\Omega, t) B^{\mathbf{T}}(\Omega, t)]_{ij} p(\Omega, t) \tag{4.78}$$

which then results in

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} P_0(\Gamma, \Delta, t) &= \frac{1}{2N} \left[\frac{\partial^2}{\partial \Gamma^2} \Delta + 2 \frac{\partial^2}{\partial \Delta^2} \Delta^2 \right] P_0(\Gamma, \Delta, t) \\ &\quad + 2\mu \left[\frac{\partial}{\partial \Gamma} (\Gamma - \ell/2) + 2 \frac{\partial}{\partial \Delta} (\Delta - \ell/4 + \Delta/4\mu N) \right] P_0(\Gamma, \Delta, t) \end{aligned} \tag{4.79}$$

To emphasize one more time, $P_0(\Gamma, \Delta, t)$ is the distribution of intra-population trait statistics over different population realizations, and should be distinguished from the trait distribution in a single population, $\rho(\phi)$.

We can analytically compute the stationary solutions of eq. (4.79) for the marginal distributions, $P_0(\Gamma; \langle\Delta\rangle)$ and $P_0(\Delta; \langle\Gamma\rangle)$.

1. *Trait average,* Γ

   The marginal stationary probability distribution for the intra-population trait average is,

   $$P_0(\Gamma; \langle\Delta\rangle) = \frac{1}{Z_\Gamma} \exp[\frac{-2(\Gamma - \ell/2)^2}{\ell(1 - 4\mu N)}] \tag{4.80}$$

   with $Z_\Gamma \approx \sqrt{\pi(1 - 4\mu N)\ell/2}$ as the normalization factor. We obtain the stationary solution of the marginal distribution $P_0(\Gamma; \langle\Delta\rangle)$ by using the result of the

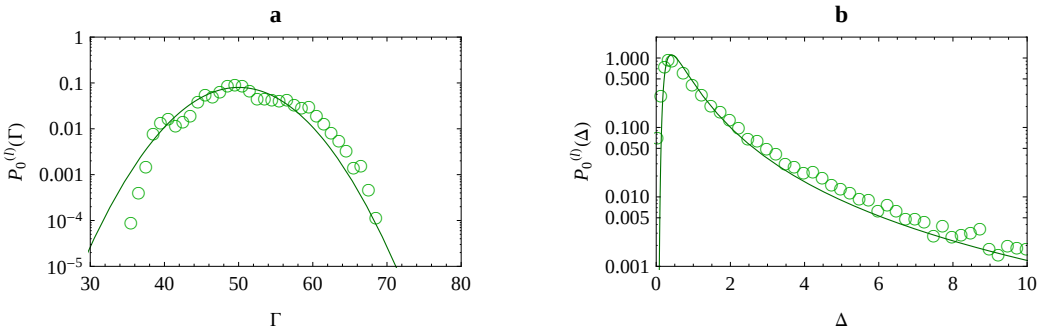

Figure 4.4: Equilibrium distribution of linked traits in neutrality. (a) The analytical estimate for the marginal distribution of intra-population trait average $P_0^{(l)}(\Gamma)$ (full line) vs. the numerical result for simulated evolution (open circles) of linked traits; eq. ((4.80). (b) Same comparison as in (a) for the trait diversity observable Δ; eq. (4.82). The Fokker-Planck equation in eq. (4.79) can accurately describe the stationary state the macroscopic trait observable in linked genomes. The simulation parameters are chosen: $N = 200$, $\ell = 100$ bp and $\mu N = 0.0125$.

following eq. (4.84) to integrate over the intra-population trait diversity. The marginal distribution $P_0(\Gamma; \langle\Delta\rangle)$ is in agreement with numerical results for simulated evolution of a linked genome; see Fig. 4.4(a). Simulation are base on the forward Wright-Fisher process similar to what we discussed in the Materials and Method of Section. 3.7.

Cross-population mean and variance of the trait average Γ are,

$$\langle\Gamma\rangle = \ell/2 \qquad \langle\Gamma^2\rangle - \langle\Gamma\rangle^2 = \ell(1 - 4\mu N)/4 \tag{4.81}$$

Comparing the Γ statistic under linkage to the free-recombining case in eq. (4.41), shows that linkage properties do not influence the statistics of the linear macroscopic observable, Γ.

2. *Trait diversity,* Δ

The marginal stationary probability distribution for the intra-population trait diversity is,

$$P_0(\Delta; \langle\Gamma\rangle) = \frac{1}{Z_\Delta} \Delta^{-3-4\mu N} \exp\left[-\frac{\mu N \ell}{\Delta}\right] \tag{4.82}$$

with the normalization factor,

$$Z_\Delta = (\mu N \ell)^{-2-4\mu N}\Gamma[2+4\mu N] \tag{4.83}$$

The marginal distribution $P_0(\Delta;\langle\Gamma\rangle)$ is in agreement with numerical results for simulated evolution of a linked genome; see Fig. 4.4(b). Cross-population mean and variance of the intra-population trait diversity Δ are,

$$\langle\Delta\rangle = \mu N\ell(1-4\mu N)+\mathcal{O}((\mu N)^3) \quad , \quad \langle\Delta^2\rangle-\langle\Delta\rangle^2 = \mu N\ell^2(1/4-2\mu N)+\mathcal{O}((\mu N)^3) \tag{4.84}$$

The distribution of the intra-population trait diversity under linkage grows linearly with the genome length, whereas for the free-recombining case, it grows sub-linearly as $\ell^{1/2}$. This is due to the long-range correlations between the loci under linkage which is not present between the free-recombining loci.

4.2.5.2 Evolution of linked loci under selection

Long-range correlations in a linked polygenic trait cause complicated microscopic dynamics at the level of individual loci. Detailed balance is not satisfied in the stationary description of the high-dimensional allele frequency vector $\mathbf{x}_\alpha$. The extensive self-averaging of the microscopic complexities however, can result in a tractable macroscopic picture. Here, we have reduced the high-dimensional space of multi-locus allele frequencies onto a 2-dimensional space of trait observables. Nonetheless, the joint stationary description for these macroscopic observables in eq. (4.79) can still have a non-vanishing current $\mathbf{J}P(\Gamma,\Delta)$ on a 2-dimensional simplex. However, the marginal stationary distributions $P(\Gamma;\langle\Delta\rangle)$ and $P(\Delta;\langle\Gamma\rangle)$, which are the 1-dimensional projection of the joint distribution, satisfy detailed balance.

