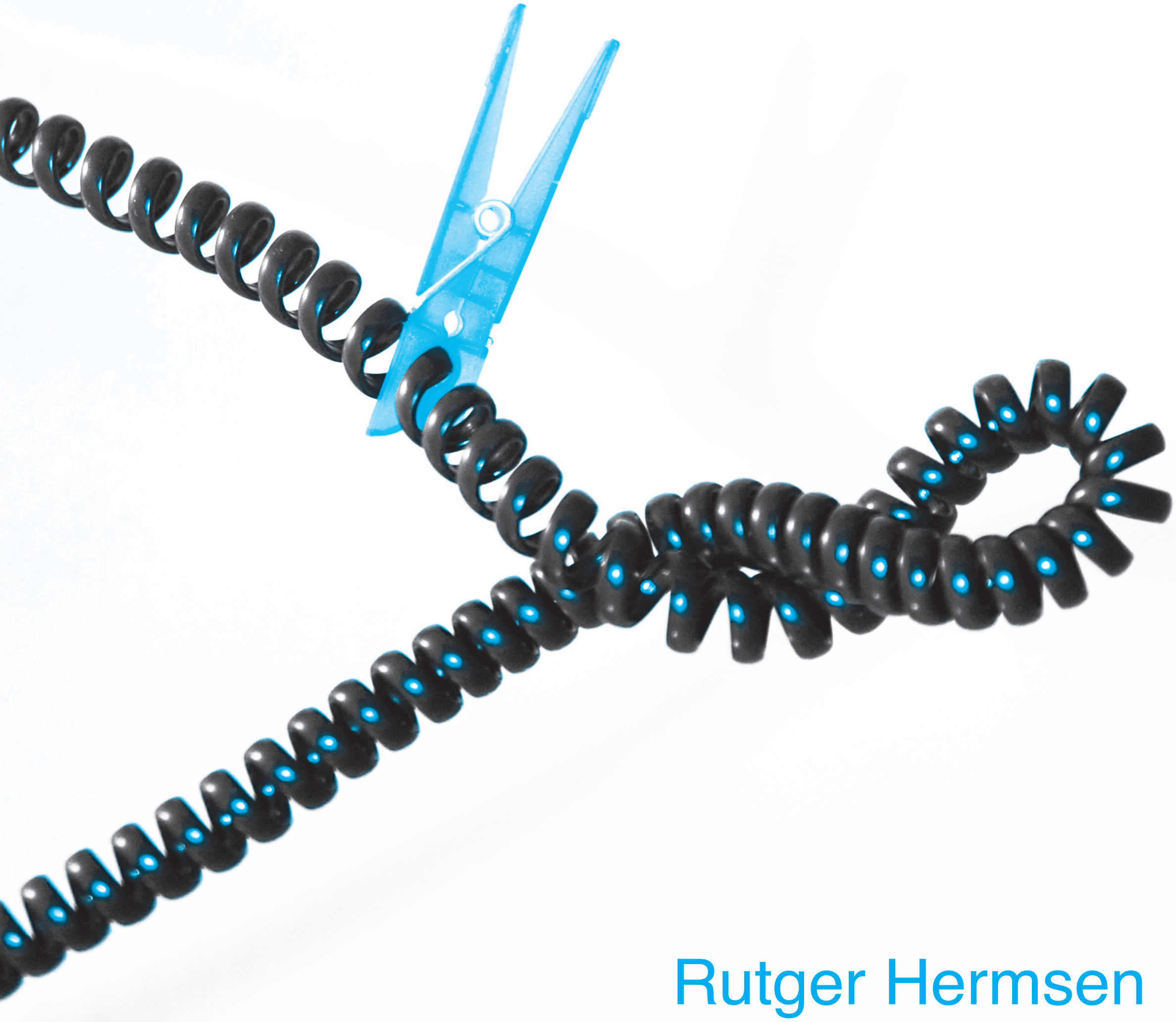
Transcription Regulation
and
Genome Organization



Rutger Hermsen

Transcription Regulation
and Genome Organization

This thesis was reviewed by:

Prof. Dr. U. Gerland
Prof. Dr. P. Hogeweg
Prof. Dr. F. C. MacKintosh
Dr. S. Tans
Dr. S. A. Teichmann

ISBN-13: 978-90-77209-27-1

ISBN-10: 90-77209-27-1

Nederlandse titel: Transcriptieregulatie en de organisatie van genomen

The work described in this thesis was performed at the FOM-Institute for Atomic- and Molecular Physics, Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands. The work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The work of the FOM-institute is part of the research program of the Stichting Fundamenteel Onderzoek der Materie (FOM) and is supported by NWO.

VRIJE UNIVERSITEIT

Transcription Regulation and Genome Organization

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op dinsdag 28 oktober 2008 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Rutger Hermsen

geboren te Leiderdorp

promotor: prof.dr. P.R. ten Wolde

Publications:

Part of the work presented in this thesis has been published in the following publications:

Hermsen, R, Tans, S, and ten Wolde P. R., 2007. Transcriptional regulation by competing transcription factor modules. *PLoS Comput Biol*, **2**(12):e164

Hermsen, R., ten Wolde, P. R., and Teichmann, S. A., 2008. Chance and necessity in chromosomal gene distributions. *Trends in Genetics*, **24**(5):216–219

Hermsen, R., and ten Wolde, P. R., 2008. The role of terminator loss in the evolution of genomes. *Submitted*

Contents

1 Introduction 1

1.1 Transcription regulation . . . 3
  1.1.1 Binding the DNA . . . 3
  1.1.2 Regulation of transcription initiation in prokaryotes . . . 9
1.2 Evolutionary population genetics . . . 14
  1.2.1 A Wright–Fisher model with selection . . . 14
  1.2.2 Fixation probability . . . 15
  1.2.3 Limits . . . 17
  1.2.4 Fixation rates . . . 18

2 Regulation by modules 19

2.1 Introduction . . . 20
2.2 Model of transcriptional regulation . . . 23
2.3 Evolutionary design of logic gates . . . 25
2.4 Results: *cis*-regulatory constructs . . . 26
  2.4.1 Homo-cooperative auxiliary sites provide steep responses . . . 26
  2.4.2 Hetero-cooperativity provides conditional responses . . . 28
  2.4.3 Competition between modules . . . 29
2.5 Discussion and outlook . . . 30
2.A Detailed description of the model of transcription regulation . . . 34
2.B Affinities of auxiliary sites . . . 38
2.C Minimal models for the complex gates . . . 42
2.D Extending the model . . . 45

3 Auto-regulation and bi-stability in transcription regulation 47

3.1 Introduction . . . 48
3.2 Models and methods . . . 49
  3.2.1 The genetic switch . . . 49
  3.2.2 Logic gates . . . 52
3.3 Results . . . 53
  3.3.1 The genetic switch . . . 53
  3.3.2 Logic gates . . . 58
3.4 Discussion, conclusions and outlook . . . 61

4 Chance and necessity in gene distributions 63
4.1 Probability distributions of intergenic distances 64
4.2 Models 64
4.2.1 The Ideal-Gas model 64
4.2.2 The Tonks-Gas model 66
4.2.3 The Constant-Force model 66
4.3 Results 70
4.3.1 The CF model fits the C and T distribution of *S. cerevisiae* 70
4.3.2 Typical UCR and DCR sizes for other fungi 70
4.3.3 *S. cerevisiae* contains many bi-directional UCRs 70
4.3.4 *E. coli* has many bi-directional terminators 72
4.3.5 The operon structure of *E. coli* is evident from the T distribution 72
4.4 Concluding remarks 73
4.A The constant-force model 74
4.A.1 Assumptions 74
4.A.2 Derivation of the distance distributions: CF interaction with fixed range 74
4.A.3 CF interactions with different ranges 75
4.B More detailed models 77
4.B.1 Polydisperse UCRs and DCRs 77
4.B.2 Alternative potentials 78
4.B.3 Yeast DCRs 78
4.B.4 Higher eukaryotes 79
4.C Fitting procedure 79
4.D Statistical tests 81
4.D.1 Co-expressed divergent pairs in *S. cerevisiae* are closer together than expected 82
4.D.2 Closely spaced divergent genes functionally related 83
4.D.3 Greater than expected number of convergent regions in short convergent regions 84
4.D.4 Terminators in C regions more often bi-directional 84
4.D.5 Putative bi-directional terminators occur in short regions 85
4.D.6 Closely spaced operons not more often functionally related 85

5 Terminator loss 87
5.1 Introduction 88
5.2 Model 90
5.2.1 A daisy-chain genome 90
5.2.2 Four elementary types of mutations 91
5.2.3 Kimura–Ohta population genetics 91
5.2.4 Selection on genome length and gratuitous transcription . 92
5.3 Results 92

5.3.1 Spontaneous formation of operons 93
5.3.2 Bi-directional terminators 94
5.3.3 Spacing of genomic elements 95
5.4 Discussion 97
5.A Methods 99
5.A.1 Simulating population genetics 99
5.A.2 Parameters 99
5.A.3 Initialization 100
5.B Derivation of Equation 5.5 101

6 Evolution of intergenic regions 103
6.1 Introduction 104
6.2 Short description of the model 104
6.2.1 Model assumptions 105
6.2.2 Master equation for spacers 106
6.2.3 The joint probabilities 107
6.2.4 Merging and splitting 107
6.3 Results 108
6.3.1 The data: *E. coli* versus *S. dysenteriae* and *S. Typhi* . . . 108
6.3.2 The model can fit the data 109
6.3.3 Required number of mutations far too high 110
6.3.4 Larger insertions and deletions 111
6.3.5 Outliers are candidates for merging or splitting 113
6.4 Conclusions 117
6.A The data for the scatterplots 118
6.B Details of the model 118
6.B.1 The evolution of spacers 118
6.B.2 The evolution of the UTRs 120
6.B.3 Combining the UTRs and the untranscribed spacers . . . 121
6.C Parameter choices 124

Summary 137

Samenvatting 143

Acknowledgments 149

Curriculum vitae 151

Chapter 1

Introduction

These days, the words DNA, genes, chromosomes, evolution and mutants have become commonplace. Movies such as Jurassic Park, cartoons like the Teenage Mutant Ninja Turtles or Pokemon, and forensic TV series such as CSI have introduced the main concepts of genetics and genomics to a broad audience. At the time of this writing, the Dutch newspapers discuss the genome of Marjolein Kriek, the first woman who's complete DNA has been sequenced, and headlines question whether the Dutch Cabinet will reach consensus on the issue of embryo selection by genetic screening. Genetically modified vegetables can be bought in all supermarkets and transgenic bacteria produce insulin for diabetes patients. It is clear that, only forty years after the discovery of the genetic code and the establishment of the *central dogma of molecular biology* (see Box 1.1), the genetics revolution has a major impact on society.

And yet, the fields of genetics and genomics are still in their infancy. If we aim to understand the properties, behavior and evolution of living systems, unraveling the genetic code and recording sequences are only the beginning. To understand why, it is useful to compare cells to computers. Very much like computers, which respond to input signals such as the keyboard, the mouse and messages from the internet, cells respond to environmental clues such as temperature changes, chemical gradients and mechanical stress. Based on such signals, cells make complex decisions and adjust their behavior accordingly. Therefore cells, like computers, are information processing machines. In this analogy, the discovery of DNA as the main biological medium for information storage is similar to the discovery of a computer's *hard drive*. And the unraveling of the genetic code can perhaps be compared to the revelation that information on the hard drive is encoded in *bits* that are organized in *bytes*. Clearly, such discoveries

Box 1.1: The central dogma

The central dogma of molecular biology is a statement about the information flow of hereditary information in all living organisms. The dogma states that information is stored in and inherited through deoxyribonucleic acid (DNA) molecules. Pieces of these molecules can be copied or *transcribed* into ribonucleic acid (RNA) molecules. This task is performed by multi-subunit enzymes called RNA polymerase (RNAp). In many cases, the produced RNA is subsequently used as a template for the production of a specific protein; the RNA is then called messenger RNA (mRNA). In all prokaryotes and in a number of eukaryotes, one unit of transcription (*transcription unit* or TU) can contain multiple genes. In such cases, the co-transcribed set of genes is called an *operon*. In the process of *translation*, molecular machines called *ribosomes* read the chain of nucleotides that constitute the mRNA. This sequence of nucleotide specifies the chain of amino acids that make up the protein. Each triplet of nucleotides specifically codes for one amino acid in the chain. The protein is assembled by the ribosome while reading the RNA.

Information therefore (normally) flows in one direction only: DNA specifies RNA, and RNA specifies protein.

would be important breakthroughs for hypothetical scientists trying to unravel the internal workings of a computer, but nevertheless only small steps towards understanding the behavior of a complete PC running a full-fledged operating system. A similar gap separates knowing DNA sequences or even lists of gene functions from understanding cellular behavior.

In order to bridge the gap, at least two things are required. First: more information. Indeed, both small-scale and massive high-throughput experiments all around the world now produce vast amounts of data on molecular functions, genomic organization, gene expression, regulatory DNA sites, phylogeny, interactions between proteins, localization of gene products and many, many other aspects of living systems. Inevitably, analyzing, interpreting and coupling these data sets has become a major part of biology and has lead to a new discipline: *bioinformatics*. Second: a deeper understanding of biological *mechanisms*. Even if the gene products and cellular components involved in a given process are known, it is usually far from trivial to understand *how* these constituents give rise to the observed behavior or which functions they convey. Due to the complexity of the systems under study, a profound understanding often requires the formulation and analysis of quantitative models. In essence, such models are often not very different from models that have been developed for physical systems in the course of the past centuries. Consequently, formalisms from physics are now successfully being applied to biological problems. The work presented in this thesis is intended as a contribution in this direction.

In this work, we focus on the phenomenon of *transcription regulation*. This term refers to the mechanisms used by cells to dynamically adjust the rate at which genes are transcribed (see Box 1.1). Transcription regulation is one of the main mechanisms allowing cells to respond to extra-cellular and intra-cellular signals. The overarching question connecting all chapters in this thesis can be summarized as: what are the mechanisms involved in transcription regulation, and to what extent do these process shape the organization of genomes? In our attempt to answer these questions, we mainly focus on prokaryotes (in particular the bacterium *Escherichia coli*) and on one of the less complex eukaryotic organisms: *Saccharomyces cerevisiae* (baker's yeast).

Below, we introduce a number of concepts and results that are used in the subsequent chapters. It is not our intention to present an exhaustive review of each of the subjects, but rather to provide readers that are not familiar with these topics with a minimal introduction to the essential concepts. First, we discuss the process of transcription regulation and the framework used in quantitative models describing these processes. Next, in Section 1.2, we briefly introduce some basic results from the field of evolutionary population genetics.

1.1 Transcription regulation

Transcription regulation is one of the main mechanisms that allow cells to adjust their behavior to external and internal clues. By regulating the rate at which genes are transcribed, cells can rapidly adapt their protein and RNA content to their capricious environments. As these proteins and RNA molecules function as enzymes and structural building materials, they determine the properties of the cell to a large extent. In unicellular organisms such as most bacteria, transcription regulation is responsible for an enormous range of processes such as circadian clocks, DNA repair, the cell cycle, and responses to different food sources and temperature changes. In multicellular organisms, it is also a crucial factor in cell differentiation.

The processes of transcription regulation are mediated by a special class of proteins called transcription factors (TFs). TFs are capable of binding rather specifically to certain DNA sequences, called TF-binding sites or *operators*. As a result, they can interfere with the transcription process of particular genes. If a TF has the capacity to increase the rate at which a gene is transcribed, it is called an *activator*. Conversely, if it tends to lower the rate of transcription, it is termed a *repressor*. Whether a TF is an activator or a repressor depends on the structure of the TF, but also on the location of its binding sites. Some TFs, in particular in bacteria, can act as an activator or as a repressor depending on the context and are therefore called *dual regulators*.

Many TFs are sensitive to some external signal. For instance, small molecules (ligands) such as cyclic AMP or sugars may bind to TFs. In such examples, the conformation of the TF changes allosterically upon binding and consequently also its DNA-binding affinity or specificity alters. Repression or activation of certain genes can thus be relieved or induced depending on physical and chemical signals.

The genes coding for the transcription factors themselves are typically regulated as well, by yet other TFs. The set of these regulatory dependencies between TFs and their target genes invokes the picture of a *network* of genes. Such a network can be depicted as a graph. Transcription units (TUs; see Box 1.1) are then represented as nodes, and regulatory interactions between two TUs are depicted as directed links between the corresponding nodes. Such networks are called transcription regulatory networks (TRNs).

1.1.1 Binding the DNA

Transcription factors and RNA polymerase (RNAp) bind to specific sites on the DNA. The chromosome of *E. coli* contains about 4.6 million base pairs; therefore, there are more than 9 million sites where each molecule could bind. Hence in order to bind to the correct site with high probability, DNA-binding molecules need to recognize their target sites with high precision. In other words, DNA-binding molecules need to be highly *specific*. This raises many questions about the nature of the binding reactions. How can binding sites be this specific? Which physical parameters are relevant to the specificity? How quickly can a molecule find its target site and which mechanisms are involved in these kinetics? Many of these questions can, and have been, studied using physical models. Below, we summarize some basic facts that underly the model in Chapters 2 and 3.

DNA binding is mediated by electrostatic interactions and hydrogen bonding

Most information about DNA-binding molecules is due to detailed chemical experiments, many of which date back to the late seventies. These experiments have shown that DNA binding is mainly mediated by electrostatic interactions and hydrogen bonding of the protein residues with the major and minor grooves of the DNA helix (von Hippel and Berg, 1986). The structures of transcription factors are indeed specifically adapted to fit in these grooves.

The first three-dimensional structures of transcription factors became available in the early eighties and were all examples from *E. coli* (Baumberg, 1999). The structures of these TFs all fell into the class of so-called Helix–Turn–Helix (HTH) motives. The binding domain of TFs in this class consists of two alpha helices separated by a tight turn. The spacing of these helices, about 34Å, is equal to the spacing between the consecutive major grooves of the DNA. Indeed, TFs of the HTH type, such as CRP, λcro and λcI, bind to their operator sequences by inserting the HTH domain in the major grooves of the operator.

Since the discovery of the HTH structure, many other motives have been discovered (Harrison, 1991). Most of them rely on alpha helices, but exceptions are also known (Baumberg, 1999). Some of the motives occur exclusively in eukaryotic species, such as the so-called zinc fingers (Alberts *et al.*, 1994).

Binding sites: specific, pseudo-specific & unspecific binding

In the literature, often a distinction is made between specific and unspecific binding; unfortunately, these terms are not always properly defined. As this distinction is relevant in our models, we discuss it here.

Experiments have shown (*e.g.* Kabata *et al.* (1993)) that transcription factors and RNAp can bind to DNA in two different modes. In the first mode, the binding affinity for a given site depends strongly on the base-pair sequence of the site and is therefore called (sequence) specific. The distinction between different sequences is made mainly through hydrogen-bonds (von Hippel and Berg, 1986). The nucleotides of the DNA and the amino acids in the TF can form hydrogen-bonds only if the hydrogen-bond donors and receptors in the TF and the nucleotides are properly aligned. Recent work on the *lac* repressor LacI has shown that it folds in a particular way when it is bound specifically (Kalodimos *et al.*, 2004); in this state it wraps around the DNA and makes contact with the major grooves.

Specific binding occurs at the physiologically functional binding sites, but can also occur at non-functional sites if their sequence happens to be similar to the binding sequence of the DNA-binding molecule. We choose to call such sites pseudo-operators, following Berg *et al.* (1981), Winter and Von Hippel (1981) and Winter *et al.* (1981). If we wish to distinguish binding to operators from binding to pseudo-operators, we refer to the latter as pseudo-specific binding.