Given neutral equilibrium at the level of the marginal distributions, evolution in an arbitrary time-independent fitness landscape $f(\phi)$ of a gradient-form follows, (Mustonen and Lässig, 2010),

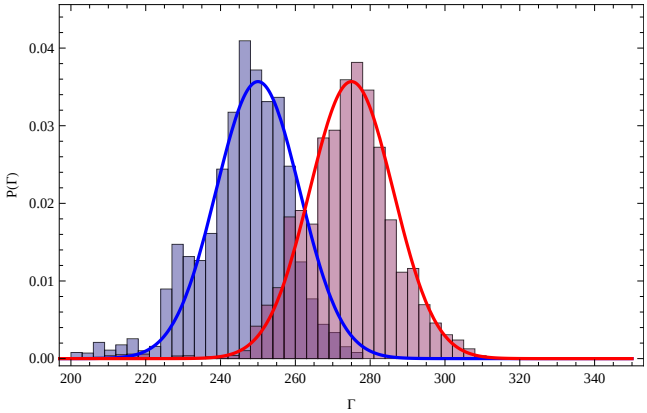

Figure 4.5: Equilibrium distribution of linked traits under directional selection. Directional selection $f(\phi) = \alpha\phi$ only affects the intra-population trait average Γ. We compare the stationary distribution of trait average in neutrality $P_0(\Gamma)$ (in blue) to the distribution under directional selection $Q_{lin}(\Gamma)$ (in red). The analytical estimates of eq. (4.87) (solid curves) are in agreement with simulation results from the evolution of linked loci in linear fitness landscape and in neutral conditions (bar histograms). The parameters are $N = 500$, $\mu N = 0.0125$ and $\ell = 1000$ bp. The rescaled slope of the fitness landscape in red histogram is, $N\alpha = 0.1$. Directional selection shifts the Γ distribution, but does not change its width.

$$Q(\Gamma; \langle\Delta\rangle) = \frac{1}{Z}\, P_0(\Gamma; \langle\Delta\rangle)\, \mathrm{e}^{2NF(\Gamma;\langle\Delta\rangle)} \tag{4.85}$$

$$Q(\Delta; \langle\Gamma\rangle) = \frac{1}{Z} P_0(\Delta; \langle\Gamma\rangle)\, \mathrm{e}^{2NF(\Delta;\langle\Gamma\rangle)} \tag{4.86}$$

where Z is the appropriate normalization factor. The Boltzmann factor that relates each of the two distributions is the rescaled mean population fitness $2NF(\Gamma_\phi, \Delta_\phi)$ (eq. (4.52)) projected on the corresponding plane in the $\Gamma - \Delta$ coordinate. Similar to the analysis of free-recombining loci in Section. 4.2.4.2, we will characterize the evolutionary dynamics of the linked genome in two types of fitness landscapes: (i) linear fitness which is additive in loci and (ii) quadratic fitness which is the minimal landscape that allows epistatic interactions.

Linear Fitness (Directional selection). Directional selection on trait ϕ acts as $f_{lin}(\phi) = \alpha\ \phi$ with a non-zero slope α. The mean population fitness in this landscape has the form, $F_{lin}(\Gamma) = \alpha\ \Gamma$; see eq. (4.54). This type of selection only affects the

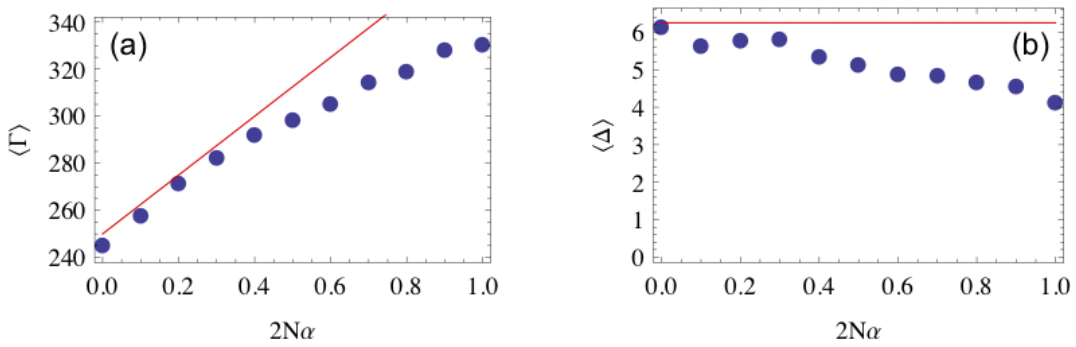

Figure 4.6: Effect of directional selection on the trait statistics of linked loci. Dependency of the mean estimates for (a) intra-population trait average $\langle\Gamma\rangle$ and (b) intra-population trait diversity $\langle\Delta\rangle$ on the strength of directional selection α. Directional selection changes the mean of the trait average but does not influence the trait diversity. The theoretical estimates (solid lines) are in good agreement with simulation results for evolution of a linked genome in a linear fitness landscape (blue dots) in the regime of weak to moderate selection. The deviation between numerical and analytical results is due to the limited number of available loci in a finite genome that can satisfy the high selection criteria. This separation gets smaller with increasing genome size ℓ. Parameters are $N = 100$, $\ell = 500$ and $2N\mu = 0.025$.

cross-population mean of the trait average; see eq. (4.55). The marginal distribution under selection follows,

$$\begin{aligned} Q_{lin}(\Gamma;\langle\Delta\rangle) &= \frac{1}{Z}\, P_0(\Gamma;\langle\Delta\rangle)\mathrm{e}^{2N\alpha\Gamma} \\ &= \frac{1}{Z}\exp\left[-\frac{2\,(\Gamma-\ell/2)^2}{\ell(1-4\mu N)}+2N\alpha\Gamma\right] \end{aligned} \tag{4.87}$$

Z is the appropriate normalization factor. In Fig. 4.5 we compare the marginal distribution $Q_{lin}(\Gamma;\langle\Delta\rangle)$ in eq. (4.87) to the numerical results for simulated evolution of a linked genome in a linear fitness landscape. The analytical predictions are in perfect agreement with the numerical results. As shown in Fig. 4.5 and in eq. (4.87), the mean phenotype under selection is normally distributed with a similar variance to that of the neutral evolution but with a shifted average $\langle\Gamma\rangle_s$,

$$\begin{aligned}\langle\Gamma\rangle_s &= \langle\Gamma\rangle_0 + \frac{\alpha N}{2}[\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2] \\ &= \ell/2 + \alpha N\ell\ (1-4\mu N)/2 \end{aligned} \tag{4.88}$$

As before, $\langle\cdot\rangle_s$ refers to averages over the population ensemble that evolved under selection pressure and $\langle\cdot\rangle_0$ denotes the averages over the neutral ensemble. The higher moments of the Γ distribution remain unchanged in the linear landscape. Directional selection affects the statistics of the intra-population trait average similarly in both linked and free-recombining genomes. Trait diversity Δ is also not affected by directional selection; see Fig. 4.6 for comparisons with numerical simulations.

Quadratic fitness. This type of fitness landscape, $f_{quad}(\phi) = \omega(\phi - \phi^\star)^2$ has non-zero slope, $2\omega(\phi - \phi^\star)$ and curvature, 2ω. The mean population fitness in a quadratic landscape has the form, $F_{quad}(\Gamma, \Delta) = \omega(\Gamma - \phi^\star)^2 + \omega\Delta$. The marginal distributions of intra-population trait average and trait diversity under quadratic selection follow,

$$\begin{aligned} Q_{quad}(\Gamma; \langle\Delta\rangle) &= \frac{1}{Z}\ P_0(\Gamma; \langle\Delta\rangle)\ \mathrm{e}^{2N\omega(\Gamma-\phi^\star)^2} \\ &= \frac{1}{Z}\ \exp\left[-\frac{2\ (\Gamma - \ell/2)^2}{\ell(1-4\mu N)} + 2N\omega(\Gamma-\phi^\star)^2\right] \end{aligned} \tag{4.89}$$

$$\begin{aligned} Q_{quad}(\Delta; \langle\Gamma\rangle) &= \frac{1}{Z}\ P_0(\Delta; \langle\Gamma\rangle)\ \mathrm{e}^{2N\omega\Delta} \\ &= \frac{1}{Z}\ \Delta^{-3-4\mu N} \exp\left[-\frac{\mu N\ell}{\Delta} + 2N\omega\Delta\right] \end{aligned} \tag{4.90}$$

Z is the appropriate normalization factor. Fig. 4.7 compares the analytical estimates of eq. (4.89) and eq. (4.90) to the numerical results for simulated evolution of a linked genome in a quadratic fitness landscape. Quadratic selection affects statistics of both intra-population trait average Γ and trait diversity Δ,

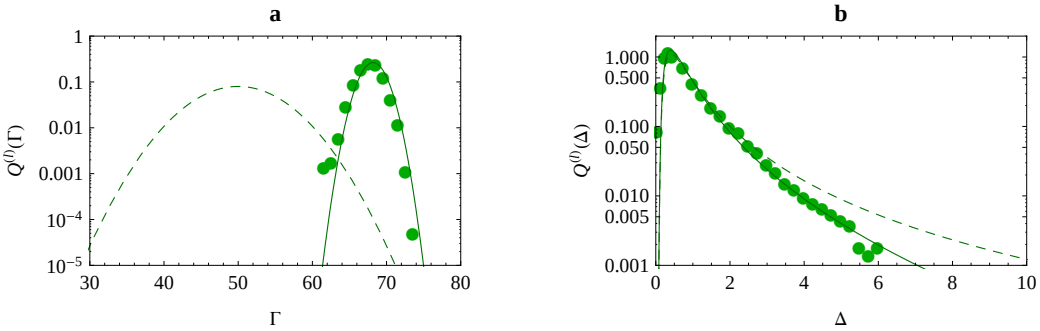

Figure 4.7: Equilibrium distribution of linked traits under stabilizing selection. Quadratic selection, $f(\phi) = \omega(\phi - \phi^\star)^2$ affects both intra-population trait average Γ and trait diversity Δ. (a) The analytical estimate for the marginal distribution of the trait mean $Q^{(l)}(\Gamma)$ under stabilizing selection (full line) and the neutral expectation (dashed line) vs. the numerical result for simulated evolution in a quadratic fitness landscape (filled circles) of linked traits. The simulation results are in excellent agreement with the analytical prediction in eq. (4.89). (b) Same comparison as in (a) for the trait diversity $Q^{(l)}(\Delta)$ under stabilizing selection; eq. 4.90). The parameters are, $N = 200$, $\mu N = 0.0125$ and $\ell = 100$ bp with the rescaled fitness parameter, $2\omega N = -0.2$. The fitness peak $\phi^\star$ is set to 0.7ℓ. Quadratic fitness with negative curvature reduces the trait diversity and stabilizes the trait values in the population.

- *Trait average* ($\Gamma_0 \to \Gamma_s$)

$$\begin{aligned} \langle\Gamma\rangle_s &= \phi^\star - \frac{\phi^\star - \langle\Gamma\rangle_0}{1 - 4\omega N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right)} \\ &= \frac{\ell/2 - \omega N\ell\phi^\star}{1 - \omega N\ell} \end{aligned} \tag{4.91}$$

$$\begin{aligned} \langle\Gamma^2\rangle_s - \langle\Gamma\rangle_s^2 &= \frac{\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2}{1 - 4\omega N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right)} \\ &= \frac{\ell/4}{1 - \omega N\ell} \end{aligned} \tag{4.92}$$

Evolution in a quadratic fitness landscape with negative curvature $\omega < 0$ (i.e., stabilizing selection), reduces the spread of the intra-population trait average distribution. Fig. 4.8 shows the dependency of Γ statistics on the fitness parameter ω and compares the analytical estimates to the simulation results for the evolu-

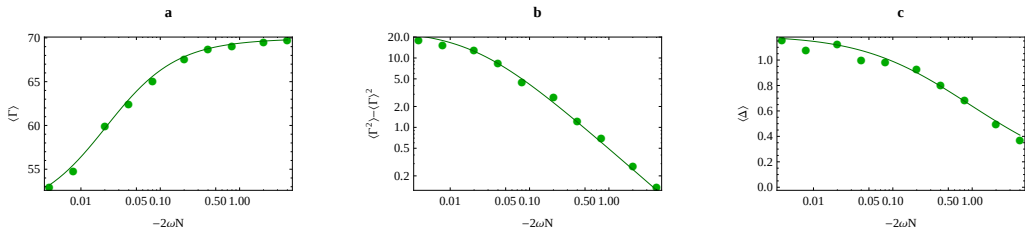

Figure 4.8: Effect of stabilizing selection on the trait statistics of linked traits. Dependency of (a) cross-population mean of the trait average $\langle\Gamma\rangle$, (b) cross-population variance of the trait average (trait divergence) $\langle\Gamma^2\rangle - \langle\Gamma\rangle^2$ and (c) cross-population mean of the trait diversity $\langle\Delta\rangle$ on the strength of stabilizing selection $2\omega N$ in a quadratic fitness landscape. Stabilizing selection reduces both trait diversity in the population and divergence across populations. Analytical predictions in eq. (4.91), eq. (4.92) and, eq. (4.93) (solid lines) are in good agreement with simulation results for evolution in a quadratic fitness landscape (full circles). Parameters are $N = 200$, $\ell = 100$ and $2N\mu = 0.025$ and the fitness peak $\phi^\star = 0.7\ell$.

tionary dynamics of the linked loci.

- *Trait diversity* ($\Delta_0 \to \Delta_s$)

$$\langle\Delta\rangle_s = \sqrt{\frac{-\mu N\ell}{2\omega N}}\ \frac{k_{1+4\mu N}[\sqrt{(-8\mu N\ell\ \omega N)}]}{k_{2+4\mu N}[\sqrt{(-8\mu N\ell\ \omega N)}]} \tag{4.93}$$

$$\langle\Delta^2\rangle_s - \langle\Delta\rangle_s^2 = \frac{-\mu N\ell}{2\omega N}\left[\frac{k_{4\mu N}[\sqrt{-8\mu N\ell\ \omega N}]}{k_{2+4\mu N}[\sqrt{-8\mu N\ell\ \omega N}]] } - \left(\frac{k_{1+4\mu N}[\sqrt{(-8\mu N\ell\ \omega N)}]}{k_{2+4\mu N}[\sqrt{(-8\mu N\ell\ \omega N)}]}\right)^2\right] \tag{4.94}$$

where $k_n(z) = \text{BesselK}[n, z]$ is the modified Bessel function of the second kind which satisfies the differential equation, $z^2y'' + zy' - (z^2 + n^2)y = 0$. In the regime of weak selection ($|\omega N| \ll 1$), average diversity under selection $\langle\Delta\rangle_s$ in eq. (4.93) simplifies,

$$\begin{aligned} \langle\Delta\rangle_s &= \langle\Delta\rangle_0 + 2\omega N\langle\Delta\rangle_0^2 + \mathcal{O}[(\mu N\ell)^3(\omega N)^2] \\ &\approx \mu N\ell + 2(\mu N\ell)^2\omega N \end{aligned} \tag{4.95}$$

Fig. 4.8 compares the analytical estimates for statistics of Δ to the numerical results for simulated evolution of a linked genome in a quadratic fitness landscape. Quadratic selection with negative curvature, sharpens the Γ distribution, $P_0(\Gamma)$ and shortens the long-tail of the Δ distribution $P_0(\Delta)$, i.e., the trait composition in the population is stabilized. The exact analytical derivations in this section couple the characteristics of the fitness landscape to the statistics of the trait distribution. In this way, we can infer the shape of the fitness landscape through measurements of the macroscopic trait observables.