The second mode of binding hardly depends on sequence and is therefore called non-specific. This mode relies mainly on electrostatic interactions with the DNA backbone (Mossing and Record, 1985; Kalodimos *et al.*, 2004) and is considerably weaker than specific binding to wild-type operators. Due to these non-specific interactions, a large fraction of the DNA-binding molecules that are not bound at their functional sites are bound non-specifically to the DNA; in the case of the TF LacI, for instance, only 10% of the proteins are free in the cytoplasm (Kao-Huang *et al.*, 1977). This behavior may have a functional role. Assuming a DNA-binding molecule can “slide” along the DNA once it is non-specifically bound — which is observed in, for instance, the case of RNAp (Kabata *et al.*, 1993) — the problem of finding its functional site on the DNA is reduced from a three-dimensional problem to a one-dimensional one (Richter and Eigen, 1974; Berg *et al.*, 1981; Winter and Von Hippel, 1981; Winter *et al.*, 1981). If the speed at which target sites are found is relevant, models predict that a balance between 1D and 3D diffusion is optimal. Indeed, experiments suggest that the association rate of LacI may be 100–1000 times higher than the estimated 3D-diffusion-limited rate (Riggs *et al.*, 1970; Berg *et al.*, 1981). However, whether this can indeed be attributed to sliding along the DNA is still subject of debate (*e.g.* Gowers and Halford (2003)).

Physical models of TF–DNA binding

The modeling of protein–DNA interactions in terms of statistical physics was pioneered in classical papers by Otto Berg, Robert Winter and Peter von Hippel, starting from the early eighties (Berg *et al.*, 1981; Winter *et al.*, 1981; von Hippel and Berg, 1986; Berg and von Hippel, 1987; Berg, 1988; Berg and von Hippel, 1988; Berg, 1990, 1992). Their models have recently been supplemented by others (*e.g.* Stormo and Fields (1998); Sengupta *et al.* (2002); Gerland and Hwa (2002);

Gerland *et al.* (2002)). For our purposes, rather minimal models suffice.

In the simplest cases, we are interested in reactions of the kind

$$X + O \xleftrightarrow{K_\mathrm{d}} XO, \tag{1.1}$$

where O is the DNA site to which the molecule X binds, and K_d is the dissociation constant of the binding, defined as the concentration of X at which the equilibrium occupancy of O is one half. If we denote the concentration of X by [X], the fractional occupancy of O in equilibrium as [XO] and define $[\mathrm{O}] \equiv 1 - [\mathrm{XO}]$, then the law of mass action states that in equilibrium:

$$K_\mathrm{d} = \frac{[\mathrm{X}]\,[\mathrm{O}]}{[\mathrm{XO}]}. \tag{1.2}$$

Next, we define the affinity q of X as

$$q \equiv \frac{[\mathrm{X}]}{K_\mathrm{d}} = \frac{[\mathrm{XO}]}{[\mathrm{O}]}. \tag{1.3}$$

From Equations 1.2 and 1.4 and the relation $[\mathrm{O}] + [\mathrm{XO}] = 1$ it immediately follows that

$$[\mathrm{XO}] = \frac{q}{1+q}. \tag{1.4}$$

This relation holds for specific, unspecific and pseudo-specific binding alike.

Equation 1.4 can also be derived from a statistical mechanics viewpoint. We call the partition sum of a single molecule of X that is not bound Z_X; formally, Z_X is a sum (or integral) of Boltzmann factors, where the summation runs over all internal and external degrees of freedom of the molecule in the unbound state. The partition sum of the unbound operator we denote by Z_O. Similarly, Z_XO is defined as the partition sum over all states of the complex XO. In this partition sum, the different binding modes (specific or unspecific) are both included.

We now denote the number of molecules of type X present in the cell by n. Assuming that molecules that are not bound do not mutually interact (*i.e.* behave like ideal particles), the partition sum Z_off of states in which none of the molecules is bound at O can be written as

$$Z_\mathrm{off} = Z_\mathrm{O}\frac{(Z_\mathrm{X})^n}{n!}, \tag{1.5}$$

whereas Z_on, the partition sum of all states in which a molecule *is* bound, equals

$$Z_\mathrm{on} = Z_\mathrm{XO}\frac{(Z_\mathrm{X})^{n-1}}{(n-1)!}. \tag{1.6}$$

The probability that the system is in the bound state, p_{on}, then follows as

$$p_{\mathrm{on}} = \frac{Z_{\mathrm{on}}}{Z_{\mathrm{on}} + Z_{\mathrm{off}}} = \frac{Z_{\mathrm{on}}/Z_{\mathrm{off}}}{1 + Z_{\mathrm{on}}/Z_{\mathrm{off}}}. \tag{1.7}$$

Comparison with Equation 1.4 leads to the identification $q = Z_{\mathrm{on}}/Z_{\mathrm{off}}$, and

$$K_{\mathrm{d}} = \frac{1}{V} \frac{Z_{\mathrm{O}} Z_{\mathrm{X}}}{Z_{\mathrm{XO}}}, \tag{1.8}$$

where V is the volume of the cell. Note that Z_{X}, Z_{O} and Z_{XO} each include an integral over the center of mass positions of X, O and XO respectively. Each of these partition sums should therefore scale with the volume of the cell, V. Defining

$$Z^*_{\mathrm{X}} \equiv \frac{Z_{\mathrm{X}}}{V}, \qquad Z^*_{\mathrm{O}} \equiv \frac{Z_{\mathrm{O}}}{V}, \qquad Z^*_{\mathrm{XO}} \equiv \frac{Z_{\mathrm{XO}}}{V}, \tag{1.9}$$

leads to

$$K_{\mathrm{d}} = \frac{Z^*_{\mathrm{O}} Z^*_{\mathrm{X}}}{Z^*_{\mathrm{XO}}}, \tag{1.10}$$

demonstrating that K_{d} is independent of the volume of the cell.

Now also the free energy of binding follows immediately as

$$\Delta F_{\mathrm{b}} \equiv -k_{\mathrm{B}} T \log\left(\frac{Z_{\mathrm{on}}}{Z_{\mathrm{off}}}\right) = -k_{\mathrm{B}} T \log\left(\frac{[\mathrm{X}]}{K_{\mathrm{d}}}\right). \tag{1.11}$$

Note that, by this definition, the binding energy is negative if the free energy of the bound state is lower than that of the unbound state; large negative binding energies thus correspond to strong binding. Clearly, ΔF_{b} depends on the concentration of X.

Competition between binding sites

We mentioned that, in general, DNA-binding molecules can bind to all DNA sites. This implies that any physiologically functional binding site has to compete with other possible sites. In the previous section we ignored this; here we estimate, using a simple model, the influence of this genomic background on the occupancy of a given site. We use these results in Chapters 2 and 3.

Every site i on the DNA has its own binding free energy ΔF_i. As in the previous section, we define this binding free energy such that it includes both the specific and unspecific binding modes. The partition sum of all states in which the molecule is bound at some site on the "background" DNA, Z_{back}, can therefore be written as:

$$Z_{\mathrm{back}} = \sum_i \mathrm{e}^{-\beta \Delta F_i}. \tag{1.12}$$

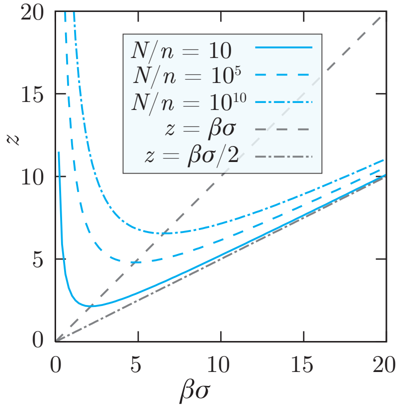


Figure 1.1: A DNA binding site for a certain molecule can only have a large occupancy if it binds the molecule much better than other sites on the DNA with which it competes. The figure shows the number of standard deviations z that the binding energy of the site needs to deviate from the mean binding energy in order to obtain an occupancy of $1/2$, as a function of the standard deviation σ itself. Plots are given for different values of N/n. For each N/n, the minimal value of z lies on the line $z = \beta\sigma$, as is shown graphically. For large σ, all lines asymptotically converge to $z = \beta\sigma/2$.

Here, as usual, $\beta \equiv 1/k_{\rm B}T$, where $k_{\rm B}$ is the Boltzmann constant and T is the temperature. In general, we do not know the binding energies of all these background states. However, we do know that these energies are the result of multiple contributions of individual contacts between the residues of the DNA-binding molecule and base pairs in the operator. It is therefore plausible that the probability distribution of binding energies in the background can be approximated by a normal distribution $\mathcal{N}(\beta\Delta F|\beta\mu, \beta\sigma)$, in which $\beta\mu$ and $\beta\sigma$ are the mean and the standard deviation. Given this distribution and the fact that the number of sites is very large ($N \approx 10^7$ in *E. coli*), we can estimate $Z_{\rm back}$ as follows:

$$Z_{\rm back} = N \left\langle \mathrm{e}^{-\beta\Delta F_i} \right\rangle_{\{\text{all sites } i\}} \approx N \int_{-\infty}^{\infty} \mathcal{N}(x|\beta\mu, \beta\sigma)\, \mathrm{e}^{-x}\, \mathrm{d}x \tag{1.13}$$

$$= N \int_{-\infty}^{\infty} \mathcal{N}(x| -\beta\mu, \beta\sigma)\, \mathrm{e}^{x}\, \mathrm{d}x = N \int_{0}^{\infty} \mathcal{L}(y| -\beta\mu, \beta\sigma)\, y\, \mathrm{d}y \tag{1.14}$$

$$= N \exp\left(-\beta\mu + \frac{(\beta\sigma)^2}{2}\right). \tag{1.15}$$

In the second line we introduced the coordinate transformation $y = \exp(x)$, and the lognormal distribution $\mathcal{L}\,(y| -\beta\mu, \beta\sigma)$. In the final line, we use the fact that the mean of a lognormal distribution characterized by parameters $-\beta\mu$ and $\beta\sigma$ is given by $\exp(-\beta\mu + (\beta\sigma)^2/2)$.

We again denote the number of molecules of X contained in the cell by n.

As we noted, many DNA-binding molecules are bound to the DNA most of the time; therefore, we use the simplifying approximation that X is *always* bound somewhere on the DNA. In that case, the occupancy of a site with binding energy ΔF_b, taking into account the competition with the background states, can be written as

$$p_\mathrm{on} = \frac{R}{1+R}, \tag{1.16}$$

$$R \equiv \frac{Z_\mathrm{on}}{Z_\mathrm{off}} = \frac{n\exp(-\beta\Delta F_\mathrm{b})}{Z_\mathrm{back}} = \frac{n}{N}\exp\left(-\beta(\Delta F_\mathrm{b}-\mu) - \frac{(\beta\sigma)^2}{2}\right). \tag{1.17}$$

This result shows that the occupancy of the operator does not only depend on the *mean* binding free energy of the background states, but also on the *variance* of the background free energies. In order to have a reasonable occupancy of the operator, we should have $R \approx 1$ (corresponding to $p_\mathrm{on} = 0.5$); now we compute the number of standard deviations $z(\sigma)$ that ΔF_b needs to deviate from the mean μ in order to obey $R = 1$:

$$z(\sigma) \equiv \left|\frac{\Delta F_\mathrm{b}-\mu}{\sigma}\right| = \frac{1}{\beta\sigma}\log\left(\frac{N}{n}\right) + \frac{\beta\sigma}{2}. \tag{1.18}$$

We thus establish that, given an occupancy of one half, z depends strongly on the standard deviation. In Fig. 1.1 we plot $z(\sigma)$ for several values of N/n. For very low or very high standard deviations, z becomes very large; between these limits, z goes through a minimum. It can be shown that, at this point,

$$z_\mathrm{min} = \beta\sigma_\mathrm{min} = \sqrt{2\log\left(\frac{N}{n}\right)}. \tag{1.19}$$

For $N \approx 10^7$ and $n \approx 100$, we arrive at $z_\mathrm{min} \approx 4.8$. Apparently, in this simplified model, a binding site should have a binding energy that is at least 4.8 standard deviations below average in order to have an occupancy of one half. This is sufficient only if the standard deviation has the optimal value of $4.8\,k_\mathrm{B}T$; otherwise, the binding energy needs to be even more exceptional. If we relax our assumption that X is always on the DNA, this only leads to even higher estimates.

Indeed, in *E. coli*, the standard deviation of binding energies of TFs to the background is of the order of $5\,k_\mathrm{B}T$ (see for instance Mustonen and Lassig (2005) for estimates of the distribution for CRP binding energies in *E. coli*).

Of course, the exact functional forms and values derived above do depend on our assumption that the binding free energies of the background states are normally distributed. However, the general message is much less sensitive to these assumptions. This message is: since the background partition sum is proportional to the mean of the Boltzmann factors $\exp(-\beta\Delta F)$ over the sites, the partition sum is dominated by the lowest energies (best binders). As a result, the occupancy depends strongly on the variance of the energies.

Binding energy

In the case of sequence-specific binding, the binding free energy of a site depends on its sequence $\vec{s}$. Several experiments on TFs have suggested that, approximately, the individual nucleotides $s_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ contribute independently and additively to the binding free energy. This is only approximately true (*e.g.* Fields *et al.* (1997) and Roulet *et al.* (2002)), but nevertheless allows for a good approximation (Benos *et al.*, 2002). We can therefore write

$$\Delta F(\vec{s}) = \sum_{i=1}^{l} E_i(s_i), \tag{1.20}$$

where l is the length of the binding site in base pairs, which is typically between 6 and 20 bp, and $E_i(b)$ denotes the contribution of base pair b at position i to the total free energy of binding.

For a few DNA-binding molecules, the function $E_i(b)$ has been measured using detailed mutational studies (*e.g.* Fields *et al.* (1997)). In other cases, these energies can be estimated using known binding sequences. Qualitatively, the estimation procedure can be understood as follows. The set of known binding sequences is different from a random set of sequences in that it only contains sequences that have a (relatively) high affinity for the TF. This implies that the mean binding energy of the sequences in the set of known binding sequences must be lower than expected at random. (Remember that we defined better binders to have a lower binding energy.) Consequently, one should expect that, in the set of known binding sequences, at any given position i, base pairs b that contribute strongly to the binding energy (*i.e.* for which $E_i(b)$ is low) are statistically over-represented. Conversely, if, in a large set of binding sequences, a base b occurs more often at position i than expected at random, this is evidence for a low value of $E_i(b)$.

This qualitative argument can be turned into a quantitative one (Berg and von Hippel, 1987; Djordjevic *et al.*, 2003; van Nimwegen, 2007). Assuming only that the ensemble of binding sequences for a certain TF is characterized by an average binding energy $\langle \Delta F \rangle$, the maximum entropy principle (Jaynes and Bretthorst, 2003) predicts that the occurrence frequency of a pair (i, b) in the set depends exponentially on $E_i(b)$ (van Nimwegen, 2007). Alternatively, the same conclusion follows if one assumes that the energies of binding sites are all in a rather narrow range around some value $\langle \Delta F \rangle$; in this case, the argument is equivalent to the derivation of the Boltzmann factor in the canonical ensemble. The energies $E_i(b)$ can hence be related to the logarithm of the occurrence frequencies. A more profound Bayesian framework can be used to compute the likelihoods of energy values given the set of binding sequences; this framework is reviewed in van

Nimwegen (2007).

Mandel-Gutfreund and Margalit (1998) estimated the contribution of each amino-acid−nucleotide interaction to the binding free energy between TFs and DNA, using the coordinates of 53 solved protein–DNA co-crystals. To this end, they counted the number of times a certain amino acid was in close contact with a base pair, including hydrogen bonds and hydrophobic interactions. Such counts can again be used to estimate the relative energies of such interactions, leading to a 20×4 matrix estimating the interaction energies for each amino-acid−base-pair combination. We use this matrix in Chapters 2 and 3 to model TF–DNA interactions in a coarse-grained manner.

1.1.2 Regulation of transcription initiation in prokaryotes

We now turn to the process of transcription regulation. We limit our discussion to mechanisms in prokaryotes. In eukaryotes, transcription regulation is known to be considerably more complex. Nevertheless, most of the basic mechanisms discussed below are expected to play a role in eukaryotes as well.

Transcription is the process that copies the information on the DNA into RNA molecules. This multistep process is catalyzed by RNA polymerase, which is a multi-subunit enzyme. Its binding site, the (core) promoter, is mainly defined by two stretches of six nucleotides located 10 and 35 base pairs upstream of the start of transcription; these sites are called the -10 and -35 hexamers. When bound to the promoter, RNAp first forms a complex called the closed complex. Next, RNAp locally catalyzes the melting of a stretch of DNA about 12 to 16 base pairs long. At this point, the resulting RNAp–DNA complex is called the open complex. The polymerase subsequently initiates transcription. At first, the so-called initiation complex transcribes at a low rate; but after the transcription of about 10 base pairs RNAp releases one of its components (the σ factor, which was responsible for the initial recognition of the promoter) and changes conformation. The resulting elongation complex then starts transcribing the DNA at a much higher rate: about 50 bp per second (Browning and Busby, 2004; Alberts *et al.*, 1994; Baumberg, 1999).

In theory, each step in this process could be influenced by other proteins interacting with the transcription machinery; such interactions can therefore be used to *regulate* the rate of transcription. But most of the transcription factors that have been studied to date mainly influence the very first steps of the process: the recognition of the promoter by the RNAp–σ-factor holoenzyme and the initiation process. The models used and developed in this thesis to describe and understand transcription regulation all deal with processes of this kind.

The first physical-chemical model of transcription regulation were introduced in the early eighties by Gary K. Ackers and Madeline A. Shea (Ackers *et al.*, 1982; Shea and Ackers, 1985). Their work focussed on the regulation of the genes of the bacteriophage λ, which is a virus infecting *E. coli*. Even though quite some work has been done in this field since then, the main assumptions underlying the

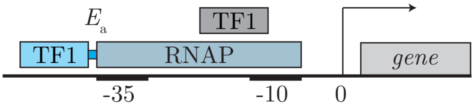


Figure 1.2: Cartoon illustrating the standard mechanisms for regulation of transcription initiation. In order to start transcription, RNA polymerase needs to bind to the core promoter, consisting of (at least) the -10 and -35 consensus hexamers. Transcription factors binding to sites overlapping with the core promoter (these sites are indicated in gray) can reduce the occupancy of the promoter by RNAp, thereby lowering the transcription rate. On the other hand, TFs binding close to the core promoter (in blue) can recruit RNAp to its site. Often, the recruitment is due to a direct contact between the activator and one of the sub-domains of RNAp. The energy associated with this contact, E_{a}, effectively leads to an increased affinity of RNAp for its site if the activator is bound.

models have not changed (Bintu *et al.*, 2005a,b; Buchler *et al.*, 2003).

Most models of transcription regulation rely on the assumption that the rate of transcription of a given transcription unit is proportional to the equilibrium fractional occupancy of the core promoter, *i.e.* the fraction of time an RNAp molecule is bound to the promoter. This is a good approximation if the equilibration time of the RNAp–promoter binding is very short compared to the rate at which isomerization of RNA polymerase from the closed to the open complex takes place. This is usually the case, and the transcription rate A of the transcription unit in the absence of any regulation is then simply given by

$$A \propto p_{\mathrm{on}} = \frac{Z_{\mathrm{on}}}{Z_{\mathrm{on}} + Z_{\mathrm{off}}} = \frac{q_{\mathrm{p}}}{1 + q_{\mathrm{p}}} = \frac{1}{1 + q_{\mathrm{p}}^{-1}}, \tag{1.21}$$

given the affinity q_{p} of RNAp for the promoter (see Equation 1.4 and 1.16). This shows that the transcription rate of the gene is a Hill function of the concentration of RNAp, with a Hill coefficient equal to 1.

By binding to the DNA on specific sites, transcription factors can influence the fractional occupancy of the core promoter by RNAp and hence the transcription rate of the transcription unit. This influence can have a positive sign, in which case we speak of *activation*, or a negative one, which is called *repression*. Below, we introduce the main mechanisms of repression and activation and describe them quantitatively.

Repression

The most direct mechanism to repress transcription is by steric hindrance (Müller-Hill, 1998). In this case, a TF binds to a site close to or overlapping with the core promoter and thus directly impedes the binding of RNAp to the promoter (see Fig. 1.2). If we denote the affinity of the repressor for its operator by q_{r} and assume that, at any time, either the repressor or RNAp can be bound, but not both, then we can adjust Equation 1.21 to describe this repression mechanisms formally:

$$\begin{aligned} Z_{\text{off}} &= 1 + q_{\text{r}}, \\ Z_{\text{on}} &= q_{\text{p}}, \\ A \propto p_{\text{on}} &= \frac{Z_{\text{on}}}{Z_{\text{on}} + Z_{\text{off}}} = \frac{1}{1 + q_{\text{p}}^{-1}(1 + q_{\text{r}})}. \end{aligned} \tag{1.22}$$

Clearly, the repression becomes more significant with increasing repressor concentration. If the repressor concentration vanishes ($q_{\text{r}} = 0$), we recover Equation 1.21.

It is also possible to indirectly repress a gene by interfering with an activation system. We call such a process anti-activation and discuss this and other indirect repression mechanisms in Chapter 2. Other repression systems involve, for instance, looping of DNA; these are not treated here.

Activation

Like repression, activation can be achieved in several ways. The most standard method is recruitment of RNAp to the core promoter (see Fig. 1.2). Here, the activator binds to a binding site located such that it does not hinder the binding of RNAp. Instead, the bound activator touches one of the subunits of RNAp (*e.g.* the αCTD subunit or the σ factor) when it is bound, effectively stabilizing the bound state of RNAp. Writing the affinity of the activator as q_{a} and the contact energy between RNAp and the activator as E_{a}, we can again compute the occupancy of the core promoter:

$$\begin{aligned} Z_{\text{off}} &= 1 + q_{\text{a}}, \\ Z_{\text{on}} &= q_{\text{p}} \left(1 + \omega' q_{\text{a}}\right), \\ A \propto p_{\text{on}} &= \frac{Z_{\text{on}}}{Z_{\text{on}} + Z_{\text{off}}} = \frac{1}{1 + q_{\text{p}}^{-1}\mathcal{R}}, \end{aligned} \tag{1.23}$$

where we made use of the definitions

$$\omega' \equiv \mathrm{e}^{-\beta E_{\text{a}}}, \tag{1.24}$$

$$\mathcal{R} \equiv \frac{1 + q_{\text{a}}}{1 + \omega' q_{\text{a}}}. \tag{1.25}$$

Evidently, if q_{a} is large, $\mathcal{R} \approx 1/\omega'$, so that effectively the activator increases the affinity q_{p} of RNAp with a factor ω'. In the other limit, where $q_{\text{a}} = 0$, we again retrieve Equation 1.21.

Activation can also occur by other mechanisms than direct recruitment of RNAp. A notable example is the transcription factor MetR, which, at the *merT* promoter of *E. coli*, manages to activate transcription by twisting the DNA between the recognition hexamers of the core promoter. This way, it properly aligns these hexamers for RNAp binding (Summers, 1992). Effectively, mechanisms such as these can be treated by the same equations as above; E_a should then be interpreted as the difference between the free energy of the state where both the activator and RNAp are bound and the sum of the binding energies of both molecules independently.

In Chapter 2 several indirect activation mechanisms are discussed, including anti-repression.

Complex regulation and cooperativity

Most of the best-known promoters are relatively simple. They are regulated by one or two transcription factors and each of these factors binds to few sites at the promoter. As we point out in Chapter 2, an analysis of the known binding sites shows that this is not typical; many promoters contain multiple binding sites of several transcription factors. Clearly, the expression of certain genes is determined by many signals. This indicates that the “decisions” made by the bacterium can be rather complex. In almost all cases, the multi-dimensional response functions of these promoters have not been measured in detail.

One important reason why transcription factors often have multiple binding sites close to each other, is because this allows the transcription rate of the target gene to respond to changes in the TF concentration in a more stepwise fashion (less gradual). This requires that the TF molecules bind *cooperatively* to these binding sites. One speaks of cooperative binding if the binding of one molecule facilitates the binding of another one. This is often due to a direct contact between the bound TFs, but it is also possible that the binding of one of the molecules changes the conformation of the local DNA such that the affinity of the other molecule increases.

As an example, we quantify the rate of transcription as a function of TF concentration for a simple cooperative repression system. We assume that a TF binds cooperatively to two operators with affinity q_r, one of which overlaps with the core promoter. Without cooperativity, the Boltzmann factor of the state in which both operators are occupied would be q_r^2; with cooperativity, this Boltzmann factor is increased by a certain factor we call ω; typical values of ω are 0–100. Then, the transcription rate can be computed as

$$\begin{aligned} Z_\mathrm{off} &= 1 + 2q_\mathrm{r} + \omega q_\mathrm{r}^2, \\ Z_\mathrm{on} &= q_\mathrm{p}\left(1 + q_\mathrm{r}\right), \\ A \propto p_\mathrm{on} &= \frac{Z_\mathrm{on}}{Z_\mathrm{on} + Z_\mathrm{off}} = \frac{1}{1 + q_\mathrm{p}^{-1}\mathcal{R}}, \end{aligned} \tag{1.26}$$

with

$$\mathcal{R} \equiv \frac{1 + 2q_\mathrm{r} + \omega q_\mathrm{r}^2}{1 + q_\mathrm{r}}. \tag{1.27}$$

Due to the cooperativity, the regulatory factor $\mathcal{R}$ now contains a large quadratic term in the numerator; this results in a steeper response of p_on to the repressor

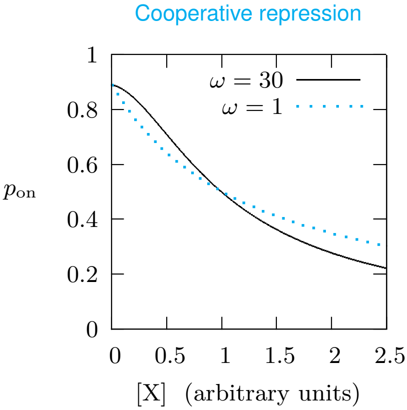


Figure 1.3: Cooperative repression. Here, a system is considered where a repressor can bind to two operators, one of which overlaps with the core promoter. The two lines show p_{on} as a function of the repressor concentration (in arbitrary units) for two values of the cooperativity factor ω, which quantifies the interaction between repressors bound to the operators. Both curves have been scaled such that they reach $p_{\mathrm{on}} = 0.5$ at $[\mathrm{X}] = 1$. Clearly, the curve for $\omega = 30$ (a realistic value) shows a steeper switching than the one for $\omega = 1$ (corresponding to no cooperativity).

concentration. Fig. 1.3 shows this effect. Additional binding sites and fine-tuning of their affinities can increase the steepness even more; we show examples in Chapter 2.

Some promoters behave as logic gates

Bacteria often have to make logical decisions. A famous example is the regulation of the sugar utilization system in the bacterium *E. coli* (Müller-Hill, 1996). In order to uptake and digest different sugars, such as glucose, lactose, galactose and arabinose, *E. coli* needs to produce particular sets of proteins that catalyze the required metabolic processes. However, these sugars are not always present in the environment. As the production of the proteins requires an investment in terms of energy and other resources, it would be quite inefficient to express these genes constitutively. Hence, *E. coli* decides when to transcribe the genes and when not to, depending on the availability of the sugars.

In order to make an informed decision, *E. coli* measures the concentrations of various sugars in the environment. This is achieved by particular TFs that each specifically bind (derivatives of) one of the sugars. The concentrations of the TF–ligand complexes inside the cell therefore reflect the availability of the particular sugars in the environment. The genes required for the uptake and digestion of the sugars are directly regulated by these TFs. Since the DNA binding affinities of the TFs change upon ligand binding, the transcription rates of the sugar genes are dynamically adjusted to the concentrations of the sugars.

However, glucose is *E. coli*'s preferred source of energy, because it allows for the highest instantaneous growth rate. The genes coding for the uptake and degradation of the other sugars are therefore also turned off if there is sufficient glucose in the environment. For instance, the lactose operon is transcribed at a high rate only if there is lactose and no glucose in the environment. This illustrates that the promoter of the *lac* genes effectively functions as a *logic gate*, that is, as a device that integrates several input signals (sugar availability reflected in TF concentrations) to produce one output signal (the transcription rate or expression level of the regulated transcription unit) according to the Boolean logic function ANDN (A **AND** **N**ot B). In general, many of the decisions taken by cells can be categorized using the language of Boolean logic (Buchler *et al.*, 2003; Kuhlman *et al.*, 2007). In Table 2.1 the names of the different logic gates with two inputs are summarized.

It should be kept in mind, however, that descriptions in terms of Boolean logic are a simplification. In many cases, the actual response of a gene is more complex, and with reason. In fact, the *lac* system is again a good example. Detailed measurements show that, if the *lac* operon is in the "off" state, it is still being transcribed at a low but non-zero basal rate. For the functioning of the system this is crucial: in order to measure if lactose is present in the environment, it is critical that at all times at least some lactose could be taken up by the cell. This requires that a small amount of the enzyme LacY, the transport protein (permease) responsible for the uptake of galactosides, is being synthesized constitutively.

In Chapter 2 we study complex mechanisms of transcription regulation and show how overlapping binding sites and cooperativity can be combined to construct all transcriptional logic gates with two inputs.

1.2 Evolutionary population genetics

Any living system is the result of a long process of evolution. Evolutionary population genetics is the field that studies how the genetic composition of a population of organisms evolves under evolutionary forces such as natural selection, stochastic reproduction (drift), migration and sexual selection. In several chapters we try to understand the organization of the genome from an evolutionary perspective; then the field of evolutionary population genetics becomes crucial.

Many of the important results of population genetics date back to the 1950s and 1960s. Nowadays, it is an advanced field of mathematical biology. Through a multitude of models the influence of the many relevant biological parameters have been studied. Obviously, this is not the place to review the field. Instead, below we derive the basic results that are relevant in the remainder of this thesis, based on the well-known Wright–Fisher model. For an excellent and accessible introduction we refer to the lecture notes of Joseph Felsenstein, which are freely available online[1] or to any of the many textbooks available.

In Chapter 5 we present simulation results exploiting an expression for the probability of fixation of a mutant in a population of haploids, as a function of the fitness effect of the mutation and the population size. This relation can be derived in several ways. Below, we introduce a version of the so-called Wright–Fisher model and then derive the fixation probability using the diffusion approximation first used by Kimura (1962).

1.2.1 A Wright–Fisher model with selection

The Wright–Fisher model is a simplified model for the population genetics of populations consisting of haploid organisms. It assumes that the population has a fixed size N and reproduces in discrete, non-overlapping generations. Each individual organism o in the population, with genome g_o, has a fitness F_o, which is defined as the expectation value of the number of offspring that the organism will have in the next generation. As the population size is fixed, the average fitness has to be 1. Therefore it is useful to introduce the selection coefficient s_o as $s_o \equiv F_o - 1$.

Given the population in generation t, the population of the next generation, $t+1$, is chosen by stochastically sampling genomes from the population in generation t. The probability that a randomly selected individual from generation $t+1$ is a child of organism o from generation t is chosen to be proportional to the fitness of o:

$$p_o = \frac{1+s_o}{\sum_{o'} 1+s_{o'}} = \frac{1+s_o}{N}. \tag{1.28}$$

The number of offspring of o, denoted by k, is therefore a random variable K_o with a binomial probability distribution:

$$P_{\,\mathrm{bin}}(K_o = k \mid N) = \binom{N}{k} p_o^k (1-p_o)^{N-k}. \tag{1.29}$$

The mean of this distribution is $\mu_o = Np_o = 1 + s_o$, which shows that indeed the expectation value of the number of offspring of o is equal to its fitness. The variance of the distribution is $\sigma_o^2 = Np_o(1-p_o)$.

1.2.2 Fixation probability

In a finite population, any mutant will on the long run either go extinct, or take over the population. In the latter case, we say that the mutant becomes "fixed" in the population. The probability for a mutation to eventually become fixed depends on the selection coefficient of the mutant: clearly, mutations that increase the fitness of an organism are more likely to become fixed than ones that decrease

---

[1] Joseph Felsenstein, *Theoretical Evolutionary Genetics*
http://evolution.genetics.washington.edu/pgbook/pgbook.html

it. We now derive the fixation probability for a mutant in the framework of the Wright–Fisher model, using the diffusion approximation introduced by Kimura (1962).

We assume that, due to a mutation in one of the individuals, the population now contains two versions of a certain locus in the genome: `A` and `a`. Initially, in generation t_0, the population contains i individuals with the new, mutated version `A` of the locus. These organisms have selection coefficient $s_{\mathtt{A}}$. The remaining $(N-i)$ individuals have selection coefficient $s_{\mathtt{a}} = 0$.

The number of individuals with `A` is likely to change in the next generation. Assuming i is given, the number of individuals bearing `A` in the next generation, j, is a random variable with a probability distribution $P(j|i)$. This distribution is determined by Equation 1.29, but we do not compute it yet. We do note however that, if $s_{\mathtt{A}}$ is small, $P(j|i)$ is expected to be very small except when $|i-j|$ is small; we need this below.

We introduce the probabilities u_i that the locus `A` becomes fixed in the population given that it has i copies in generation t_0. For these probabilities the following relation holds:

$$u_i = \sum_j P(j|i) u_j. \tag{1.30}$$

Diffusion approximation

It is convenient to now introduce the following notation. The variable $p \equiv i/N$ is the fraction of population that carries `A`. We use $\Delta p \equiv (j-i)/N$ to denote the change in p after one generation, and $U(p) \equiv u(i/N)$ as the fixation probability, which is a function of the initial fraction of the population bearing `A`. Then Equation 1.30 can be rewritten as:

$$U(p) = \sum_{\Delta p} P_p(\Delta p) U(p + \Delta p), \tag{1.31}$$

where $P_p(\Delta p)$ is the probability of a change Δp after one generation given the initial fraction is p.

Now we introduce an approximation of Equation 1.31 by considering $U(p)$ as a continuous function of a now continuous variable p. Hence, summations are replaced by integrals. Also, we use the fact that $P_p(\Delta p)$ is small except when Δp is small to approximate $U(p+\Delta p)$ in Equation 1.31 by a second order Taylor approximation:

$$U(p) \approx U(p) \int_\epsilon P_p(\epsilon)\,\mathrm{d}\epsilon + \frac{\partial U(p)}{\partial p} \int_\epsilon P_p(\epsilon)\,\epsilon\,\mathrm{d}\epsilon + \frac{1}{2}\frac{\partial^2 U(p)}{\partial p^2} \int_\epsilon P_p(\epsilon)\,\epsilon^2\,\mathrm{d}\epsilon. \tag{1.32}$$

Note that $\int_\epsilon P_p(\epsilon)\,\mathrm{d}\epsilon = 1$. The factor $\int_\epsilon P_p(\epsilon)\,\epsilon\,\mathrm{d}\epsilon$ should be identified as the mean change in p after one generation given the initial state p and will be called $M(p)$. Similarly, the factor $\int_\epsilon P_p(\epsilon)\,\epsilon^2\,\mathrm{d}\epsilon$ is the second moment of the distribution

$P_p(\epsilon)$, and is called $V(p)$. This simplifies relation 1.32 considerably:

$$M(p)\frac{\partial U(p)}{\partial p} + V(p)\frac{1}{2}\frac{\partial^2 U(p)}{\partial p^2} \approx 0. \tag{1.33}$$

This can be recognized as a diffusion equation with drift that can be solved formally:

$$U(p) = \frac{\int_0^p G(x)\,\mathrm{d}x}{\int_0^1 G(x)}, \tag{1.34}$$

where

$$G(x) \equiv \exp\left(-2\int_c^x \frac{M(y)}{V(y)}\,\mathrm{d}y\right). \tag{1.35}$$

The constant c is a meaningless factor that cancels in Equation 1.34.

Diffusion approximation applied to the Wright–Fisher model

In order to finish the derivation of the fixation probability as a function of the selection coefficient $s_{\mathtt{A}}$ of the mutant and the population size N, we have to derive $M(p)$ and $V(p)$ for the Wright–Fisher model. For this it is useful to introduce the probability q that a randomly selected individual in generation $t_0 + 1$ has genotype `A`:

$$q = \frac{i(1+s_{\mathtt{A}})}{i(1+s_{\mathtt{A}}) + (N-i)(1+s_{\mathtt{a}})} = \frac{p(1+s_{\mathtt{A}})}{p(1+s_{\mathtt{A}}) + (1-p)(1+s_{\mathtt{a}})}. \tag{1.36}$$

This means that the number of individuals carrying `A` in generation $t_0 + 1$ is binomial:

$$P(j|i) = \binom{N}{j} q^j (1-q)^{N-j}. \tag{1.37}$$

The mean of this distribution is $\langle j \rangle = Nq$ and the variance is $\sigma^2 = Nq(1-q)$. Therefore the mean change $M(p) = M(i/N)$ can be derived as

$$M(p) = \frac{1}{N}\left(\sum_j (j-i)P(j|i)\right) = \frac{1}{N}(Nq - i) = \frac{s_{\mathtt{A}}p(1-p)}{1+s_{\mathtt{A}}p}. \tag{1.38}$$

The second moment of the change, $V(p)$, can be expressed as

$$V(p) = \frac{1}{N^2}\left(\sum_j (j-i)^2 P(j|i)\right) = \frac{1}{N^2}\left(\langle j^2 \rangle - 2i\langle j \rangle + i^2\right). \tag{1.39}$$

Using the fact that the variance σ^2 of $P(j|i)$ obeys $\sigma^2 = \langle j^2 \rangle - \langle j \rangle^2$, we can write

$$V(p) = \frac{1}{N^2}\left(\sigma^2 + \langle j \rangle^2 - 2i\langle j \rangle + i^2\right) \tag{1.40}$$

$$= \frac{q(1-q)}{N} + q^2 - 2pq + p^2. \tag{1.41}$$

This allows one to compute the ratio

$$\frac{M(p)}{V(p)} = \frac{Ns_{\mathtt{A}}(1+ps_{\mathtt{A}})}{1+s_{\mathtt{A}}+N(1-p)ps_{\mathtt{A}}^2} \tag{1.42}$$

$$= Ns_{\mathtt{A}} - N(1-p)s_{\mathtt{A}}^2 + \mathcal{O}(s_{\mathtt{A}}^3). \tag{1.43}$$

In the second line we used a Taylor expansion around $s_{\mathtt{A}} = 0$; assuming that $s_{\mathtt{A}} \ll 1$, we can ignore the terms of order $s_{\mathtt{A}}^2$ and higher so that $M(p)/V(p) \approx Ns_{\mathtt{A}}$. If we insert this into Equation 1.34 and perform the integration, we arrive at:

$$U(p) = \frac{1-\mathrm{e}^{-2Ns_{\mathtt{A}}p}}{1-\mathrm{e}^{-2Ns_{\mathtt{A}}}}. \tag{1.44}$$

This is the result we were aiming for: it expresses the probability of fixation of a mutation `A` with selection coefficient $s_{\mathtt{A}}$, given that currently a fraction p of the population carries the mutation. A special case of this equation is the situation when exactly one mutant is introduced to the population, so that $i = 1$ and $p = 1/N$. We call the fixation probability of that mutant, as a function of $s_{\mathtt{A}}$, $P_{\mathrm{K}}(s_{\mathtt{A}})$:

$$P_{\mathrm{K}}(s_{\mathtt{A}}) = U(1/N) = \frac{1-\mathrm{e}^{-2s_{\mathtt{A}}}}{1-\mathrm{e}^{-2Ns_{\mathtt{A}}}}. \tag{1.45}$$

We will use this result several times in the subsequent chapters; it is often referred to as the Kimura–Ohta fixation probability function for haploid organisms.