4.3 Inference of selection strength from phenotypic polymorphism

In this part, we present the most practical result of this section: inference of fitness parameters by comparing the statistics of the phenotypic polymorphisms to the neutral expectations. Our effort was to quantify the neutral dynamics of the linked genome despite the mathematical difficulties that arise from the microscopic details of multi-loci statistics. We then study the effect of selection on these traits and characterize the modified statistics of the macroscopic population observable, trait average Γ and diversity Δ. We distinguish between intra- and inter- population statistics and introduce a framework through which the phenotypic trait in neutrality and under the influence of natural selection are related. Using the results in Section. 4.2.4.2 for free-recombining genome and in Section. 4.2.5.2 for asexual populations, we can now infer the shape of the fitness landscape on which the evolutionary dynamics has been directed.

4.3.1 Free-recombining genome

Although the free-recombining model does not address the genomic composition of actual biological systems, it is still a useful approximation for traits with loci located far apart or on different chromosomes. The two types of analytical fitness landscapes that we analyzed in Section. 4.2.4.2 are coupled to the trait observables and thus measurable in the following way,

Directional Selection, $(f(\phi) = \alpha\ \phi)$

In a free-recombining genome, directional selection only affects the cross-population mean trait distribution by shifting its average. In this way, we can infer the rescaled slope of the fitness landscape $N\alpha$ from eq. (4.56),

$$N\alpha = \frac{1}{2}\ \frac{\langle\Gamma\rangle_s - \langle\Gamma\rangle_0}{\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2} \tag{4.96}$$

Quadratic Selection, $(f(\phi) = \omega\ (\phi - \phi^\star)^2)$

Evolution in a quadratic fitness landscape affects the statistics of intra-population trait average and trait diversity . The rescaled curvature of the fitness function $N\mathcal{C} = 2\omega N$ is coupled to both trait average Γ and trait diversity Δ and can be inferred in two ways. From the analysis of Γ-statistics in eq. (4.61),

$$N\mathcal{C} = 2N\omega = \frac{1}{2}\ \frac{(\mathbf{var}\ \Gamma)_s - (\mathbf{var}\ \Gamma)_0}{(\mathbf{var}\ \Gamma)_s\ (\mathbf{var}\ \Gamma)_0} \tag{4.97}$$

with $\mathbf{var}\ \Gamma = \langle\Gamma^2\rangle - \langle\Gamma\rangle^2$, is the variance of the trait average in the population ensemble under selection $(\cdot)_s$ or in neutrality $(\cdot)_0$. Similarly, from the analysis of Δ-statistics in eq. (4.62),

$$N\mathcal{C} = \frac{\langle\Delta\rangle_s - \langle\Delta\rangle_0}{\langle\Delta^2\rangle_0 - \langle\Delta\rangle_0^2} \tag{4.98}$$

and the location of the fitness peak $\phi^\star$ can be inferred from eq. (4.60),

$$\frac{\phi^\star - \langle\Gamma\rangle_s}{\phi^\star - \langle\Gamma\rangle_0} = 1 - 2\mathcal{C}N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right) \tag{4.99}$$

4.3.2 Fully linked genome

Directional Selection, $(f(\phi) = \alpha\ \phi)$

Due to the additivity of phenotype ϕ, directional selection affects the linked genome in the same way that it affects the free-recombining genome. We can infer the rescaled slope of the fitness landscape $N\alpha$ from eq. (4.88),

$$N\alpha = \frac{1}{2}\ \frac{\langle\Gamma\rangle_s - \langle\Gamma\rangle_0}{\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2} \tag{4.100}$$

Quadratic Selection, $(f(\phi) = \omega\ (\phi - \phi^\star)^2)$

Similar to the free-recombining case, the parameters of a quadratic fitness landscape are coupled to both Γ-statistics and Δ-statistics. The change in the spread of the Γ-distribution yields (eq. (4.92)),

$$N\mathcal{C} = 2N\omega = \frac{1}{2}\ \frac{(\mathbf{var}\ \Gamma)_s - (\mathbf{var}\ \Gamma)_0}{(\mathbf{var}\ \Gamma)_s\ (\mathbf{var}\ \Gamma)_0} \tag{4.101}$$

The curvature of the quadratic fitness function is coupled to the cross-population average of the phenotype diversity $\langle\Delta\rangle$ which in asexual population is given by combinations of Bessel functions in eq. (4.93). In the regime of weak selection ($\omega N \ll 1$), this function simplifies and we can infer the rescaled curvature of the fitness landscape from eq. (4.95),

$$N\mathcal{C} = 2N\omega = \frac{\langle\Delta\rangle_s - \langle\Delta\rangle_0}{\langle\Delta\rangle_0^2} \tag{4.102}$$

and the location of the fitness peak $\phi^\star$ follows from eq. (4.91),

$$\frac{\phi^\star - \langle\Gamma\rangle_s}{\phi^\star - \langle\Gamma\rangle_0} = 1 - 2\mathcal{C}N\left(\langle\Gamma^2\rangle_0 - \langle\Gamma\rangle_0^2\right) \tag{4.103}$$

One can immediately realize that relations to infer the shape of the fitness landscape in free-recombining genome in Section. 4.3.1 are mostly identical to that of the linked genome in Section. 4.3.1. It should not however be forgotten that the trait statistics are different in these two cases and hence, different selection pressures are required to reduce the phenotypic diversity of the linked and the free-recombining genomes by the same amount. These estimates allow us to infer the shape of the time-independent fitness landscape, by comparing the distributions of the phenotype statistics to the neutral expectations and hence serve as a valuable means for empirical and genomic analysis; see the Discussion on genomic applications of the polygenic analysis in Section. 4.4.

4.4 Discussion & Outlook

Evolution of polygenic traits in linked and free-recombining genomes

In this chapter, we study the phenotypic evolution of polygenic traits in both free-recombining (with perfect reassortment) and linked (asexual) genomes. We characterize the phenotypic composition of a population by its intra-population statistics, trait average Γ and trait diversity Δ. In this way, we map the microscopic locus-based statistics of the trait onto the macroscopic population observables. This description proves to be very essential in eliminating the small-scale microscopic complications, yet it is informative enough to explain standing variation in populations. We estimate the likelihood of the trait statistics Γ and Δ for a linear phenotype which is comprised of ℓ nucleotide loci with equal phenotypic contributions. We first characterize the stochastic neutral evolution of the trait statistics and evaluate the stationary marginal distributions for trait average Γ, $P_0(\Gamma; \langle\Delta\rangle)$ and for trait diversity Δ, $P_0(\Delta; \langle\Gamma\rangle)$ under mutation-drift balance. These results are summarized in Table. 4.1. The main difference between the free-recombining and linked loci lies in their distribution of trait diversity $P_0(\Delta; \langle\Gamma\rangle)$. Long-range correlations in linked genome results in a broad distribution of the trait diversity $P_0(\Delta; \langle\Gamma\rangle)$ with a power-law tail. In a free-recombining genome however, the Δ distribution is Gaussian as expected from central limit theorem. This result has further consequences especially in the presence of selection.