1.2.3 Limits

The Kimura–Ohta fixation probability derived in the previous section can be simplified in several limiting cases:

- ✓ If Ns is large and negative, the fixation probability approaches zero exponentially:

  $$P_{\mathrm{K}}(s) = \frac{1-\mathrm{e}^{-2s}}{1-\mathrm{e}^{-2Ns}} \approx \mathrm{e}^{-2|s|(N-1)}, \qquad (\text{for } Ns \ll -1). \tag{1.46}$$

  As the population size of bacterial populations is typically very large (estimates are at least of the order of 10^5 to 10^7), this is already the case for rather small values of $|s|$. In other words: deleterious mutations usually do not get fixed.

- ✓ If $|Ns| \ll 1$, or in other words, if the mutation is nearly neutral, the exponents in the equation can be approximated up to linear order, so that

  $$P_{\mathrm{K}}(s) = \frac{1-\mathrm{e}^{-2s}}{1-\mathrm{e}^{-2Ns}} \approx \frac{s}{Ns} = \frac{1}{N}, \qquad (\text{for } |Ns| \ll 1). \tag{1.47}$$

  This makes sense, as in the neutral case each individual in the population should have an equal probability to become fixed in the population; this directly leads to a fixation probability of $1/N$.

- ✓ If $1 \ll sN \ll N$ (weak selection), then Equation 1.45 reduces to:

  $$P_{\mathrm{K}}(s) = \frac{1-\mathrm{e}^{-2s}}{1-\mathrm{e}^{-2Ns}} \approx s, \qquad (\text{for } 1 \ll sN \ll N). \tag{1.48}$$

  So: if s is small but non-zero, the fixation probability is approximately equal to the fitness difference s.

- ✓ If $Ns \gg 1$,

  $$P_{\mathrm{K}}(s) = \frac{1-\mathrm{e}^{-2s}}{1-\mathrm{e}^{-2Ns}} \approx 1-\mathrm{e}^{-2s}, \qquad (\text{for } Ns \gg 1). \tag{1.49}$$

  This shows that the fixation probability for a considerably advantageous mutation approaches 1 exponentially.

1.2.4 Fixation rates

Given these fixation probabilities and mutation rates, we can derive the rates at which certain mutations are fixed in the population. Suppose that a particular mutation occurs with rate μ in each organism in the population, and results in a selection coefficient s. Then the rate at which this mutation occurs in the total population is μN, and the rate at which it is fixed in the population is therefore given by

$$\mu_f(s) = \mu N P_{\mathrm{K}}(s). \tag{1.50}$$

The approximations from the previous section can be applied again. For neutral, weakly selected and strongly selected mutations, the rates can be approximated by, respectively:

$$\mu_f(s) \approx \begin{cases} \mu & \text{if } |Ns| \ll 1, \\ \mu N s & \text{if } 1 \ll Ns \ll N, \\ \mu N(1-\mathrm{e}^{-2s}) & \text{if } Ns \gg 1. \end{cases} \tag{1.51}$$

The rate at which neutral mutations become fixed in the population apparently does not depend on the population size. On the other hand, fitness differences get more and more significant with increasing population size.

Chapter 2

Transcription regulation by transcription factor modules

The designs of both eukaryotic and prokaryotic *cis*-regulatory regions are usually highly complex. They frequently consist of both repetitive and overlapping transcription factor binding sites. To unravel the design principles of such promoter architectures, we have designed *in silico* prokaryotic transcriptional logic gates with predefined input–output relations using an evolutionary algorithm. The resulting *cis*-regulatory designs are often composed of modules that consist of tandem arrays of binding sites to which the transcription factors bind cooperatively. Moreover, these modules often overlap with each other, leading to competition between them. Our analysis thus identifies a new signal integration motif that is based upon the interplay between intra-modular cooperativity and inter-modular competition. We show that this signal integration mechanism drastically enhances the capacity of *cis*-regulatory domains to integrate signals. Our results provide a possible explanation for the complexity of promoter architectures and could be used for the rational design of synthetic gene circuits.

2.1 Introduction

Cells continually have to make logical decisions. Many of these decisions are taken in the *cis*-regulatory regions of genes, which can function as analog implementations of logic gates (Buchler *et al.*, 2003; Istrail and Davidson, 2005; Yuh *et al.*, 1998). A classical example is the lactose system in the bacterium *Escherichia coli*, where the *lac* operon is strongly expressed only if the concentration of active CRP, due to the absence of glucose, is high and that of active LacI, due to the presence of lactose, is low. This network can be interpreted as a logic gate with two input signals, namely the concentrations of the transcription factors (TFs) CRP and LacI, and one output signal, the expression level of the operon; indeed, this gate could be classified as an ANDN gate. The lactose system has been studied in much detail both theoretically and experimentally and is now fairly well understood (Jacob and Monod, 1961a; Müller-Hill, 1996; Ptashne and Gann, 2002; Setty *et al.*, 2003). However, even in prokaryotes, many *cis*-regulatory regions are much more complex than that of the lac operon. Fig. 2.1, taken directly from the EcoCyc database v9.5 (Keseler *et al.*, 2005), shows four typical examples. The *cis*-regulatory regions often contain long tandem arrays of TF binding sites. Moreover, many TFs can both activate and repress the same operon. Perhaps most strikingly, TF binding sites often overlap with one another.

We have performed a statistical analysis of the importance of repetitive and overlapping binding sites in *E. coli*, based on the EcoCyc database (Keseler *et al.*, 2005). The results are shown in Fig. 2.2. We find that 37% of the TF–operon interactions are mediated by more than one binding site and 39% of the binding sites overlap with at least one other site. The question arises what kind of functionality these complex structures can convey (Müller-Hill, 1998). Here we present theoretical results that suggest that these intricate structures are a consequence of the functional requirement of *cis*-regulatory domains to integrate signals. Our results identify a new mechanism for signal integration during transcriptional regulatory control, which is based upon the interplay between cooperative binding of TFs to adjacent sites and competitive binding of TFs to overlapping sites.

To elucidate the origin of the complicated structures shown in Fig. 2.1, we have adopted a novel approach. Using an evolutionary algorithm (Francois and Hakim, 2004), we have designed prokaryotic *cis*-regulatory domains with predefined functions *in silico*. In our approach, no specific promoter architectures are specified *a priori*: the space of possible architectures is sampled in an unbiased manner. This makes it possible to discover new architectures and find the optimal design for a *cis*-regulatory domain that is consistent with a required function. The design principles of these architectures are then extracted *a posteriori*. As we will show below, this approach has allowed us to reveal new design principles of transcriptional regulation, which would have been difficult to obtain using the more conventional approach of studying particular architectures.

In order to design prokaryotic *cis*-regulatory domains, we have developed a

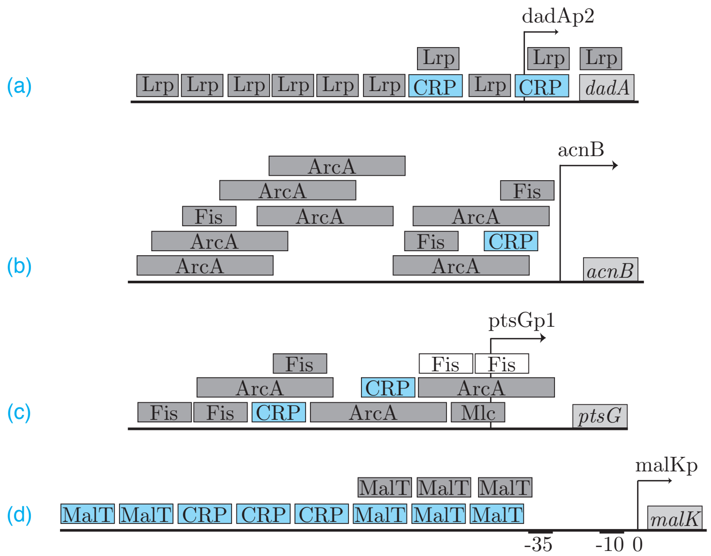


Figure 2.1: Examples of complex *E. coli* promoters. Figs. (a)–(c) are copied from the EcoCyc database (Keseler *et al.*, 2005). Fig. (d) is described in Richet and Sogaard-Andersen (1994). Blue blocks denote TF binding sites that have an activating effect; gray blocks denote repressor sites. White sites can both activate and repress transcription. Note that repetitive and overlapping binding sites occur frequently. Understanding this kind of promoters requires detailed quantitative information about binding affinities and interactions.

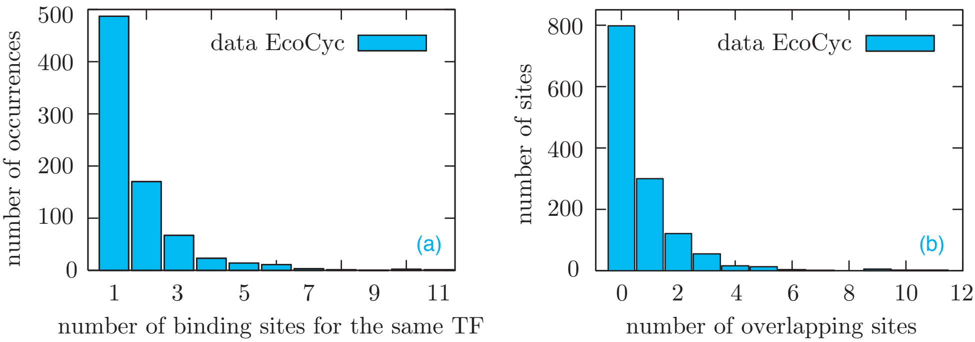


Figure 2.2: (a) Histogram of the number of binding sites responsible for each interaction between a TF and an operon, according to the EcoCyc database (Keseler *et al.*, 2005). Note that multiple sites are common; *e.g.*, the *cis*-regulatory region of *focA*, has as many as 11 binding sites for NarL. (b) Histogram of the number of binding sites overlapping with each binding site (Keseler *et al.*, 2005). For example, bin 1 with hight 300 should be interpreted as: there are 300 binding sites that overlap with exactly 1 other binding site. Overlap is common; some ArcA sites in the *sodA* regulatory region overlap with as many as 11 sites.

novel model of prokaryotic transcriptional regulation, in which the input–output relation of an operon is deduced from the amino-acid sequences of the TFs and the base-pair sequence of the *cis*-regulatory region of the operon. To go from sequence to network function (which in this case is given by the input–output relation of the *cis*-regulatory region), the model contains the following key ingredients (see Fig. 2.3):

1. Each TF can bind anywhere on the *cis*-regulatory region; conversely, this directly implies that to a given location, all TFs can bind;
2. The affinity of a TF for a certain location is determined by its DNA sequence and the amino-acids in the DNA-binding domain of the TF; the binding energies of the amino-acid−base-pair contacts are extracted from a matrix that is based on crystallographically solved protein–DNA complexes (Mandel-Gutfreund and Margalit, 1998);
3. TFs cannot overlap in space, even though binding sites *can* overlap along the DNA; TFs thus compete with each other for binding to overlapping sites;
4. TFs that bind close to each other on the DNA exhibit a cooperative interaction (Ptashne and Gann, 2002);
5. The transcription rate of operons is controlled via the mechanism of “regulated recruitment”, meaning that TFs function by stimulating or hindering the binding of RNA polymerase (RNAp) to the DNA (Ptashne and Gann, 2002). Even though this is the dominant mechanism in prokaryotes, we note that several alternative mechanisms are used as well. To describe the input–output relationship for an operon quantitatively, we employ the statistical mechanical approach developed by Shea and Ackers (Shea and Ackers, 1985) and Buchler *et al.* (Buchler *et al.*, 2003).

This model makes it possible to design *cis*-regulatory domains by performing rounds of mutation and selection in an evolutionary algorithm. Because the input–output relation is completely specified at the microscopic level of the amino-acid sequences of the TFs and the base-pair sequences of the *cis*-regulatory regions, new architectures can be obtained by introducing mutations at the microscopic (sequence) level, while the structures are selected at the macroscopic level of the input–output relation. Importantly, neither the architectures of the *cis*-regulatory regions, nor the functional form of the gene regulatory functions, have to be specified *a priori*: in the course of our simulations, TF binding sites emerge naturally as sites with a particularly high affinity for a certain TF.

We have used our approach to design all possible transcriptional logic gates with two input signals and one output signal (see Table 2.1). These gates have been studied by Buchler *et al.* (2003) using a rational design approach. Our simulations, however, unravels new design principles. In spite of the simplicity

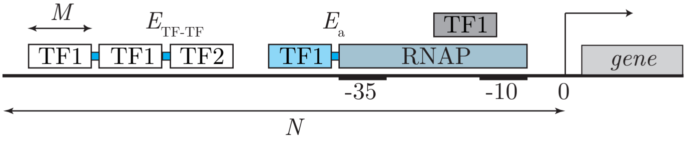


Figure 2.3: Illustration of the model. The *cis*-regulatory region consists of $N = 100$ base pairs directly upstream of the transcription start site. In *E. coli*, most TFs bind to this region, although binding sites are also found downstream of the transcription start site; mechanisms requiring such downstream sites are excluded by our model. A TF binding domain counts M amino acids, which can bind $M = 10$ base pairs (Madan Babu and Teichmann, 2003; Pérez-Rueda and Collado-Vides, 2000). When two TFs bind within a distance less than $k = 3$ base pairs, they interact with energy $E_{\mathrm{TF-TF}}$; this is indicated by a blue connection between the TFs. When a TF binds close to the RNAp, we assume an interaction energy $E_{\mathrm{TF-P}}$. The core promoter, consisting of the -10 and -35 hexamers, is indicated; when the RNAp binds to it, it blocks both hexamers and the spacer between them. The TF that binds overlapping with the RNAp is gray, to indicate that it represses transcription by steric hindrance; the blue TF is an activator, since it recruits RNAp. The white TFs bind too far upstream from the core promoter to influence the transcription rate.

of the model, quite complex functionality can emerge. In particular, we find that promoter architectures are often constructed from modules that consist of tandem arrays of binding sites to which TFs can bind cooperatively (see Fig. 2.4). Furthermore, these modules often overlap, leading to competition between them. We show that the intricate interplay between intra-modular cooperativity and inter-modular competition allows for a wide range of regulatory functions.

2.2 Model of transcriptional regulation

We assume that the transcription rate of an operon is proportional to the fraction of time RNAp is bound to the promoter (Shea and Ackers, 1985; Buchler *et al.*, 2003; Bintu *et al.*, 2005a,b). The model we use to compute this quantity is illustrated in Fig. 2.3. The RNAp–σ complex binds only to the -10 and -35 hexamers, called the core promoter, and we determine its binding energy by comparing the core promoter to a large set of real *E. coli* promoters (Lisser and Margalit, 1993; Berg and von Hippel, 1987, 1988; Berg, 1988)(see Appendix 2.A for details). We ignore the fact that, in some promoters, the affinity of the RNAp for the promoter is enhanced by interactions of its α C-terminal domain with DNA upstream of the -35 hexamers. TFs can bind to any site in the *cis*-regulatory region. Whenever a TF binds to the DNA, each amino acid interacts with exactly one base pair, and the total binding free energy is the sum of the contributions of each amino-acid–base-pair contact. This is known to be a reasonable approximation for many TFs (Fields *et al.*, 1997; Stormo and Fields, 1998; Benos *et al.*, 2002; Berg and von Hippel, 1987, 1988), even though exceptions have also been documented (Roulet *et al.*, 2002). The binding energies associated with each amino-acid–base-pair contact are extracted from a matrix

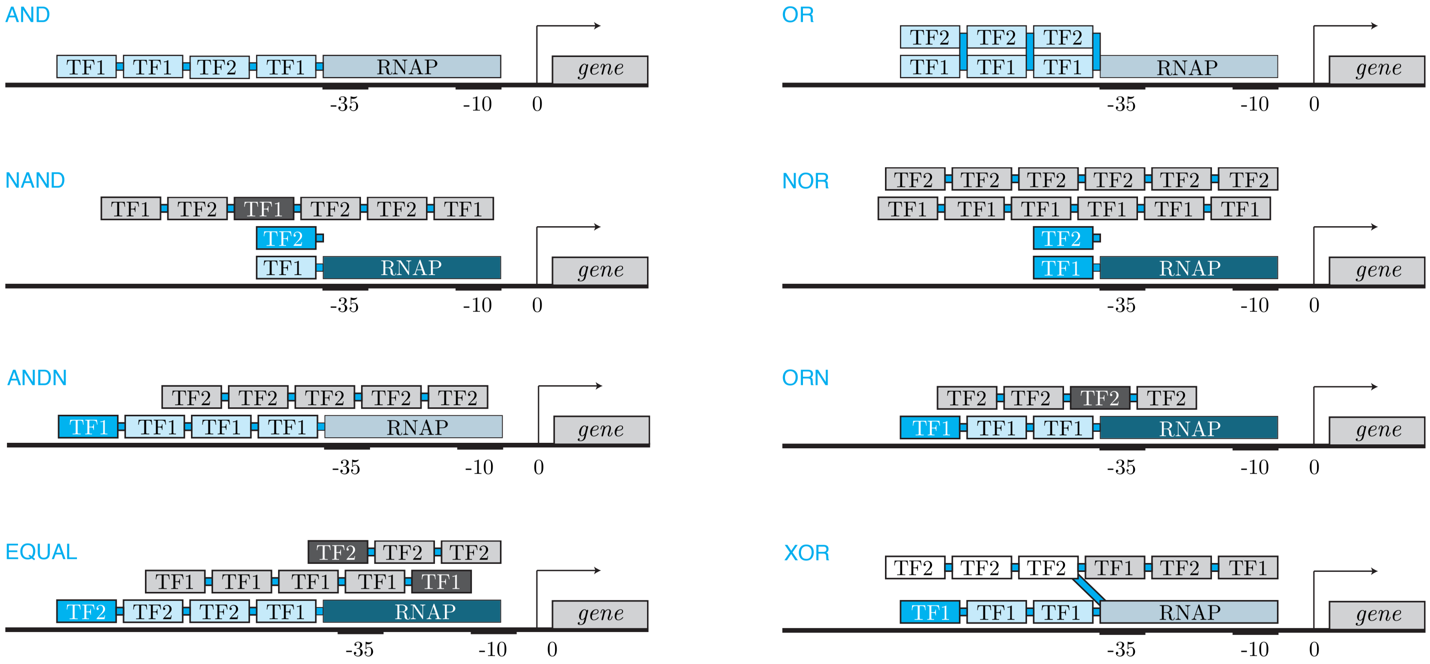


Figure 2.4: Cartoons of *cis*-regulatory constructs emerging from our *in silico* design of transcriptional logic gates. The boxes indicate the TF binding sites; blue indicates that a TF acts as an activator, gray that it acts as a repressor, and white that the action of the TF depends upon the concentrations of the two TFs. Weak binding sites ($K_\mathrm{D} > 2 \times 10^3$ nM) have a light color, strong ones are dark. Blue connections between TFs signify cooperative interactions. The designs show that the logic gates are constructed as overlapping arrays of cooperative binding sites. Each layer acts as a module, either activating or repressing transcription. Signals are integrated via the interplay between intra-modular cooperativity and inter-modular competition.

TFs		operon activity							
c_1	c_2	AND	ANDN	XOR	OR	NOR	EQU	ORN	NAND
low	low	off	off	off	off	on	on	on	on
low	high	off	off	on	on	off	off	off	on
high	low	off	on	on	on	off	off	on	on
high	high	on	off	off	on	off	on	on	off

Table 2.1: Truth tables of transcriptional logic gates. Logic gates are devices that perform elementary binary computations, mapping multiple input signals to one output signal. Here we consider transcriptional logic gates with two inputs and one output. The table specifies, for each gate, the status of the operon (“on” or “off”) for all TF concentrations c_1 and c_2 (“low” or “high”). Gates that are identical to one of the shown gates up to TF labels, and those that depend on only one TF, are disregarded. In our simulations, concentrations above (below) 500nM are considered high (low). The acronyms of the gates summarize their function: the operon of an AND gate should only be transcribed when both c_1 *and* c_2 are high. The acronym ORN stands for “or not”: the gate is “on” if c_1 is high *or* c_2 is *not* high. The EQU gate is “on” if the input concentrations are *equal* (either both low or both high). The activity of the “NOR” (Not OR) gate is, in all conditions, opposite to the activity of the “OR" gate.

based on crystallographically solved protein–DNA complexes (Mandel-Gutfreund and Margalit, 1998). The results, however, do not depend critically upon the precise values of the matrix elements; random matrices with the same mean and standard deviation give similar results. Note that some real TFs can bind ligands or can become phosphorylated; in that case the TF concentration in our model corresponds to the concentration of the DNA-binding form of the TF.

The model allows for two types of TF–TF and TF–RNAp interactions (see Fig. 2.3) (Ptashne and Gann, 2002). Firstly, we include steric hindrance: molecules cannot overlap in space. Secondly, we include a cooperative interaction of energy $E_{\mathrm{TF-TF}}$ between any pair of TFs when they bind within a distance of k base pairs. Likewise, if a TF and RNAp bind close together, we assume a synergetic energy $E_{\mathrm{TF-P}}$ (Busby and Ebright, 1994). We thus assume that TFs can bind cooperatively with themselves, with RNAp and with other TFs. We note that, although some TFs are known to have all these properties (for instance MalT and MelR), it is unlikely that this is the case for all TFs. Our results will show, however, that combinations of some of these properties allow for a myriad of promoter functions. In our simulations we used $k = 3$ and $E_{\mathrm{TF-TF}} = E_{\mathrm{TF-P}} = 3.40\,\mathrm{k_B T}$ or $2.0\,\mathrm{kCal/mol}$ (so that $\exp(\beta E_{\mathrm{TF-TF}}) = 30.0$) (Buchler *et al.*, 2003).

Such protein–protein interactions could have various origins. Many TFs interact by direct contact between patches on their surface. We note that these interactions are very weak and are therefore not likely to be detected in large-scale experiments such as those of Butland *et al.* (2005). However, cooperativity can also result from bending, stretching or super-coiling of the DNA by one of the molecules, affecting the binding affinity of the other (Berg *et al.*, 2004; Ptashne and Gann, 2002). At the level of description of our model, such mechanisms can be described in the same way as cooperativity by direct contact. This means that most effects of *local* chromosome structure are implicitly included in the model. However, the model does not allow action at a distance. Therefore, mechanisms involving *global* chromosome structure, such as DNA looping, are not included. Also, mechanisms that rely on direct interactions between the RNAp and TFs bound further upstream, for instance through contact with the flexible RNAp α C terminal domain, are not possible in our model, even though it could be extended to incorporate such effects (Buchler *et al.*, 2003).

We use the statistical mechanical approach developed by Shea and Ackers (1985) and Buchler *et al.* (2003) to describe the input–output relationship for an operon in a quantitative way (see Appendix 2.A and Section 1.1.2 of the Introduction chapter). In order to compute the influence of each TF on the transcription rate in a tractable way, we have developed a fast algorithm that efficiently takes into account all TF–DNA, TF–TF and TF–RNAp interactions (see Appendix 2.A).

2.3 Evolutionary design of logic gates

We have used the model and an evolutionary algorithm to design transcriptional logic gates consisting of one operon, regulated by two TFs. Typically, 250 gates, with initially random DNA and amino-acid sequences, were subjected to cycles of mutation and selection. In each cycle, point mutations were introduced; the probability of a mutation occurring within a given *cis*-regulatory region or TF was 0.85 and 0.3 respectively — but the results do not depend strongly on these values. Next, the top 20% of the gates were selected and the others were removed. To complete the cycle, we finally refilled the empty slots by copying randomly chosen genotypes from the selected gates.

In order to select the top 20% of the gates, we define a fitness function that quantifies the quality of the gate. The transcription rate A of a gate depends on the concentrations c_1 and c_2 of the two TFs: $A = A(c_1, c_2)$. First, we compute the transcription rate for 16 values of (c_1, c_2) in the range 0–1000 nM; for the AND gate in Fig. 2.5, these 4×4 values are depicted as dark dots. For each of these points, we determine how far A deviates from a goal function $G(c_1, c_2)$, which is defined by the logic gate we are trying to obtain. Next, we compute the sum of the squares of these deviations. If this quantity is small, then the fitness is considered high. (Refer to Appendix 2.A for more details.) Our fitness function selects for rather steeply-switching gates, since the switching is required to take place between $c_i = 333\,\mathrm{nM}$ (considered low) and $c_i = 667\,\mathrm{nM}$ (considered high). We also implicitly assume that all conditions are equally important; each of the 16 points has an equal weight in the fitness function. In reality, this is not necessarily the case: the fitness cost of a gene being “on” at a wrong time, need not match the cost of one that is “off” when it should not be (see also Berg *et al.* (2004)). In order to elucidate general design principles, we select for idealized promoter functions, although, clearly, in nature the input–output relations can be more intricate; an example is the *lac* promoter, which is not a perfect ANDN gate (Setty *et al.*, 2003).

2.4 Results: *cis*-regulatory constructs

Figs. 2.4 and 2.5 show typical simulation results for the gates in Table 2.1. Clearly, the architectures can be quite complex. Interestingly, the final constructs do not depend much on the initial conditions; this can be regarded as a simple example of convergent evolution. Moreover, they are remarkably similar to the structures found in *E. coli*, as we now describe.

2.4.1 Homo-cooperative auxiliary sites provide steep responses

We can distinguish two kinds of binding sites. Binding sites from where the TFs directly interact with the RNAp are called *primary* sites. Primary activator sites are located right next to the -35 hexamer of the core promoter, whereas primary repressor sites directly overlap with the core promoter. The remaining binding sites are called *auxiliary* or *secondary* sites (Müller-Hill, 1998). These sites provide cooperativity. The main function of cooperativity between identical TFs, called *homo*-cooperativity, is to create steep responses (Buchler *et al.*, 2003; Alberts *et al.*, 1994). We find that both activating and repressing binding sites are regularly supported by (tandem arrays of) auxiliary sites.

Activation

In cooperative arrays of activation sites, the auxiliary site furthest removed from the core promoter usually has the strongest affinity. This can be seen in the *cis*-regulatory regions of EQU, ORN, XOR and ANDN. Further analysis shows that this pattern enhances the steepness of response (see Appendix 2.B). The steepness is optimal if the binding affinities of the furthest site and those of the other sites differ by a factor 2 to 14, depending on the strength of the promoter, the value of the interaction energies ($E_{\mathrm{TF-P}}$ and $E_{\mathrm{TF-TF}}$) and the number of tandem repeats: this way, the steepness can be enhanced up to 27%. A similar result was presented in (Bintu *et al.*, 2005b) for systems with one auxiliary site, in the context of the regulation of the phage λ promoter P_{RM}. We therefore predict that activating auxiliary sites in real promoters regularly have a higher affinity than their primary sites.

It may be useful to repeat that we define auxiliary sites as sites that do not interact directly with the RNAp. If, in real *E. coli* promoters, any of the upstream sites does interact with RNAp, for instance via direct contact with the α C-terminal subdomain of the RNAp, then such a site is, by definition, a primary site. If such a distant primary site is accompanied by an auxiliary site,

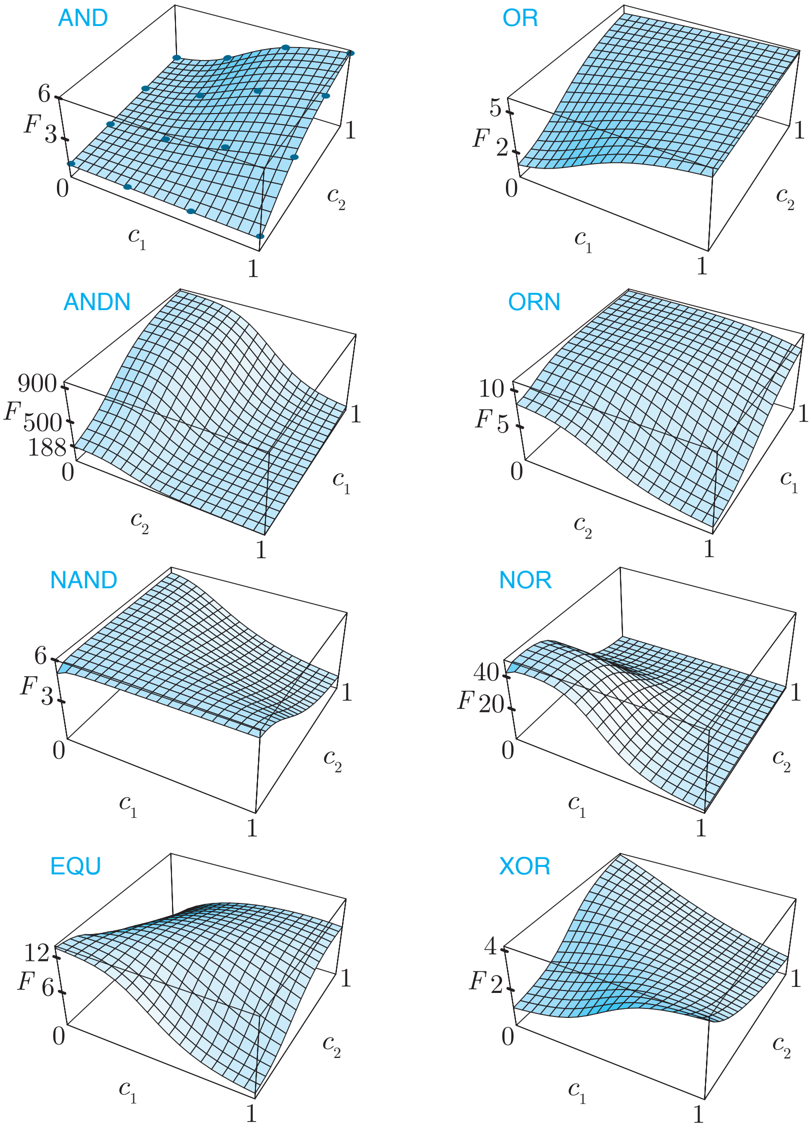


Figure 2.5: Response plots of logic gates emerging from the simulations. The quantity F on the vertical axis is the fold change of the transcription rate, defined here as $F = A(c_1, c_2)/A_{\text{min}}$, where A_{min} is the minimal transcription rate on this TF concentration domain. The concentrations c_1 and c_2 are in μM and plotted on a linear rather than logarithmic scale. The 16 dark dots in the AND gate indicate the measurement points (c_1, c_2) that are used in the fitness function.

then this auxiliary site still needs to have a higher affinity than its primary site in order to maximize the steepness of response.

In *E. coli*, homo-cooperative activation occurs regularly. For example, the TFs of the LysR family often bind to two sites, one at −65 and the other close to the −35 hexamer of the core promoter (Wagner, 2000; Schell, 1993). In some cases, the TFs bind cooperatively to these sites; in these cases the site at −65 has a stronger affinity than that near −35 (Wilson *et al.*, 1995; Lamblin and Fuchs, 1994), as one would expect from our results. Another example is the activation of the P_{RM} promoter in bacteriophage λ by CI, which binds more strongly to the auxiliary site ($\mathrm{O_{R1}}$) than to the primary activation site ($\mathrm{O_{R2}}$) (Shea and Ackers, 1985; Bintu *et al.*, 2005b). We note however, that this example is complicated by the fact that $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$ are also involved in repressing the P_{R} promoter. We will get back to this in the next subsection.

Repression

In contrast to the activation modules, the auxiliary sites in repressor complexes are usually much *weaker* than the primary ones (*e.g.* ORN and EQU). Further analysis (see Appendix 2.B) shows that the steepness of repression is optimal if the primary site has a 5 to 50 times higher affinity than the auxiliary sites (depending on the promoter strength, the values of the interaction energies and the number of tandem repeats). This pattern can increase the maximal steepness of the response by about 70%, as compared to the case where all sites have an equal affinity. As weak auxiliary sites of repressor systems are not only sufficient, but even optimal, it seems highly unlikely that evolution would maintain strong auxiliary repression sites if steep responses are beneficial (see Appendix 2.B). We therefore predict that auxiliary sites in real repressor systems should often be weak.

Indeed, most well-characterized repressor systems in *E. coli* have auxiliary operators (Müller-Hill, 1998; Rojo, 2001), many of which are weak. For example, the two cooperative Fur-binding sites that overlap the core promoter on the pColV-K30 plasmid are supported by an array of low affinity auxiliary sites (Escolar *et al.*, 2000). A second example is the duo of *dnaA* promoters, 1P and 2P (Lee and Hwang, 1997). At low concentrations DnaA represses only 1P, but at high concentrations it blocks both promoters, as a result of the cooperative binding of up to four DnaA monomers to weak binding sites overlapping the 2P region (Lee and Hwang, 1997). Other examples are the TrpR repressor on the *trp* promoter (Jeeves *et al.*, 1999) and the Fis repressor on the *aldB* promoter (Xu and Johnson, 1995). Finally, the *gltA–sdhC* intergenic region contains at least two high-affinity ArcA–P repressor sites, one overlapping the *gltA* promoter and one blocking the *sdhC* promoter; at higher ArcA–P concentrations both binding regions broaden until ArcA–P covers a region of about 230 bp, suggesting ArcA–P oligomerization on the DNA (Lynch and Lin, 1996; Shen and Gunsalus, 1997).

In the previous subsection, we mentioned the activation of P_{RM} by CI in the bacteriophage λ as an example of cooperative activation, and argued that steep activation requires that the auxiliary site $\mathrm{O_{R1}}$ should be considerably stronger than the primary site $\mathrm{O_{R2}}$. Interestingly, the same CI binding sites $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$ are also involved in repressing the P_{R} promoter. But the binding sites now have reversed roles: from the point of view of promoter P_{R}, $\mathrm{O_{R1}}$ is the primary repressor site, and $\mathrm{O_{R2}}$ is auxiliary. However, since we just argued that, in repressor systems, primary sites should be stronger than auxiliary sites, we conclude that both for steep activation of P_{RM} and for steep repression of P_{R}, site $\mathrm{O_{R1}}$ needs to be stronger than site $\mathrm{O_{R2}}$, as is indeed the case.

As a final remark on homo-cooperativity, we point out that, while cooperativity is used widely, as Fig. 2.1 shows, many of the better characterized promoters, such as the *lac* promoter, have a simpler architecture. It should be realized that the number of binding sites not only depends upon the complexity of the desired input–output relation, but also upon the required steepness of the response. If, for instance, we select for simpler gates with a weaker response function, we do obtain simpler promoter architectures (data not shown).

2.4.2 Hetero-cooperativity provides conditional responses

While the benefit of homo-cooperativity is to create steep responses, the function of cooperativity between *different* molecular species, *hetero*-cooperativity, is rather to integrate signals. It can be used whenever a response should be conditional on the presence of more than one TF. A good example is the AND gate. As with the OR gate, this gate requires a weak promoter — this ensures that the operon is not transcribed when both TFs are absent. In contrast to the OR gate, however, the AND gate should be on only when both TF1 and TF2 are present. The activation is therefore mediated by a TF1 binding site that is too weak to be functional by itself. Next to this site a stronger TF2 binding site is present. Only when TF1 and TF2 are both present, they bind cooperatively and induce activation (Buchler *et al.*, 2003). The remaining sites can bind either TF1 or TF2 and are responsible for the steepness of the response.

Activation

Hetero-cooperative activation is found regularly in naturally occurring promoters. A good example is the activation of the *melAB* operon by MelR, which binds to four sites (Wade *et al.*, 2001). A CRP binding site is present between MelR sites 2 and 3. Here, CRP binds cooperatively with the downstream MelR sites. This increases their fractional occupancy, resulting in transcription activation. Another excellent example is the *malKp* promoter (see Fig. 2.1(d)), which is discussed below (Richet and Sogaard-Andersen, 1994; Richet, 2000).

Repression

The CytR regulon provides an example of hetero-cooperative repression. CytR often binds cooperatively with cAMP–CRP to form a repression complex. Good examples are *udp* (Brikun *et al.*, 1996), *nupG* (Pedersen *et al.*, 1995), *tsx-p2*

(Gerlach *et al.*, 1991) and *deoP2* (Shin *et al.*, 2001); see also (Tretyachenko-Ladokhina *et al.*, 2001; Meibom *et al.*, 1999). Recently, it has also been shown that Lrp and H-NS act cooperatively at the *rrnB* promoter (Pul *et al.*, 2005).

2.4.3 Competition between modules

Whenever binding sites overlap, competition between TF complexes occurs. It is well known that the core promoter often overlaps with an operator; this is a standard repression mechanism (Müller-Hill, 1998). The role of overlapping TF binding sites in signal integration has been less commented on. Clearly, a repressor which binds to an operator overlapping with an activator site can be used to create anti-activation. Likewise, anti-repression occurs when a binding site overlaps with a repressor site, but not with the core promoter. But the full potential of this type of competition becomes clear only when it is combined with cooperativity. Our NOR, NAND, EQU, and XOR gates serve as instructive examples.

Sharpening repression by competitive activation

The NOR gate (see Figs. 2.4, 2.5 and Table 2.1) combines competition and homo-cooperativity. This gate contains both activator and repressor sites for each TF. The single activator sites are strong compared to the repressor sites; as a result, activation dominates at low TF concentrations. However, as the TF concentrations increase, the affinity of the repressor module grows more rapidly; this is the result of the homo-cooperativity between the repressor sites. Consequently, at high TF concentrations repression dominates. The function of the activating sites is thus to counteract repression at low concentrations, thereby increasing the switching steepness. As it turns out, whenever we select for steep repression, we also get activation. The general message is that using competing modules containing different numbers of homo-cooperative binding sites, a TF can effectively be both an activator and a repressor, depending on its concentration.

The NAND gate looks rather similar to the NOR gate, but uses hetero- instead of homo-cooperativity. Repression dominates only if both TF1 and TF2 are present in sufficient concentrations. This shows that by combining competition and hetero-cooperativity, a TF can either be an activator or a repressor, *conditionally* on the concentration of another TF.

Intra-modular cooperativity and inter-modular competition

In the EQU gate all mechanisms act in concert. In an EQU gate the operon must be on when the concentrations of both TFs are low; this requires a strong promoter. If either TF1 or TF2 is present, the operon must be off; this requires homo-cooperative repression modules, which block the binding of RNAp when either TF1 or TF2 is present. However, if both TF1 and TF2 are present in similar concentrations, the operon must be on; this requires a hetero-cooperative activation module that counteracts the effect of the homo-cooperative repression modules.

In the XOR gate, the same mechanisms act, but in an opposite manner: if both TFs are absent, the operon should be off; this requires a weak promoter. If one of the two TFs is present, the operon should be on; this demands homo-cooperative activation modules, which recruit the RNAp when only one of the two TFs is present. If both TFs are present, however, the operon should be off; this requires a hetero-cooperative repression module that neutralizes the actions of the homo-cooperative modules when both TFs are present.

In both gates, the homo-cooperative and hetero-cooperative modules have to compete with one another. This is achieved via the binding of the TFs to overlapping binding sites. Which module wins the competition depends upon the TF concentrations, the number of TFs in the modules and upon the quantitative details of the protein–protein and protein–DNA interactions. In Appendix 2.C we discuss minimal models of both gates quantitatively.