We then characterize the distribution of the macroscopic trait observables in populations which have evolved in two types of fitness landscapes: Linear (directional selection) and quadratic fitness landscape. The statistics of Γ and Δ can be computed

	$P(\Gamma)$	$P(\Delta)$	$\langle\Gamma\rangle$	$\langle\Delta\rangle$	$\mathbf{var}\Gamma$	$\mathbf{var}\Delta$
Neutrality Free Recom.	$\exp[\frac{-2(\Gamma-\ell/2)^2}{\ell}]$	$\exp[\frac{-3(\Delta-\mu N\ell)^2}{\mu N\ell}]$	$\ell/2$	$\mu N\ell(1-4\mu N)$	$\ell/4$	$\mu N\ell/6$
Sel.-Lin. Free Recom.	$P_0^{(f)}(\Gamma)\ \mathrm{e}^{2N\alpha\Gamma}$	$P_0^{(f)}(\Delta)$	$\ell/2+\alpha N\ell/2$	$\mu N\ell(1-4\mu N)$	$\ell/4$	$\mu N\ell/6$
Sel.-Quad. Free Recom.	$P_0^{(f)}(\Gamma)\ \mathrm{e}^{2N\omega(\Gamma-\phi^\star)^2}$	$P_0^{(f)}(\Delta)\ \mathrm{e}^{2N\omega\Delta}$	$\frac{\ell/2-\omega N\ell\phi^\star}{1-\omega N\ell}$	$\mu N\ell(1+\omega N/3)$	$\frac{\ell/4}{1-\omega N\ell}$	$\mu N\ell/6$
Neutrality Full Linkage	$\exp[\frac{-2(\Gamma-\ell/2)^2}{\ell(1-4\mu N)}]$	$\Delta^{-3-4\mu N}\mathrm{e}^{\frac{-\mu N\ell}{\Delta}}$	$\ell/2$	$\mu N\ell(1-4\mu N)$	$\ell/4$	$\mu N\ell^2/4$
Sel.-Lin. Full Linkage	$P_0^{(l)}(\Gamma)\ \mathrm{e}^{2N\alpha\Gamma}$	$P_0^{(l)}(\Delta)$	$\ell/2+\alpha N\ell/2$	$\mu N\ell(1-4\mu N)$	$\ell/4$	$\mu N\ell^2/4$
Sel.-Quad. Full Linkage	$P_0^{(l)}(\Gamma)\ \mathrm{e}^{2N\omega(\Gamma-\phi^\star)^2}$	$P_0^{(l)}(\Delta)\ \mathrm{e}^{2N\omega\Delta}$	$\frac{\ell/2-\omega N\ell\phi^\star}{1-\omega N\ell}$	$\mu N\ell+2\omega N(\mu N\ell)^2$	$\frac{\ell/4}{1-\omega N\ell}$	eq. (4.94)

Table 4.1: Statistics of the intra-population phenotype observables. Characteristics of intra-population trait average Γ and trait diversity Δ for the linked and free-recombining genomes. The table shows the stationary probability distributions for these macroscopic observables in neutrality, $P_0(\Gamma)$ and $P_0(\Delta)$ and their corresponding statistic: cross-population average, $\langle\Gamma\rangle$, $\langle\Delta\rangle$ and variance $\mathbf{var}\Gamma$ and $\mathbf{var}\Delta$ of the distributions . Similar information is shown for populations which have evolved in a linear fitness landscape (Sel.-Lin.), $f(\phi)=\alpha\phi$ and in a quadratic fitness landscape (Sel.-Quad.), $f(\phi)=\omega(\phi-\phi^\star)^2$. ϕ is the polygenic trait under study with ℓ constitutive loci, N is the population size and μ is mutation rate per nucleotide per generation.

exactly under such dynamics; see Table. 4.1. Characteristic properties of the fitness function are coupled to the phenotype observables, and hence become measurable in the population; see Section. 4.3. A quadratic fitness landscape, $f(\phi)=\omega(\phi-\phi^\star)^2$ with $\omega<0$, imposes a stabilizing selection on traits in the population which reduces both trait divergence across population var(Γ) and diversity within populations Δ. In this description, the stationary state of the polygenic traits in linked genome is set by

mutation-selection-drift balance; all these evolutionary forces may in general influence the trait on comparable timescales. The example of such type is discussed in Chapter. 3.

Application of the polygenic analysis to genomic data

The possibility of inferring genotype-phenotype maps from large genomic datasets requires a better characterization of evolutionary dynamics, such that it is applicable to the analysis of genomic variation. Phenotypic variations can be viewed on different timescales:

(i) Trait diversity across species. Nucleotide divergence across species has been used as a proxy for positive selection and evidence for adaptation i.e., response to the change of fitness preference during evolution (McDonald and Kreitman, 1991). Substitution patterns have also been used in analysis of molecular traits such as binding sites by presuming that the underlying fitness landscape has been maintained during the evolutionary divergence (Mustonen et al., 2008; Mustonen and Lässig, 2005; Nourmohammad and Lässig, 2011). In most of these cases however, this assumption is questionable and is mainly made due to the limitations in genomic sequence data of each population. Analysis of quantitative traits with numerous linked loci should incorporate the full dynamics with mutation-selection-drift balance at stationary state. Inference of positive selection and degree of adaptation between two species can the be characterized on those basis.

(ii) Trait variations within a population. This is the short-term variation of phenotypes between the individuals of a population which experience similar selective constraints. In this picture, the average phenotype in the population remains close to the fitness peak and the phenotypic diversity is reduced in comparison to the neutrally evolving populations; see Fig. 4.9(a). Different populations evolving in a common fitness landscape also cluster around the fitness peak, and the difference between their typical phenotypes (Fig. 4.9(a)) is much smaller than that of the neutrally evolving populations (Fig. 4.9(b)). One of the immediate applications of this approach is to characterize the population dynamics in evolution experiments with multiple independent populations. Analysis of the phenotypic polymorphism is an unbiased approach

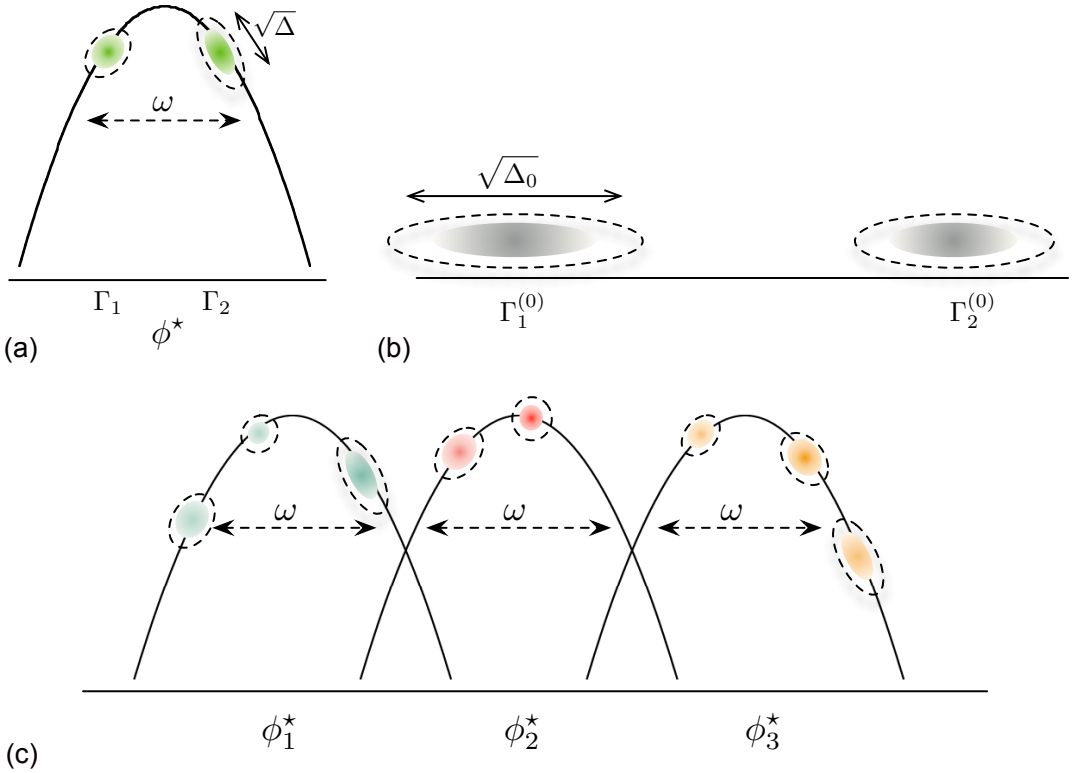

Figure 4.9: Trait variations on different time-scales. (a) Populations evolving in a quadratic fitness landscape for the trait ϕ. Each population is shown by a cloud of individuals spread around a phenotypic average Γ with a diversity Δ. The populations are clustered around the fitness peak $\phi^\star$. (b) Neutrally evolving populations. The intra-population phenotypic diversity Δ is larger than the ones evolving under quadratic selection in (a). Also, the population clouds are located further from each other than in (a) which indicates the absence of selection for a specific phenotype value ($\phi^\star$ in (a)). (c) Cross-loci phenotypic composition of a population. Different loci are pictured to evolve in fitness landscapes with similar shapes (curvature) yet different location of the peak. Populations are again shown with clouds that center around the intra-population phenotype average Γ with a spread Δ. Different fitness landscapes are associated with different loci with the phenotype preference $\phi_i^\star$ (i=1,2,3). The intra-population phenotype diversity is not sensitive to the value of $\phi^\star$ and hence the Δ-statistics can be formed from all these loci to extract the common width of these fitness functions, ω. This method is most practical for analysis of molecular phenotypes with common biophysical constraints e.g., transcription factor binding site interactions. We found this type of fitness landscape to be consistent with evolution of regulatory complexes in yeast; Chapter. 3.

to characterize the evolutionary forces that derive the population.

(iii) Trait variations across loci. This level of diversity is particularly informative

about the biophysically constrained molecular phenotypes, such as transcription factor-binding site interactions (Kinney et al., 2007; Mustonen and Lässig, 2005), nucleosome positioning (Tsankov et al., 2010) and protein folding (Fernández and Lynch, 2011). The peculiar feature of these molecular phenotypes is that the biochemical interactions, which determine the trait functionality, is shared even across various loci in the genome. For example, the binding characteristics of a transcription factor molecule dictates the nucleotide composition of its binding sites across the genome (see Chapter. 1). This feature introduces another type of phenotypic similarity which is not due to common descent but rather common functional characteristics between loci. On the other hand, these loci may often be required for differential functional outputs (e.g., promoters with different regulatory outputs yet interacting with the same transcription factor). In other words, the fitness optimum for each of these trait loci is located at the value that best matches its own regulatory output. If fitness functions for all loci are equal, they can be thought of as different realizations of the same evolutionary process, and cross-loci statistics will match the cross-population statistics of the trait observables as discussed above. If only the fitness optimum $G^\star$ differs between the loci, then the cross-loci Γ-statistics differ from that of the cross-population description, but the higher order statistics (Δ-statistics for our analysis) will remain compatible. As a result, we can form cross-loci statistics by averaging over all fitness landscapes with similar curvatures; see Fig. 4.9(c). In this way, different loci are effectively treated as separate population realizations to form the Δ-statistics. This approach is highly practical for genomic analysis of compatible traits.

In Chapter. 3 we combined the intra-population and cross-loci phenotype variations and extract information on the evolutionary dynamics of promoter complexes in yeast.

Complex fitness function and complex phenotypes

Genomic loci, especially in eukaryotes, may encode multiple phenotypes. The analysis of such pleiotropic features have been carried out mostly in the context of Fisher's geometrical model in multi-dimensional phenotype space (Fisher, 1930). In this model, phenotypes are presented as points a multi-dimensional space where the axis correspond to phenotype characters.The fitness is a decreasing function of the phenotype distances to the local optimum. Mutations in this model are stochastic events that are defined on

this phenotype space and create a new phenotype from the pre-existing ones. Clearly, this model does not incorporate the genomic information to the evolutionary dynamics of the phenotypic traits. The connection to the population composition of genotypes which map into multiple phenotypes in an organisms has not been often discussed. The analysis in this section can in principle be generalized to multi-dimensional phenotypes, but the existence of evolutionary equilibrium in that context is unresolved. The fitness landscapes also can be arbitrarily more complicated. In this section, we only discuss an equilibrium state of a population which evolve in a static fitness landscape. Adaptation and driven evolution is of course lacking from this picture. Generalization of the macroscopic framework in this chapter to integrate the time-dependent characteristics of a fitness function (Mustonen and Lässig, 2007, 2010), is our next step towards understanding the adaptive evolution of the polygenic traits.

References

Abdulrehman, D. and Monteiro, P. (2011). YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic acids Research*, 39(Database):D136–D140. 69

Arnosti, D., Barolo, S., Levine, M., and Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development*, 122(1):205–214. 44

Arnosti, D. N. and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898. 3, 45

Badis, G. and et. al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723. 2

Barton, N. and Otto, S. (2005). Evolution of recombination due to random drift. *Genetics*, 169(4):2353–2370. 79

Barton, N. and Turelli, M. (1989). Evolutionary quantitative genetics: how little do we know? *Annual Review of Genetics*, 23:337–370. 74, 83

Barton, N. H. and Coe, J. B. (2009). On the application of statistical physics to evolutionary biology. *Journal of theoretical biology*, 259(2):317–324. 48, 56, 74, 83

Baxter, G. and Blythe, R. (2007). Exact solution of the multi-allelic diffusion model. *Mathematical Biosciences*, 209:124–170. 74, 79

Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106(4):1133–1138. 73

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 27(2):573–580. 20, 24

Berg, J. and Lässig, M. (2003). Stochastic evolution of transcription factor binding sites. *Biophysics Moscow*, 48:S36–S44. 17, 26, 27

Berg, J., Willmann, S., and Lässig, M. (2004). Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4(1):42. 11, 17, 26, 27, 56, 73, 77

REFERENCES

Berg, O. and von Hippel, P. (1987). Selection on DNA-binding sites by regulatory proteins. Statistical mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193(4):723–750. 2, 6, 16

Bergman, C. and et. al. (2002). Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biology*, 3(12). 3, 15

Bergman, C. M., Carlson, J. W., and Celniker, S. E. (2005). Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, 21(8):1747–1749. 18, 24, 26, 32, 37, 39

Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124. 41, 44, 45

Boeva, V., Regnier, M., Papatsenko, D., and Makeev, V. (2006). Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, 22(6):676–684. 15, 16, 24, 37