Similar mechanisms are known to occur in *E. coli*. The *malKp* promoter (see Fig. 2.1(d)) provides a good example, although its full input–output relation is more complex than those of the logic gates studied here. In the presence of CRP, MalT binds to three tandem sites to form the activation complex (Richet and Sogaard-Andersen, 1994; Richet, 2000). In the absence of CRP, however, MalT binds with relatively high affinity to an alternative triplet of repressor sites that overlaps the activation complex, thereby repressing *malK*. As in the EQU gate presented here, the activation complex has to compete with the repression complex; the CRP concentration determines whether MalT acts as a repressor or as an activator (Richet and Sogaard-Andersen, 1994; Richet, 2000).

2.5 Discussion and outlook

We have developed a model of transcriptional regulation and applied it to the evolutionary design of transcriptional logic gates in prokaryotes. Our approach has revealed new design principles, which would have been difficult to predict using a rational design approach. In particular, our analysis stresses the importance of the interplay of the following mechanisms:

1. homo-cooperative interactions between TFs within modules;
2. hetero-cooperative interactions between TFs within modules;
3. competition between TF modules. Using these mechanisms only, a wide range of input–output relations can be produced, including the full repertoire of *cis*-regulatory logic gates with two input signals and one output signal.

The resulting constructs make extensive use of cooperative tandem binding sites. Homo-cooperativity is often used as a means of achieving high Hill coefficients. In such tandem arrays of binding sites, weak sites can be important. In repressive arrays, auxiliary sites are usually weak, while in activating arrays the auxiliary sites tend to have the highest affinity. Hetero-cooperativity allows for regulation conditional on the presence of more than one TF species. Hetero-cooperativity within modules thus plays a central role in integrating different signals; in the gates studied here, a hetero-cooperative module only becomes active if both TFs are present. Nevertheless, we wish to emphasize that while many promoters in nature exhibit long arrays of binding sites (see Figs. 2.1 and 2.2), it is not likely that all TFs of *E. coli* have the capacity to bind cooperatively into such long arrays. It should be realized, however, that for most gates, modules consisting of two or three cooperative sites are sufficient, although additional sites add to the quality of the gates and to the sharpness of the responses.

The capacity to integrate signals is dramatically enhanced by the competition between different modules, as summarized in Table 2.2. Competing modules allow the integration of signals, because (a) both homo- and hetero-cooperative modules can act as activator modules or as repressor modules, and (b) when the concentrations of the TFs vary, also the relative activities of the activating and repressing modules change. How their activities change with the TF concentrations, depends upon the strength of the TF–DNA, TF–TF, and TF–RNAp interactions. It also depends upon the degree of cooperativity: the number of binding sites in a module not only determines the steepness of the response, but also affects the concentration range in which the module is active. For instance, a large module will dominate an overlapping, but smaller one at sufficiently high TF concentrations, even when the individual TFs in the larger module have a weaker affinity for the DNA. Indeed, not only hetero-cooperativity, but also homo-cooperativity can play an essential role in signal integration (see also Fig. 2.8).

Our results provide a possible explanation for the complexity of *cis*-regulatory regions found in *E. coli*, which, indeed, often contain tandem TF binding sites and overlapping sites. Our analysis suggests that these complex architectures are a natural consequence of the basic mechanisms of transcriptional regulation and, on the other hand, the function of *cis*-regulatory domains to integrate signals. While we focus here on prokaryotes, it should be clear that similar integration mechanisms might also operate in the *cis*-regulatory domains of transcription units in eukaryotes; ample anecdotal evidence exists, *e.g.*, for the role of adjacent and overlapping TF binding sites in signal integration during embryonic development of the sea urchin (Yuh *et al.*, 1998) and *Drosophila* (Gilbert, 2003). Our results also emphasize that understanding the complex promoters observed both in our simulations and in nature, requires quantitative knowledge of binding affinities and interactions: from the binding site locations only, it is often not possible to distinguish an AND gate from an OR, nor a NAND from a NOR.

In this chapter, we have used our evolutionary design method to design cis-regulatory domains of single operons. This method, however, could also be applied to design larger networks, such as multi-input modules (Shen-Orr *et al.*, 2002). As the network size increases and regulons become larger, we expect that it will become increasingly more difficult to fulfill all constraints imposed on the

		c_1	
		low	high
c_2	high	On: homo-cooperative activation by TF2 if promoter is weak	On: hetero-cooperative activation if not activated homo-cooperatively
		Off: homo-cooperative repression by TF2 if promoter is strong	Off: hetero-cooperative repression if not repressed homo-cooperatively
	low	On: strong promoter	On: homo-cooperative activation by TF1 if promoter is weak
		Off: weak promoter	Off: homo-cooperative repression by TF1 if promoter is strong

Table 2.2: Table summarizing which homo- or hetero-cooperative activation or repression modules are needed to obtain a particular transcriptional logic gate. The table consists of four quadrants, corresponding to different TF concentrations c_1 and c_2 (each being low or high). Each quadrant is divided into two parts (white and gray), corresponding to the alternative promoter states (on or off). As an example, the AND gate is on only if both TF1 and TF2 are present; this requires a hetero-cooperative activation module. In contrast, an OR gate should be on if either TF1 or TF2 is present. This requires homo-cooperative activation modules for each of the species, because the promoter is weak (the gate must be off when both species are absent); however, since the activation modules do not compete with one another, a hetero-activation module is not required: the homo-cooperative activation modules also turn the gate on when both TFs are present. In general, the design can be most easily understood by first considering the design constraints when both TFs are absent, then the requirements when one of the two are present, and lastly the design constraints when both TFs are present. The EQU and XOR gates discussed in this chapter illustrate this perhaps most clearly. Note that the EQU gate is an example of a gate in which a hetero-activation module is required, despite the fact that the promoter is strong; the hetero-activation module is needed to counteract the two homo-cooperative repression modules when both TFs are present.

promoter and TF sequences. For these larger networks, not only *positive* design — selecting for desired TF–DNA interactions — but also *negative* design — selecting against unwanted TF–DNA interactions — may be an important design criterion. Our approach can also be extended to design feedback networks. By selecting transcription networks containing multiple genes based on their *dynamics*, we can design feedback systems like transcriptional oscillators (Francois and Hakim, 2004). In the next chapter, we use our method to design a bi-stable switch.

Here, we used our method to design transcriptional logic gates. For this reason, our evolutionary algorithm was not designed to mimic natural or directed evolution. However, with suitable modifications and extentions, our approach could also be used to study questions that are pertinent to the evolution of functional promoter regions, such as what are the pathways of evolution, and how does the evolution of logic gates depend upon factors like population size, neutral drift, and mutation rates.

Finally, the proposed signal integration mechanism of intra-modular cooperativity versus inter-modular competition could be tested experimentally by rationally designing *cis*-regulatory constructs. But perhaps more interesting would be to see whether an evolutionary design method can be used. Recently, Yokobayashi *et al.* demonstrated experimentally that directed evolution can be used to change protein–DNA and protein–protein interactions in a rationally designed, but non-functional gene circuit to obtain a functional network (Yokobayashi *et al.*, 2002). Perhaps a similar method can be used to design, by experiment, transcriptional logic gates with desired input–output relations. Since no specific promoter designs have to be imposed, it would be interesting to see whether the resulting architectures exploit the signal integration mechanism of competing binding site modules.

2.A Detailed description of the model of transcription regulation

The model needs to address four quantities: (i) the binding affinities of each transcription factor (TF) for every possible site on the *cis*-regulatory region, (ii) the affinity of RNAp–σ for the core promoter, (iii) the interactions between the molecules, and finally (iv) the transcription rates based on these affinities and interactions. We discuss each of these issues below.

Binding of TFs to DNA

TFs can bind anywhere on the *cis*-regulatory region. The affinity of a TF for a given site is determined by the DNA sequence at the site and the amino-acid sequence in the DNA binding pocket of the TF. We assume that, whenever a TF a binds to a binding site O, each amino acid interacts with exactly one base pair, and that the total binding free energy $E_{a,\mathrm{O}}$ is the sum of the contributions of each amino-acid−base-pair contact. This means that the binding free energy of a TF a with amino acids a_i to a binding site O with base pairs O_i is given by

$$E_{a,\mathrm{O}} = \sum_{i=1}^{M} U_{a_i,\mathrm{O}_i}.$$

Here $U_{\lambda\mu}$ is a 20 × 4 matrix containing the binding free energies associated with each amino-acid−base-pair contact. We used a matrix given in Mandel-Gutfreund and Margalit (1998), based on christallographically solved protein–DNA complexes (see Section 1.1.2 of the Introduction chapter).

Finally, the binding affinity $q_{a,\mathrm{O}}$ of TF a for site O follows from

$$K_{a,\mathrm{O}} = \alpha\, \mathrm{e}^{-\beta E_{a,\mathrm{O}}}, \qquad q_{a,\mathrm{O}} = \frac{c_a}{K_{a,\mathrm{O}}}. \tag{A2.1}$$

Here, $K_{a,\mathrm{O}}$ denotes the dissociation constant and c_a denotes the concentration of TF a. The proportionality factor α in Equation A2.1 is determined by the free energy of all other sites that compete with O for binding of the TF. Initially, we used $\alpha = 10^7$ nM, but also found that the results do not depend critically on this value. Later, we used Equation 1.13, an estimate for the total contribution of the background states, to compute the value of α separately for each TF. The same designs were found.

Binding of RNAp

In our model, the RNAp–σ complex binds only to the core promoter. We determine the binding free energy of RNAp–σ for a core promoter p by comparing the -10 and -35 hexamers to a large set of real *E. coli* promoters, taken from reference (Lisser and Margalit, 1993). To every base pair p_i at position i within the -10 and -35 hexamers, we assign a score s_i; it equals the fraction of real *E. coli* promoters that have p_i at that particular position, normalized by the random fraction 1/4. Next, the binding energy E_p of the RNAp to that particular core promoter can be estimated by (Berg and von Hippel, 1987, 1988; Berg, 1988):

$$E_\mathrm{p} = k_\mathrm{B}T \sum_{i \in p} \log(s_i).$$

The dissociation constant of the binding reaction, K_p, and the binding affinity of the RNAp for the promoter, q_p, now follow from the equations

$$K_\mathrm{p} = \alpha' \mathrm{e}^{-\beta E_\mathrm{p}}, \qquad q_\mathrm{p} = \frac{c_\mathrm{p}}{K_\mathrm{p}} \propto c_\mathrm{p} \prod_{i \in p} s_i. \tag{A2.2}$$

The proportionality factor α' in Equation A2.2 again includes the competition between site p and all other places the RNAp could possibly be; it should be chosen such that p_on is close to unity in case of a small number of mismatches, but decreases rapidly as mismatches accumulate. For the results shown, we used $\alpha' = 10^7\,\mathrm{nM}$.

TF–TF and TF–RNAp interactions

The interaction between the molecules consists of two parts. In the first place, we include steric hindrance: TFs and RNAp cannot overlap in space. When bound to the DNA, TFs occupy M base pairs and mutually exclude each other and RNAp. Bound RNAp is assumed to block the consensus hexamers and the spacer in-between. In the second place, we include an unspecific attractive interaction between TFs whenever they bind close to each other — that is, within a distance of $k = 3$ base pairs. To this interaction we associate an energy $E_\mathrm{TF-TF}$ of 2–4 $k_\mathrm{B}T$, such that $\omega \equiv \exp(\beta E_\mathrm{TF-TF}) \approx 30$ (Buchler *et al.*, 2003). Likewise, if a TF and RNAp bind close together, we assume a similar interaction free energy $E_\mathrm{TF-P}$; Again, $\omega' \equiv \exp(\beta E_\mathrm{TF-P}) \approx 30$.

Transcription rates

We assume, following Shea and Ackers (1985) and Buchler *et al.* (2003), that the transcription rate A of an operon is proportional to the fraction of time p_on an RNAp is bound to the core promoter. This assumption is reasonable provided the kinetics of the binding and unbinding of RNAp are sufficiently fast in comparison to the transition rate from the closed to the open complex (see Section 1.1.2). In that case the binding reaction is near equilibrium and the fractional occupancy is given by

$$A \propto p_\mathrm{on} = \frac{Z_\mathrm{on}}{Z}. \tag{A2.3}$$

Here Z_on is the partition sum of all states in which an RNAp molecule is bound, and Z is the total partition sum. This approach is used widely (Shea and Ackers, 1985; Buchler *et al.*, 2003; Bintu *et al.*, 2005a,b; Graham and Duke, 2005). We note however, that this model does not apply to all cases: for instance, some

TFs function by regulating the rate of the transcriptional steps after the initial binding of RNAp to the core promoter, and in some cases a tight binding of RNAp to the core promoter might negatively influence the transition rate to the so-called open complex.

Computing the partition sums

In the previous subsections, we explained how to compute the TF binding affinities for each possible position on the *cis*-regulatory region, the affinity of RNAp for the core promoter, and all interaction energies, given the sequences and concentrations of the molecules. This allows us in principle to compute the Boltzmann factor $W(s)$ of every state s and the hence the partition sum Z of the system. But since we assumed that TFs can bind anywhere on the DNA, the total number of states or configurations can easily become huge. In fact, a minimal network consisting of only one operon and two TFs with $N = 80$ and $M = 10$, counts more than three million distinct configurations. We developed a scheme that nevertheless allows us to compute the partition sum for a given promoter in an efficient manner.

We use the following conventions (see Fig. 2.3). We refer to the stretch of DNA ranging from base pair $i - M + 1$ to base pair i as site i. We denote the binding affinity of TF a for site i, as defined in Equation A2.1, by $q_{a,i}$. Next we define

$$Q_i \equiv \sum_a q_{a,i}.$$

Finally we consider a series Z_i of *partial* partition sums ($-N \leq i \leq 0$), defined as the partition sum of all possible states in which sites with a number bigger than i are not occupied and no RNAp is bound.

Let s be the state where TFs $a_1 \dots a_m$ are bound to sites $x_1 \dots x_m$ respectively. Then in Buchler *et al.* (2003) it is explained that the Boltzmann factor $W(s)$ of s equals

$$W(s) = \left(\prod_{u \neq v} \omega_{u,v} \right) \left(\prod_{u=1}^{m} q_{a_u,x_u} \right),$$

where

$$\omega_{u,v} = \begin{cases} \omega & \text{if site } u \text{ and } v \text{ are 0 to } k\,\text{bp apart,} \\ 0 & \text{if site } u \text{ and } v \text{ overlap,} \\ 1 & \text{else.} \end{cases}$$

This implies that for the series Z_i, the following recurrence relation holds:

$$\begin{aligned} Z_i &= Q_i Z_{i-M-k} + Q_i \omega \left(Z_{i-M} - Z_{i-M-k} \right) + Z_{i-1} \\ &= Q_i \left((1-\omega) Z_{i-M-k} + \omega Z_{i-M} \right) + Z_{i-1}, \end{aligned} \tag{A2.4}$$

with starting conditions

$$Z_i = \begin{cases} 0 & \text{for } i < -N, \\ 1 & \text{for } -N < i < -N + M. \end{cases}$$

We can express Z_{off}, Z_{on} and p_{on} in terms of the Z_i as

$$Z_{\text{off}} = Z_0, \tag{A2.5}$$

$$Z_{\text{on}} = q_{\text{p}} \left(\omega' Z_x + (1 - \omega') Z_{x-k}\right), \tag{A2.6}$$

$$p_{\text{on}} = \frac{1}{1 + Z_{\text{off}}/Z_{\text{on}}}. \tag{A2.7}$$

Here x is the base pair just next to the core promoter ($x = -37$). The conclusion is that, in order to compute p_{on}, one only needs to compute the Q_i, apply Equation A2.4 N times, and finally fill in expressions A2.5, A2.6 and A2.7. Note that the time required to compute p_{on} using this algorithm scales linearly with N, M, and the number of TFs. This shows that the scheme is fast and can therefore be applied to networks consisting of many genes and TFs.

Fitness function

In order to select the gates, we need a fitness function that quantifies their quality. We now describe the fitness function we used. The transcription rate A of a gate depends on the concentrations c_1 and c_2 of the two TFs: $A = A(c_1, c_2)$. We use concentrations in the rage 0 to 10^3 nM; concentrations below (above) $c_{\text{mid}} \equiv 500$ nM are considered low (high). Each truth table t then defines a goal function $G_t(c_1, c_2)$; the perfect analog AND gate, for instance, has the following response:

$$\begin{aligned} A(c_1, c_2) &\propto p_{\text{on}}(c_1, c_2) \\ &= G_{\text{AND}}(c_1, c_2) = \theta(c_1 - c_{\text{mid}})\theta(c_2 - c_{\text{mid}}), \end{aligned}$$

where $\theta(x)$ is the Heaviside step function. We define the fitness function R as follows. First, we compute $p_{\text{on}}(c_1, c_2)$ for 16 values of (c_1, c_2); for the AND gate in Fig. 2.5, these 4×4 values are depicted as dark dots. For each of those points, we determine how much $p_{\text{on}}(c_1, c_2)$ deviates from the goal function $G_t(c_1, c_2)$; next we compute the sum of the squares of these deviations. If this quantity is small, the fitness is considered high. The following equation summarizes the measure:

$$R = -\sum_{i,j=0}^{3} \left\{ p_{\text{on}}\left(\frac{2i}{3}c_{\text{mid}}, \frac{2j}{3}c_{\text{mid}}\right) - G_t\left(\frac{2i}{3}c_{\text{mid}}, \frac{2j}{3}c_{\text{mid}}\right) \right\}^2 .$$

2.B Affinities of auxiliary sites

One of the main functions of auxiliary binding sites is to create steep responses to changes in TF concentrations. In the results of our simulations, we observed that the auxiliary sites of repressors are often weak, while in case of activator sites they are often strong. Moreover, in activator systems, the auxiliary site furthest removed from the core promoter usually has the highest affinity. Here we demonstrate that these patterns further enhance the steepness of response.

The basic idea is as follows. If the affinity of an auxiliary site is very low, the effect of the site vanishes. On the other hand, if its affinity becomes very large, the auxiliary site will always be occupied. In that case, the auxiliary site merely increases the affinity of the primary site with a constant factor (ω in our model). The effect of this is equivalent to lowering the dissociation constant of the primary site with the same factor, which shows that in this limit the cooperativity is lost as well. Somewhere between these limits, an optimum must be present. This optimum is different for activating sites and repressing sites.

It is possible to analyse the situation for any number of auxiliary sites. Below we show the results for two auxiliary sites. (See Fig. 2.6.)

Repression

We assume that a promoter has one primary repressor site and two auxiliary sites (Fig. 2.6(a)). The primary repressor site O_1 has a dissociation constant K, whereas the auxiliary sites O_2 and O_3 have dissociation constants K/r_2 and K/r_3. The question then is: what values of r_i maximize the steepness of the response?

As before, we compute p_{on} according to Equation A2.3. The partition sums are:

$$\begin{aligned} Z_{\text{on}} &= q_{\text{p}} \left(1 + (r_2 + r_3) \left(\frac{c_{\text{r}}}{K} \right) + r_2 r_3 \omega \left(\frac{c_{\text{r}}}{K} \right)^2 \right), \\ Z_{\text{off}} &= 1 + (1 + r_2 + r_3) \left(\frac{c_{\text{r}}}{K} \right) + (r_2 \omega + r_3 + r_2 r_3 \omega) \left(\frac{c_{\text{r}}}{K} \right)^2 \\ &\quad + r_2 r_3 \omega^2 \left(\frac{c_{\text{r}}}{K} \right)^3 . \end{aligned}$$

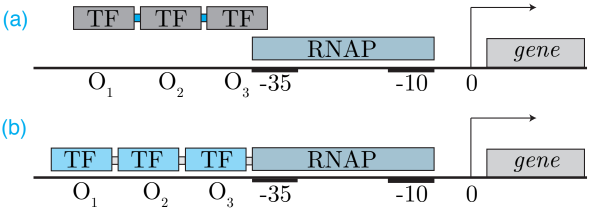


Figure 2.6: Illustration of the repression (a) and activation (b) system discussed. In both cases, the TF has three binding sites; O_1 is in both cases the primary site, while O_2 and O_3 are the auxiliary binding sites.