Britten, R. J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proceedings of the National Academy of Sciences*, 93(18):9374–9377. 15

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9):5136–5141. 3, 5, 15, 41

Comeron, J. and M, K. (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics*, 161(1):389–410. 74

Davidson, E. H. (2006). *The regulatory genome: gene regulatory networks in development and evolution*. Academic, Burlington, MA. 2, 3, 15, 38

de Vladar, H. P. and Barton, N. H. (2011a). The contribution of statistical physics to evolutionary biology. *Trends in Ecology & Evolution*, 26(8):424–432. 48, 74, 83, 88

de Vladar, H. P. and Barton, N. H. (2011b). The statistical mechanics of a polygenic character under stabilizing selection, mutation and drift. *Journal of The Royal Society Interface*, 8(58):720–739. 48, 74, 83

Desai, M. (2007). Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 17(5):385–394. 74, 92

Desai, M. and Fisher, D. (2007). The speed of evolution and maintenance of variation in asexual populations. *Current biology*. 48, 74, 92

Djordjevic, M., Sengupta, A., and Shraiman, B. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13(11):2381–2390. 5

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK. 23, 27

Falconer, D. S. (1989). *Introduction to quantitative genetics*. Halsted Press. 48, 73, 83

Fernández, A. and Lynch, M. (2011). Non-adaptive origins of interactome complexity. *Nature*, 474(7352):502–505. 73, 111

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

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford University Press, USA, 1st edition. 73, 74, 83, 111

Gallo, S. M., Li, L., Hu, Z., and Halfon, M. S. (2006). REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*, 22(3):381–383. 18, 24, 26, 32, 37, 39

Gardiner, C. (2004). *Handbook of Stochastic methods: for physics, chemistry and the natural sciences*. Springer, 3rd edition. 8, 85, 86, 95

Gerrish, P. J. and Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, 102/103:127–144. 48, 74, 92

Gershenzon, N. and Stormo, G. (2005). Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Research*, 33(7):2290–2301. 6

Gertz, J. and Cohen, B. A. (2009). Environment-specific combinatorial cis-regulation in synthetic promoters. *Molecular Systems Biology*, 5. 41, 42

Gertz, J., Siggia, E. D., and Cohen, B. A. (2008). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218. 12, 42, 45

Gruen, D. (2006). *Statistics and evolution of functional genomic sequence*. PhD thesis, Univeristät zu Köln. 15, 16, 18, 37

Halfon, M. S., Gallo, S. M., and Bergman, C. M. (2008). REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Research*, 36(Database issue):D594–8. 18, 24, 26, 32, 37, 39

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917. 27

Hancock, J. M., Shaw, P. J., Bonneton, F., and Dover, G. A. (1999). High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. *Molecular Biology and Evolution*, 16(2):253–265. 15

Harbison, C. and et. al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104. 3, 15

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., and Eisen, M. B. (2008). Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLoS Genet*, 4(6):e1000106. 41

REFERENCES

Hartl, D. and Taubes, C. (1996). Compensatory nearly neutral mutations: selection without adaptation. *Journal of Theoretical Biology*, 182(3):303–309. 73

He, X., Samee, M. A. H., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: The roles of synergistic activation, cooperative binding and short-range repression. *PLoS Computational Biology*, 6(9):e1000935. 44

Iwasa, Y. (1988). Free fitness that always increases in evolution. *Journal of Theoretical Biology*, 135(3):265–281. 56

Jaynes, E. (1957). Information Theory and Statistical Mechanics. II. *Physical Review*, 108(2):171–190. 6

Katti, M. V., Ranjekar, P. K., and Gupta, V. S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution*, 18(7):1161–1167. 40

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713. 8, 27, 48, 83

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*. 10, 73

Kimura, M. and Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763. 27, 76

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116. 1

Kinney, J. B., Tkacik, G., and Callan, C. G. J. (2007). Precise physical models of protein - DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506. 111

Kirkpatrick, M., Johnson, T., and Barton, N. (2002). General models of multilocus evolution. *Genetics*, 161(4):1727–1750. 48, 74, 83

König, P., Giraldo, R., Chapman, L., and Rhodes, D. (1996). The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell*, 85(1):125–136. 69

Kuhlman, T., Zhang, Z., Saier, M. H. J., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences*, 104(14):6043–6048. 41

Kulkarni, M. and Arnost, D. (2005). Cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. *Molecular and cellular biology*, 25(9):3411–3420. 45

Kulkarni, M. M. and Arnosti, D. N. (2005). cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. *Molecular and Cellular Biology*, 25(9):3411–3420. 3, 38

Kullback, S. and Leibler, R. A. (1951). On Information and sufficiency. *Annals of mathematical statistics*, 22(1):79–86. 19

Lam, F. H., Steger, D. J., and O'Shea, E. K. (2008). Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–250. 44

Lande, R. (1976). Natural-selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334. 48, 83

Lässig, M. (2007). From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6):S7. 5, 6, 11, 17, 27

Lavoie, H., Hogues, H., Mallick, J., Sellam, A., Nantel, A., and Whiteway, M. (2010). Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biology*, 8(3):e1000329. 69

Lee, T. I. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804. 6

Lenski, R. and Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15):6808–6814. 50

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 6945(424):147–151. 3, 15, 41

Liti, G. and et. al. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341. 69

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769):564–567. 41

Ludwig, M. Z. and Kreitman, M. (1995). Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Molecular Biology and Evolution*, 12(6):1002–1011. 41

Ludwig, M. Z., Patel, N. H., and Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125(5):949–958. 41, 73

Lusk, R. W. and Eisen, M. B. (2010). Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genetics*, 6(1):e1000829. 3

Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution*, 23(2):450–468. 3, 15

Lynch, M. and Conery, J. (2003). The origins of genome complexity. *Science*, 5649(302):1401–1404. 38, 39

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Associates Inc. 48, 73, 83

REFERENCES

Maerkl, S. J. and Quake, S. R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237. 2, 5

Markstein, M., Markstein, P., Markstein, V., and Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, 99(2):763–768. 3, 38

McDonald, J. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654. 73, 109

Messer, P. W. and Arndt, P. F. (2007). The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24(5):1190–1197. 15, 16, 18, 24, 37

Mirny, L. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences*, 107(52):22534–22539. 44

Monod, J. and Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:389–401. 1

Monteiro, P., Mendes, N., and Teixeira, M. (2008). YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*, 36(Database):D132–D136. 69

Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3:19. 26, 27

Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N., and Eisen, M. B. (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, 5(12):R98. 26, 27

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36(12):1331–1339. 2, 6

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9. 60

Mustonen, V., Kinney, J., Callan, C., and Lässig, M. (2008). Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences*, 105(34):12376–12381. 17, 30, 40, 70, 73, 78, 109

Mustonen, V. and Lässig, M. (2005). Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences*, 102(44):15936–15941. 10, 11, 27, 29, 30, 70, 78, 109, 111

Mustonen, V. and Lässig, M. (2007). Adaptations to fluctuating selection in Drosophila. *Proceedings of the National Academy of Sciences*, 104(7):2277–2282. 112

Mustonen, V. and Lässig, M. (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet*, 25(3):111–119. 27

Mustonen, V. and Lässig, M. (2010). Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences*, 107(9):4248–4253. 56, 74, 77, 88, 98, 112