The concentration of repressor is denoted by c_r.

We use three different measures of the steepness of response:

1. We optimized the slope s_half of the response plots at the TF concentration at which the expression level is half maximal (c_half); we choose K such that $c_\mathrm{half} = 500\,\mathrm{nM}$. The results are shown in Fig. 2.7(a), where we use $\omega = 50$ and $q_\mathrm{p} = 10$. The figure shows that s_half can be increased considerably (69%) by optimizing the relative affinities of the auxiliary sites. The best result is obtained at $r_2 = 0.017$ and $r_3 = 0.091$, confirming that, ideally, repressive auxiliary sites are much weaker than their primary sites.

2. We fitted the response plots to Hill functions, defined as

$$H_\mathrm{r}(c_\mathrm{r}) = A\frac{1 + (c_\mathrm{r}/K)^n/f}{1 + (c_\mathrm{r}/K)^n},$$

   where f is the maximum fold change in expression level and n is the value of the Hill coefficient (Buchler *et al.*, 2003). We optimized the value of the Hill coefficient n as a function of r_2 and r_3. The resulting plots (not shown) are very similar to those found using the first method; n can be increased by 63% by choosing $r_2 = 0.011$ and $r_3 = .057$.

3. We optimize the slope s_inf at the inflection point of the response curve. Now, we choose K such that this point is at 500 nM. Fig. 2.7(c) shows that s_inf can be increased by 70% if we fine-tune the affinities of the different binding sites. Again the auxiliary sites are weak: $r_2 = 0.014$ and $r_3 = 0.11$.

All methods show that weak auxiliary sites of repressor systems are not only sufficient, but even optimal. Therefore it is highly unlikely that evolution would maintain strong auxiliary repression sites, if a steep response is beneficial. Of course, this argument only holds for auxiliary sites that do not have a second function. If an auxiliary site also functions as an anti-activator (*i.e.*, it prevents the binding of an activator by overlapping with its binding site) a higher affinity may be required.

Interestingly, our results also show that site 2 should ideally be weaker than site 3. Note, however, that site 3 can be interpreted as an activator site for site 2; this situation is therefore analogous to the activation system, which we discuss below.

Activation

Here we present the case of cooperative activation by two auxiliary TF binding sites. We use the same conventions as in the previous subsection. For this system,

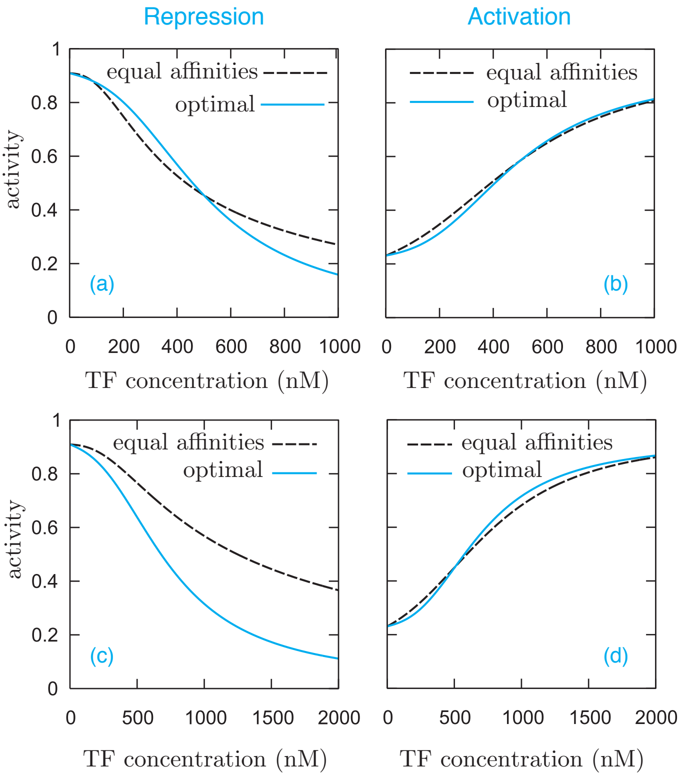


Figure 2.7: Response plots of the cooperative repressor (a and c) and activator (b and d) systems. All plots show the responses at $r_2 = r_3 = 1$ and the ones for optimized parameters. In figures (a) and (b) we optimized the slope at half maximal repression (a) or half maximal activation (b); we fixed this point at 500 nM. In figure (c) and (d) we optimized the slope at the inflection point of the curve, fixing this point at 500 nM. In the repressor system we chose $q_\mathrm{p} = 10$, while in the activation system $q_\mathrm{p} = 0.3$; in both cases $\omega = \omega' = 50$. Clearly, steepness of response of the repressor system increases considerably if the relative affinities of the binding sites are fine-tuned. The same holds for the activation system, albeit to a much lesser extent.

the partition sums become:

$$
\begin{aligned}
Z_{\mathrm{on}} &= q_{\mathrm{p}}\Big(1 + (\omega' + r_2 + r_3)\left(\frac{c_{\mathrm{a}}}{K}\right) + (r_2 r_3 \omega + r_2 \omega\omega' + r_3\omega')\left(\frac{c_{\mathrm{a}}}{K}\right)^2 \\
&\quad + r_2 r_3 \omega^2 \omega' \left(\frac{c_{\mathrm{a}}}{K}\right)^3\Big), \\
Z_{\mathrm{off}} &= 1 + (1 + r_2 + r_3)\left(\frac{c_{\mathrm{a}}}{K}\right) + (r_2\omega + r_3 + r_2 r_3 \omega)\left(\frac{c_{\mathrm{a}}}{K}\right)^2 \\
&\quad + r_2 r_3 \omega^2 \left(\frac{c_{\mathrm{a}}}{K}\right)^3,
\end{aligned}
$$

where c_{a} is the concentration of the activator TF.

Again we use three different measures for the steepness of response.

1. We optimize the slope s_{half} at the concentration c_{half} at with the expression level is half maximal. We adjust K such that $c_{\mathrm{half}} = 500\,\mathrm{nM}$. The results are shown in Fig. 2.7(b), where we use $\omega = \omega' = 50$ and $q_{\mathrm{p}} = 0.3$. The optimal parameter set, $r_2 = 1.74$ and $r_3 = 12.3$ provides an increase in s_{half} of 11%. Note that the affinity of O_3 is much higher than those of the other sites.

2. We fit the plots to the Hill function defined as:

$$H_{\mathrm{a}}(c_{\mathrm{a}}) = A\frac{f^{-1} + (c_{\mathrm{a}}/K)^n}{1 + (c_{\mathrm{a}}/K)^n}. \tag{A2.8}$$

   The results (not shown) are very similar to those obtained by the previous method. The gain in terms of n is a modest 11%.

3. We maximize the slope s_{inf} at the inflection points of the plots, adjusting K such that $c_{\mathrm{inf}} = 500\,\mathrm{nM}$ (Fig. 2.7(d)). Optimally, $r_2 = 1.97$ and $r_3 = 14.4$, which results in a 27% increase in s_{inf}.

The results show that, in order to be optimal, the auxiliary activation sites need to be as strong or stronger than the primary site. This is in stark contrast with the results for homo-cooperative repression, where we saw that the auxiliary sites need to be weak. Also, the site furthest removed from the core promoter has the highest affinity, as we found in our simulations. The increase in steepness resulting from the tuning of the binding affinities, however, is rather modest. Whether in real genetic systems the selection pressure for steep activation is usually strong enough to attain and maintain the optimal affinity ratios in a mutation–selection balance, is unclear.

2.C Minimal models for the complex gates

Some of the gates that resulted from the simulations have a rather complex design. Here, we describe simplified quantitative models for the EQU gate and the XOR gate; a simplified description provides more insight into their essential features. The other gates can be described in a similar manner.

The EQU gate

For the EQU gate, the essential ingredients of our minimal model are: a strong promoter, homo-cooperative repression for each of the two TFs, and hetero-cooperative activation when both TFs are present. For simplicity, we make the following assumptions:

1. All repression sites have an equal dissociation constant K_r; all activation sites have dissociation constant K_a.
2. The number of sites in each homo-cooperative repression module is the same and equal to n_r; the number of sites for the TF α in the hetero-cooperative activation module is $n_{\mathrm{a},\alpha}$.
3. We neglect states in which incomplete modules are bound; of a module either all sites or none of the sites are occupied.
4. The modules exclude each other on the DNA: only one of the modules can be bound at a time.
5. We assume that the TFs bind to their specific binding sites only; we thus neglect the affinities for the other binding sites on the DNA.

For this minimal model we can compute the partition sums as follows:

$$Z_\mathrm{off} = 1 + (q_{\mathrm{r},1})^{n_\mathrm{r}}\omega^{n_\mathrm{r}-1} + (q_{\mathrm{r},2})^{n_\mathrm{r}}\omega^{n_\mathrm{r}-1} + (q_{\mathrm{a},1})^{n_{\mathrm{a},1}}(q_{\mathrm{a},2})^{n_{\mathrm{a},2}}\omega^{n_{\mathrm{a},1}+n_{\mathrm{a},2}-1}, \tag{A2.9}$$

$$Z_\mathrm{on} = q_\mathrm{p}\left(1 + \omega'(q_{\mathrm{a},1})^{n_{\mathrm{a},1}}(q_{\mathrm{a},2})^{n_{\mathrm{a},2}}\omega^{n_{\mathrm{a},1}+n_{\mathrm{a},2}-1}\right). \tag{A2.10}$$

Here we used:

$$q_{r,\alpha} = \frac{c_\alpha}{K_\mathrm{r}}, \qquad q_{a,\alpha} = \frac{c_\alpha}{K_\mathrm{a}}. \tag{A2.11}$$

Note that in Z_off we do not only count states in which the repression modules are bound, but also states in which the activation sites are occupied by TFs (but with no RNAp bound). Note also that Z_off and Z_on are bivariate polynomials in the concentrations c_α. The order of these polynomials is determined by the number of binding sites in the modules; the coefficients of each term are set by the dissociation constants. Equation A2.7 shows that p_on can be written in terms of the ratio of these polynomials.

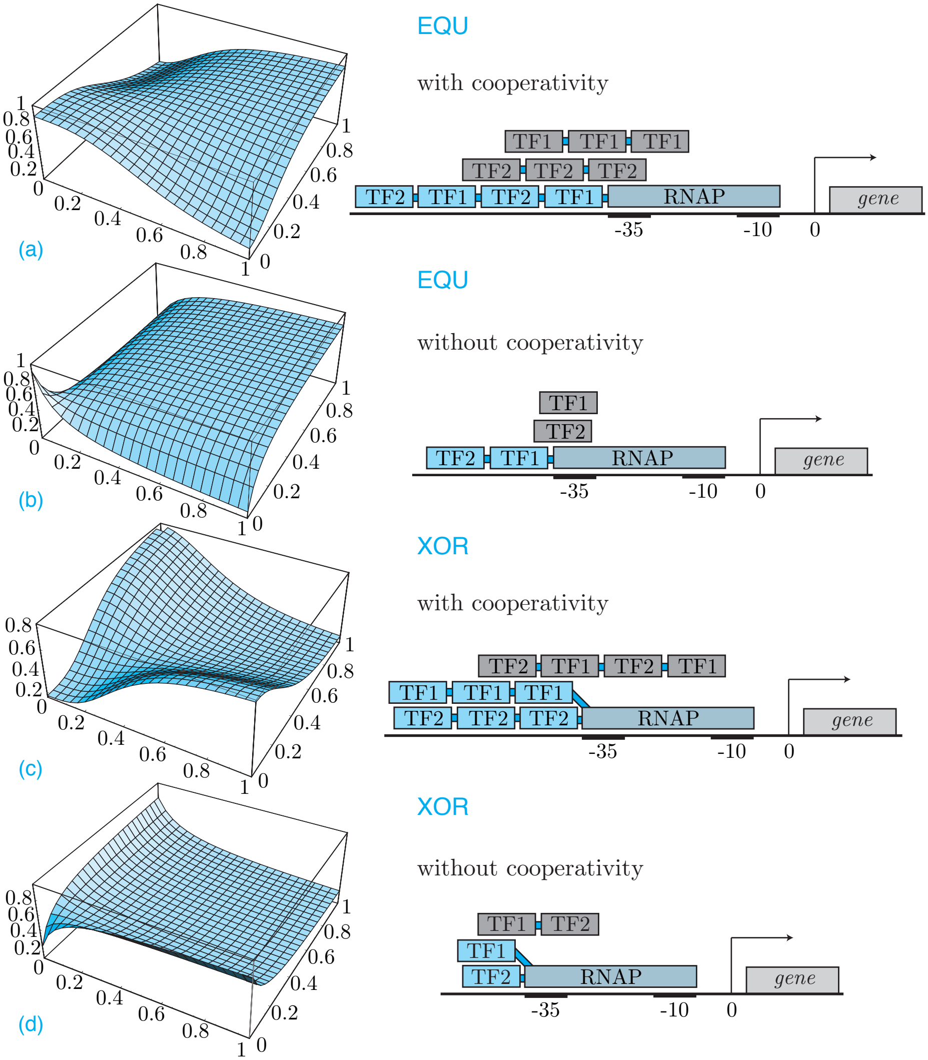


Figure 2.8: Response plots resulting from the simplified models of XOR and EQU gates. The concentration units are μM. (a) EQU gate with the following parameters: $q_\mathrm{p} = 6$, $K_\mathrm{a} = 11\mu$M, $K_\mathrm{r} = 3\mu$M, $n_{\mathrm{a},\,1} = n_{\mathrm{a},\,2} = 2$, $n_\mathrm{r} = 3$ and $\omega = 30$. (b) An EQU gate without homo-cooperative repression modules ($n_{\mathrm{a},\,1} = n_{\mathrm{a},\,2} = n_\mathrm{r} = 1$). Note that, although the values in the corners of the plot are consistent with an EQU gate, the full performance is poor. This shows that the complex behavior of the EQU gate requires homo-cooperative modules. Further parameters are: $q_\mathrm{p} = 10$, $K_\mathrm{a} = 3\mu$M, $K_\mathrm{r} = 0.01\mu$M and $\omega = 30$. (c) XOR gate with parameters $q_\mathrm{p} = 0.2$, $K_\mathrm{a} = 7\mu$M, $K_\mathrm{r} = 4\mu$M, $n_{\mathrm{r},\,1} = n_{\mathrm{r},\,2} = 2$, $n_\mathrm{a} = 3$ and $\omega = 30$. (d) Typical XOR gate with no homo-cooperative activation ($n_{\mathrm{r},\,1} = n_{\mathrm{r},\,2} = n_\mathrm{a} = 1$). The gate could hardly be classified as an XOR gate, showing that homo-cooperative activation is essential to obtain reasonable XOR gates.

We now consider the design constraints for obtaining an input–output relation that corresponds to an EQU gate. To this end, we first consider the limit in which one of the TFs is present in much larger concentration than the other. An EQU gate requires that in this limit, the expression level, and thus p_{on}, should be low. When the concentration c_1 is kept constant and c_2 is increased, p_{on} approaches a limit value that is determined by the terms of highest order in c_2 in Z_{off} and Z_{on}. It is given by

$$\lim_{c_2\to\infty} p_{\mathrm{on}} = \begin{cases} \dfrac{\omega' q_{\mathrm{p}}}{1+\omega' q_{\mathrm{p}}} & \text{if } n_{\mathrm{r}} < n_{\mathrm{a},\,2}, \\[2ex] \dfrac{\omega' q_{\mathrm{p}} (c_1)^{n_{\mathrm{a},\,1}}}{(K_{\mathrm{a}}/K_{\mathrm{r}})^{n_{\mathrm{r}}} (K_{\mathrm{a}}/\omega)^{n_{\mathrm{a},\,1}} + (c_1)^{n_{\mathrm{a},\,1}} (1+\omega' q_{\mathrm{p}})} & \text{if } n_{\mathrm{r}} = n_{\mathrm{a},\,2}, \\[2ex] 0 & \text{if } n_{\mathrm{r}} > n_{\mathrm{a},\,2}. \end{cases}$$

If $n_{\mathrm{r}} < n_{\mathrm{a},\,2}$, then p_{on} will approach unity, instead of zero as required: since for the EQU gate the promoter should be strong, $\omega' q_{\mathrm{p}}/(1+\omega q_{\mathrm{p}}) \approx 1$. If $n_{\mathrm{r}} = n_{\mathrm{a},\,2}$, the expression level depends on K_{a}, K_{r}, n_{r} and $n_{\mathrm{a},\,1}$; a judicious choice of their value can allow for an expression level that is consistent with an EQU gate. If, however, $n_{\mathrm{r}} > n_{\mathrm{a},\,2}$, then the expression level in the above limit is much more robust to the precise parameter values: if the concentration of TF1 is kept constant, then at sufficiently high concentration of TF2, TF2 will always repress transcription, as required for an EQU gate.

We now consider the scenario in which both concentrations become large. If we keep $c_1 = c_2$ and increase both concentrations, then, as long as $n_{\mathrm{a},\,1} + n_{\mathrm{a},\,2} > n_{\mathrm{r}}$, the limit value is again

$$\lim_{c_1,c_2\to\infty} p_{\mathrm{on}} = \frac{\omega' q_{\mathrm{p}}}{1+\omega' q_{\mathrm{p}}}. \tag{A2.12}$$

A good way to construct an EQU gate is therefore to choose the modules such that $n_{\mathrm{a},\,1} + n_{\mathrm{a},\,2} > n_{\mathrm{r}}$ (so that the operon is transcribed when c_1 and c_2 are both high), but to take $n_{\mathrm{r}} > n_{\mathrm{a},\,1}$ and $n_{\mathrm{r}} > n_{\mathrm{a},\,2}$ (so that the operon is repressed when only one of the two TF concentrations is high). One obvious choice is $n_{\mathrm{a},\,1} = n_{\mathrm{a},\,2} = 2$ and $n_{\mathrm{r}} = 3$. This result is shown in Fig. 2.8(a). It is seen that this gate can indeed be classified as an EQU gate.

The EQU gate that results from our simulations (see Fig. 2.4) deviates slightly from this design (Fig. 2.8(a)): the number of repressor sites of TF1 is higher than expected on the basis of the assumptions of the minimal model, so that the requirement $n_{\mathrm{a},\,1} + n_{\mathrm{a},\,2} > n_{\mathrm{r}}$ is not fulfilled. However, there are three points worthy of note:

1. most of the repressor sites are very weak; the extra repressor sites only play a major role at much higher TF concentrations than shown in Fig. 2.5;
2. the assumption of the minimal model that the modules mutually exclude each other completely, while instructive, is not entirely consistent with

the assumptions underlying the full model discussed in this chapter: it is possible for the complete hetero-cooperative activation module to bind, while simultaneously the repressor sites that do not overlap with the activation module, are also occupied;

3. whereas the previous points concern the simplicity of the assumptions of the minimal model, this point is more fundamental. In our simulations, we selected the gates not just based on their behavior in the limits of high concentrations: we also selected for a *steep* repression curve. The resulting gate is thus a compromise between the requirement of a steep response — favoring a high number of repression sites — and maximal activation when both TFs are present.