Neher, R. A. and Shraiman, B. I. (2009). Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16):6866–6871. 79

Neher, R. A. and Shraiman, B. I. (2011). Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics*, 83(4):1283–1300. 48, 74, 79, 83

Nourmohammad, A. and Lässig, M. (2011). Formation of regulatory modules by local sequence duplication. *PLoS Computational Biology*, q-bio.PE. vi, 12, 43, 109

Ondek, B., Gloss, L., and Herr, W. (1988). The SV40 enhancer contains two distinct levels of organization. *Nature*, 333(6168):40–45. 3, 15

Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Research*, 35(Database):D127–D131. 40, 69, 70

Park, S. and Krug, J. (2007). Clonal interference in large populations. *Proceedings of the National Academy of Sciences*, 104(46):18135–18140. 48, 74, 92

Poelwijk, F., Kiviet, D., Weinreich, D., and Tans, S. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386. 73

Ptashne, M. and Gann, A. (2002). *Genes and signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. 1, 3, 15, 38, 44, 45

Raveh-Sadka, T. and Levo, M. (2009). Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Research*, 19:1480–1496. 44

Ren, B. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309. 6

Rice, S. (1990). A geometric model for the evolution of development. *Journal of theoretical biology*. 73

Rouzine, I. (2010). Multi-site adaptation in the presence of infrequent recombination. *Theoretical population biology*. 79

Rouzine, I., Brunet, É., and Wilke, C. (2008). The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theoretical population biology*, 73(1):24–46. 48, 74, 92

Schiffels, S., Szollosi, G. J., Mustonen, V., and Lässig, M. (2011). Emergent neutrality in adaptive asexual evolution. *Genetics*, 189(4):1361–1375. 48, 74, 92

Segal, E. and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current Opinion in Structural Biology*, 19(1):65–71. 19

Sella, G. and Hirsch, A. E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27):9541–9546. 56, 74

REFERENCES

Shea, M. (1985). The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *Journal of Molecular Biology*, 181(2):211–230. 41, 44

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE*, 5(3):e9722. 6

Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, 5:129. 3, 15

Sinha, S. and Siggia, E. D. (2005). Sequence turnover and tandem repeats in cis-regulatory modules in Drosophila. *Mol Biol Evol*, 22(4):874–885. 15, 16, 18, 37

Small, S., Arnosti, D. N., and Levine, M. (1993). Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development*, 119(3):762–772. 3, 38

Smith, J. (1970). Natural selection and the concept of a protein space. *Nature*, 225:563–564. 73

Stanojevic, D., Small, S., and Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science*, 254(5036):1385–1387. 3

Stark, A. and et. al. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232. 3

Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23(3):109–113. 2, 6, 19, 26

Tanay, A. and Siggia, E. D. (2008). Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biology*, 9(2):R37. 15, 16, 18, 24, 37

Teixeira, M. and et. al. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic acids Research*, 34(Database):D446–D451. 69

Thornton, K. and Long, M. (2002). Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. *Molecular Biology and Evolution*, 19(6):918–925. 40

Tsankov, A., Thompson, D., Socha, A., and Regev, A. (2010). The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biology*, 8(7):e1000414. 111

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 324(5931):1213–1216. 15, 37

Weinreich, D., Delaney, N., and DePristo, M. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114. 73

Wunderlich, Z. and Mirny, L. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–40. 4, 21, 35

Zusammenfassung

Die seit kurzem bestehende Verfügbarkeit riesiger genomischer Datenmengen und die Verbesserung von Technologien zur Sequenzierung eines Genoms ermöglichen es, dass die Populationsgenetik sich von zumeist abstrakt-theoretischen Grundlagen hin zu einer quantitativen Molekültheorie weiterentwickelt. Funktionseinheiten in der DNA sind jedoch normalerweise Kombinationen aus interagierenden Nukleotidsegmenten und die evolutionären Kräfte, die sich auf diese Segmente auswirken, können zu sehr komplizierten Populationsdynamiken führen. Es ist das Ziel, diese Interaktionen so zu beschreiben, dass die makroskopischen Eigenschaften unabhängig von den mikroskopischen Details dargestellt werden, wie in der statistischen Mechanik.

In dieser Doktorarbeit beschäftige ich mich mit der Evolutionsdynamik von regulierenden Sequenzen, die die Produktion von Proteinen in den Zellen steuern. Eine der wichtigsten Regulationsarten tritt durch das Zusammenspiel von Proteinen, die *Transkriptionsfaktoren* genannt werden, mit *Bindungsstellen* in der DNA-Sequenz auf. Die Stärke dieser Interaktionen beeinflusst die Fitness des Individuums in der Population. Da man diesen Zusammenhang herleiten kann, ist dies ein ideales Modellsystem fr die quantitative Analyse von genetischer Evolution.

Verglichen mit Prokaryoten und Hefe ist die genetische Regulation bei höheren Eukaryoten viel komplexer. Die Informationen fr die Regulation sind in Module mit mehreren Bindungsstellen aufgeteilt, die mit einer gemeinsamen Funktion in Zusammenhang stehen. In Kapitel. 2 zeigen wir, dass die Bildung von Bindungsstellenkomplexen üblicherweise durch lokale Sequenzduplikationen geschieht und nicht aus dem Nichts durch einzelne Punktmutationen entsteht. Weiterhin zeigen wir, dass die zugrunde liegende regulatorische “Grammatik” mit diesem Mechanismus in Einklang steht, sodass die Duplikationen einen Anpassungsvorteil bedeuten.

Regulatorische Komplexe ähneln einem Vielteilchensystem, dessen Funktion sich aus den kollektiven Dynamiken seiner Elemente herausbildet. In Kapitel. 3 entwickeln wir ein thermodynamisches Modell um die tatsächliche Affinität von Bindungsstellenkomplexen zu mehreren Transkriptionsfaktoren mit zusammenwirkender Bindung zu charakterisieren. Diese Affinitäten sind der Phänotyp oder das Merkmal eines Bindungskomplexes, auf den Selektion einwirkt, und wir charakterisieren ihre Evolution. Aus den Polymorphismusdaten des Hefegens leiten wir eine "Fitness-Landschaft" anhand des Verhältnisses von Fitness zu Bindungswahrscheinlichkeit ab unter Verwendung der neuartigen Methode, die in Kapitel. 4 entwickelt wird. Durch diese Methode der quantitativen Merkmalsanalyse können langfristige Korrelationen zwischen Bindungsstellen, wie sie in asexuellen Populationen auftreten, verarbeitet werden. Mit unserer "Fitness-Landschaft" treffen wir quantitative Voraussagen zur erhaltenen Phänotyp-Menge, sowie zur Menge der ausgleichenden Veränderungen zwischen den Bindungsstellen.

Unsere Ergebnisse weisen einen neuen Weg hin zum Verständnis der regulatorischen "Grammatik" des eukaryotischen Genoms basierend auf quantitativen Evolutonsmodellen. Sie beweisen, dass eine Kombination von theoretischen Modellen, experimentellen Hochdurchsatzmessungen und die Analyse von genetischen Variationen für das richtige quantitative Verständnis von biologischen Systemen notwendig ist.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstndig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit − einschließlich Tabellen, Karten und Abbildungen −, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie − abgesehen von unten angegebenen Teilpublikationen − noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von (Name des anleitenden Dozenten oder der anleitenden Dozentin) betreut worden.

Köln, August 2012