Fig. 2.8(c) shows the result for an EQU gate with no homo-cooperative repression modules ($n_{\mathrm{a},\,1} = n_{\mathrm{a},\,2} = n_{\mathrm{r}} = 1$). In the limit that $c_1 \to 0$ and $c_2 \to \infty$ and in the limit that $c_1 \to \infty$ and $c_2 \to 0$, the expression level approaches zero, as required for an EQU gate. Nevertheless, the input–output relation differs markedly from the gate with homo-cooperativity (Fig. 2.8(a)); indeed, one could argue that the gate without homo-cooperativity does not classify as an EQU gate. This shows that homo-cooperativity does not only allow for a steep response, but also can play an important role in signal integration.

The XOR gate

For the XOR gate, the essential ingredients of the minimal model are: a weak promoter, homo-cooperative activation by each of the two TFs, and hetero-cooperative repression when both TFs are present. We make the same simplifying assumptions as in the previous section. However, here the number of sites in each of the activation complexes is denoted by n_{a}, while the number of sites of TF α in the hetero-cooperative repression complex is $n_{\mathrm{r},\alpha}$. This results in the following expressions:

$$\begin{aligned}
Z_{\mathrm{off}} &= 1 + (q_{\mathrm{r},1})^{n_{\mathrm{r},\,1}} (q_{\mathrm{r},2})^{n_{\mathrm{r},\,2}} \omega^{n_{\mathrm{r},\,1}+n_{\mathrm{r},\,2}-1} + (q_{\mathrm{a},1})^{n_{\mathrm{a}}} \omega^{n_{\mathrm{a}}-1} + (q_{\mathrm{a},2})^{n_{\mathrm{a}}} \omega^{n_{\mathrm{a}}-1}, \\
Z_{\mathrm{on}} &= q_{\mathrm{p}} \left(1 + \omega' (q_{\mathrm{a},1})^{n_{\mathrm{a}}} \omega^{n_{\mathrm{a}}-1} + \omega' (q_{\mathrm{a},2})^{n_{\mathrm{a}}} \omega^{n_{\mathrm{a}}-1}\right).
\end{aligned}$$

When both TFs are absent, the operon should be off; therefore an XOR gate needs a weak promoter. When increasing c_2 at constant c_1, or c_1 at constant c_2, activation should occur. The limit value of p_{on} for $c_\alpha \to \infty$ depends on n_{a} and $n_{\mathrm{r},\alpha}$; if $n_{\mathrm{a}} > n_{\mathrm{r},\alpha}$, activation wins the competition with repression. In the limit of high concentrations of both TFs ($c_1 = c_2$, $c_1 \to \infty$), the XOR should be off. This is satisfied if $(n_{\mathrm{r},\,1} + n_{\mathrm{r},\,2}) > n_{\mathrm{a}}$. One option is therefore to choose $n_{\mathrm{r},\,1} = n_{\mathrm{r},\,2} = 2$ and $n_{\mathrm{a}} = 3$. Fig. 2.8 shows the result for this minimal model.

The XOR gate that results from our simulations (see Fig. 2.4 and 2.5), again deviates slightly from this design. The number of repressor sites of TF2 is higher than anticipated, so that the requirement $n_{\mathrm{a}} > n_{\mathrm{r},\,2}$ is not fulfilled. As for the

EQU gate, on the one hand this is due to the simplicity of the minimal model, while on the other hand it is a result of the selection for a steep response.

Fig. 2.8(d) shows the result for an XOR gate without homo-cooperative activation modules — the activation when either TF1 or TF2 is present, is non-cooperative ($n_{\mathrm{r},\,1} = n_{\mathrm{r},\,2} = n_{\mathrm{a}} = 1$). It is seen that the performance of the gate is poor. This again shows that homo-cooperativity can be a useful mechanism for shaping complex input–output relations.

2.D Extending the model

Our simplified model can easily be extended.

One clear limitation of the model is that all transcription factors interact with the same energy. In reality, the interaction between some TFs is strongly cooperative, whereas others do not seem to recruit each other at all. This limitation of the model can be removed if we allow the TF–TF interaction energies between all TFs to evolve independently. An elegant way to implement this, is to endow each TF with two interaction surfaces: one at the "front" of the TF and one at the "back". These surfaces can be represented as lists of amino acids with length L. We assume that TFs that bind close together interact through direct contact between their interaction surfaces. Then we can model the TF–TF interactions analogous to the TF–DNA interactions: we assume that the interaction energy is the sum of independent amino-acid−amino-acid interactions, *i.e.*

$$E_{a,b} = \sum_{i=1}^{L} V_{a_i,b_i}.$$

Here a and b are the amino-acid vectors describing the interaction surfaces of both transcription factors, and $V_{\lambda\mu}$ is a 20×20 matrix containing the binding free energies associated with each amino-acid−amino-acid contact. The recursive algorithm can be adjusted to take into account the different interaction energies without significant changes in its complexity.

In the current model, each TF has length M. It would be possible to allow for varying TF sizes. This again requires adjustments to the algorithm, but in this adjusted scheme the computation time will still scale linearly with all relevant parameters.

Chapter 3

Auto-regulation and bi-stability in transcription regulation

The transcription regulatory network of *Escherichia coli* contains very few feedback loops. Yet, feedback can be of great importance. For instance, a surprisingly large fraction of the transcription factors in *E. coli* regulate their own expression. A second example is the famous bacteriophage λ switch: this bi-stable system functions by virtue of feedback loops.

In Chapter 2 we used a statistical-mechanical model of transcription regulation and an evolutionary algorithm to design transcriptional logic gates. There, feedback was excluded by the method. Here, we again design small functional networks, but use an adjusted scheme in which feedback *can* be exploited. First, we apply the method to construct a bi-stable switch. Interestingly, the resulting designs are very reminiscent of the genetic switch of bacteriophage λ. Second, we again design transcriptional logic gates, but now allow them to use auto-regulation. We analyze the design principles of the resulting logic gates and confirm that they use auto-regulation to shape their response functions.

3.1 Introduction

In control theory, feedback is a crucial concept. Feedback loops can be used in control networks to construct, among others, memory modules and oscillators, or to increase the robustness of systems. Indeed, loops are exploited in regulatory networks ranging from neural and metabolic networks to the cruise controls of cars. Therefore it is perhaps surprising that in prokaryotic transcription networks feedback loops are actually very *rare* (Shen-Orr *et al.*, 2002). There are, however, a few exceptions to this rule.

A first exception is the famous genetic switch of the bacteriophage λ (Ptashne, 2004). The core of this switch is provided by two genes, *cI* and *cro*, coding for two transcription factors (TFs), Cro and CI, that repress each others transcription. As a result, the system has two stable states: one in which the concentration of CI is high and *cro* is repressed, and another in which the concentration of Cro is high and *cI* is repressed. The two steady states are crucial for the survival and spreading tactics of the phage, which rely on its ability to switch between two different life styles (Jacob and Monod, 1961b; Ptashne, 2004).

When it first infects an *E. coli* cell, the phage often incorporates itself in the DNA of the bacterium and stays there in a dormant state. During every division of the *E. coli* cell, the phage is copied along with the DNA of *E. coli* and transmitted to the daughter cells. This stage is called the *lysogenic* state. At the molecular level, the lysigenic state corresponds to the condition in which the repressor CI is present in high concentrations and all other genes are turned off. This peaceful life style may last for many cell generations. But if the *E. coli* DNA is damaged, for instance by UV radiation, the situation drastically changes. In response to the damage, the *E. coli* cell activates the stress (SOS) response system. This involves an increase in the cellular concentration of the protein RecA; this enzyme plays a key role in the SOS response but as a side effect also cleaves CI. Hence, the repression of *cro* gene is released and the genetic switch flips. This has dramatic consequences: it activates a cascade in which the λ phage enters the aggressive *lytic* state. Now, the DNA of the phage is copied many times and the heads and tails of the viral particle are produced. About 90 minutes later, the *E. coli* cell bursts open, releasing many copies of the phage that are ready to infect a new cell.

Another feedback motive in transcription regulation is auto-regulation. Shen-Orr *et al.* (2002) showed that, in *E. coli*, many TFs regulate their own transcription; in fact, many more than would be expected in randomized networks. As the data set originally used by Shen-Orr *et al.* is now outdated, we analyzed the occurrence of auto-regulation in the currently known transcription regulatory network using the RegulonDB database (Salgado *et al.*, 2001). We found 95 TFs with known binding sites in the intergenic region directly upstream of the operon that codes for them; this amounts to 59% of the known transcription factors. Of these 95 TFs, 71 auto-repress and 32 auto-activate (8 TFs have binding sites for auto-activation as well as for auto-repression). Clearly, auto-repression occurs more often than auto-activation, but both are strongly over-represented.

Why is auto-regulation this common? In regard to auto-repression a number of functions have been suggested. In the first place, auto-repression can contribute to the robustness of the expression level with respect to fluctuations in the transcription rate (Savageau, 1974; Becskei and Serrano, 2000). The second suggestion is that negative feedback can be used to increase the response speed of the regulated gene (Rosenfeld *et al.*, 2002). In the third place, negative feedback can lead to oscillations in the presence of time delays (Elowitz and Leibler, 2000). All three effects have been demonstrated *in vivo* (Becskei and Serrano, 2000; Rosenfeld *et al.*, 2002; Elowitz and Leibler, 2000). On the flip side, negative feedback can reduce the sensitivity to input signals (Hornung and Barkai, 2008).

Positive feedback tends to have an effect opposite to negative feedback: it slows down responses and tends to amplify intrinsic fluctuations. At first sight, it seems unlikely that these qualities are often beneficial. Yet, a slow response can be used as a low frequency filter combining a sensitive response to persisting signals with a filtering of fast fluctuations in the input signal (Hornung and Barkai, 2008). A special feature of auto-activation that may occasionally be useful is that it can lead to bi-stability (refer to Box 3.1 for details).

Here, we propose that auto-activation may be useful in *shaping the response functions of* cis-*regulatory constructs*. Transcription factors that auto-regulate are typically regulated by other TFs too: in fact, we found 23 auto-regulating transcription factors (72% of the auto-activators) that are known to have at least two additional inputs (other TFs). The response of the regulated TF to changes in the input concentrations must therefore be the result of an interplay between regulation and auto-regulation. Conversely, this strongly suggests that auto-regulation could be used to shape these responses.

In Chapter 2 we studied the mechanisms of transcriptional regulation; we used an evolutionary algorithm to design transcriptional logic gates and extracted design principles from the results. In that study, feedback was excluded by the model. Here, we use an adapted version of the method to design small networks that do exploit feedback.

First, we use the method to design a genetic switch consisting of two genes, each coding for TFs. We describe the dynamics of the TF concentrations as ordinary differential equation and select for networks that have two stable states. During the evolutionary process, binding sites for both TFs can emerge on the *cis*-regulatory regions of both genes. Surprisingly, the resulting switches are conceptually very similar to the phage λ switch: apart from mutual repression, they exploit auto-activation.

Second, we again design transcriptional logic gates, but this time allow the designs to use feedback. It turns out that many of the resulting designs do use auto-regulation; we describe two mechanisms that are at work in several gates.

Box 3.1: Conditions for bi-stability due to positive feedback

If a TF activates its own transcription, this can lead to a bi-stable system. However, this depends crucially on the shape of the function $p_{\mathrm{on}}(c)$, which characterizes the response of the transcription rate as a function of the activator concentration c. Here, we derive strong conditions for bi-stability, inspired by the formalism of Cherry and Adler (2000).

We assume that the dynamics of the concentration c can be described by:

$$\frac{\mathrm{d}c}{\mathrm{d}t} = a\, p_{\mathrm{on}}(c) - d\, c. \tag{3.1}$$

Equilibrium values of c then should satisfy $p_{\mathrm{on}}(c) = (d/a)c$, or equivalent, the equilibria are those values where the functions $p_{\mathrm{on}}(c)$ and $y(c) \equiv (d/a)c$ cross. Since $p_{\mathrm{on}}(0) \geq 0 = y(0)$ and $p_{\mathrm{on}}(a/d) \leq 1 = y(a/d)$, the Intermediate Value Theorem implies that such a crossing should occur at least once in the interval $c \in [0, a/d]$ (assuming $p_{\mathrm{on}}(c)$ is continuous). In order to have multiple equilibria, at least three crossings are required, two of which correspond to stable equilibria and one of which is unstable. In an unstable equilibrium, $p'_{\mathrm{on}}(c) \geq d/a$). If there is one unstable equilibrium in the interval $c \in \langle 0, a/d \rangle$, this directly implies that there are (at least) two stable ones too; therefore, the system is bi-stable if there is one value c^* for which $p_{\mathrm{on}}(c^*) = (d/a)c^*$ and $p'_{\mathrm{on}}(c^*) > d/a$. Conversely, if the system if bi-stable, there is a value c^* for which $p_{\mathrm{on}}(c^*) = (d/a)c^*$ and $p'_{\mathrm{on}}(c^*) \geq d/a$.

To derive if a given function $p_{\mathrm{on}}(c)$ can give rise to multi-stability for *at least some* values of d and a, we can combine the two requirements above. We then obtain that bi-stability is possible if there is a value $c > 0$ for which $c^* p'_{\mathrm{on}}(c^*) > p_{\mathrm{on}}(c^*)$ or, equivalently,

$$\sup_{c>0} \left(\frac{c\, p'_{\mathrm{on}}(c)}{p_{\mathrm{on}}(c)} \right) > 1. \tag{3.2}$$

Conversely, if the system is bi-stable, $\sup_{c>0} \left(\frac{c\, p'_{\mathrm{on}}(c)}{p_{\mathrm{on}}(c)} \right) \geq 1$.

Applied to the case where $p_{\mathrm{on}}(c)$ is a simple Hill function,

$$p_{\mathrm{on}}(c) = \frac{c^n}{K^n + c^n}, \tag{3.3}$$

one can show that

$$\sup_{c>0} \left(\frac{c\, p'_{\mathrm{on}}(c)}{p_{\mathrm{on}}(c)} \right) = \sup_{c>0} \left(\frac{nK^n}{K^n + c^n} \right) = n. \tag{3.4}$$

From this we conclude that, if $p_{\mathrm{on}}(c)$ is a Hill function, bi-stability is possible for certain values of a, d and K, provided $n > 1$.

We note that the above analysis relies on a deterministic description in terms of real-valued concentrations. This is not always justified; in models that take into account the stochastic character of interactions and discrete particle numbers, bi-stability could occur even at $n = 1$ (Lipshtat *et al.*, 2006).

3.2 Models and methods

Below, we discuss the models we used to design the genetic switch and the logic gates.

3.2.1 The genetic switch

To design switches, we consider networks consisting of 2 genes, *tf1* and *tf2*, that each code for a transcription factor: TF1 and TF2 respectively. Each of these TFs can regulate the transcription of both genes by binding to their promoter regions. Therefore the transcription rate of each gene is a function of the concentrations $\tilde{c}_i$ of both TFs.

Transcription rates

In order to compute the transcription rate of each gene, we use the same formalism as in Chapter 2. To summarize, each TF can bind to all sites on both *cis*-regulatory regions, with affinities that depend on the sequence of the site and the sequence of the TF (cf. Section 2.2). Steric hindrance prevents TFs from binding to overlapping sites, yet TFs binding close to each other (within a distance of $k = 3\,\mathrm{bp}$) experience a cooperative interaction energy $E_{\mathrm{TF-TF}}$. As before, the transcription rate of a gene i is assumed to be proportional to the fractional occupancy of its (core) promoter, denoted by $p_{\mathrm{on}}^{(i)}(\tilde{c}_1, \tilde{c}_2)$. The calculation of these occupancies is completely analogous to our method in Chapter 2.

Evolutionary algorithm

As in Chapter 2, we use an evolutionary algorithm to design our networks. We again subject a population of ≈ 200 elementary networks to rounds of mutation and selection. However, our selection procedure is quite different from the one used before. In Chapter 2, we designed transcriptional logic gates and therefore directly selected for specific functional forms of $p_{\mathrm{on}}(\tilde{c}_1, \tilde{c}_2)$. Now, we are not directly interested in these input–output relations, but instead base our selection scheme on the dynamical properties of the two-gene networks.

Dynamics of the concentrations

We model the dynamics of the concentrations of both TFs by a system of non-linear ordinary differential equations:

$$\frac{\mathrm{d}\tilde{c}_1}{\mathrm{d}\tilde{t}} = a\, p_{\mathrm{on}}^{(1)}(\tilde{c}_1, \tilde{c}_2) - d\, \tilde{c}_1, \tag{3.5}$$

$$\frac{\mathrm{d}\tilde{c}_2}{\mathrm{d}\tilde{t}} = a\, p_{\mathrm{on}}^{(2)}(\tilde{c}_1, \tilde{c}_2) - d\, \tilde{c}_2. \tag{3.6}$$

Here we assumed first-order degradation of both TFs, with degradation constant d, and that the protein production rate is proportional to $p_{\mathrm{on}}^{(i)}(\tilde{c}_1, \tilde{c}_2)$ with proportionality constant a. Note that this deterministic description neglects all sources of stochasticity and ignores time delays in the transcription and translation process.

It is convenient to introduce new, dimensionless variables. The concentrations of the TFs are maximal if the genes are transcribed at their maximal rate, which means that $p_{\text{on}}^{(i)}(\tilde{c}_1, \tilde{c}_2) = 1$. In this case the steady state concentrations are easily derived from Equation 3.5: $\tilde{c}_{\text{max}} = a/d$. We define the new dimensionless concentrations $c_i(t)$ as

$$c_i(t) \equiv \frac{d}{a}\tilde{c}_i(\tilde{t}), \qquad \text{where} \quad t \equiv d\,\tilde{t}. \tag{3.7}$$

As $c_{\text{max}} = (d/a)\tilde{c}_{\text{max}} = 1$, the scaled concentrations c_i are guaranteed to stay within the interval $[0, 1]$; their dynamics are described by:

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = p_{\text{on}}^{(1)}(c_1, c_2) - c_1, \tag{3.8}$$

$$\frac{\mathrm{d}c_2}{\mathrm{d}t} = p_{\text{on}}^{(2)}(c_1, c_2) - c_2. \tag{3.9}$$

The nullclines of this system of differential equations are implicitly given by the relations $p_{\text{on}}^{(1)}(c_1^*, c_2) = c_1^*$ and $p_{\text{on}}^{(2)}(c_1, c_2^*) = c_2^*$. The intersections of these nullclines are the equilibrium points of the system. In a bi-stable system, at least three equilibrium points are required, one of which is unstable (see Cherry and Adler (2000)). Our evolutionary algorithm is designed to shape the functions $p_{\text{on}}^{(i)}(c_1, c_2)$ such that two stable equilibrium points are obtained.

Fitness function

We select for two stable states: one in which $c_1 = c_{\text{max}} \equiv 1$ ("high") and one in which $c_1 = c_{\text{min}} \equiv 0$ ("low"). We therefore use a fitness function that heuristically measures the stability of these states. To that end, we numerically propagate the system of differential Equations 3.8 for two different initial conditions. In the first initial condition, c_1 is rather high, and we monitor if c_1 stays high. In the second initial condition, c_1 is quite low and we measure if it stays low.

More precisely, we first use the starting condition $(c_1, c_2) = (0.8, 0.2)$. In this case, we want $c_1(t)$ to approach c_{max} as time progresses; in order to quantify to what extent this happens, we record $c_1(t)$ at discrete time points $t_1 = 1.5, t_2 = 3.0, \ldots, t_8 = 12.0$ and compute $F_1 \equiv \sum_{i=1}^{8}(c_{\text{max}} - c_1(t_i))^2$. If this number is low, the system performs well on the first test. The second time we propagate the system with initial condition $(c_1, c_2) = (0.2, 0.8)$ and require $c_1(t)$ to converge to c_{min}. This time we compute $F_2 \equiv \sum_{i=1}^{8}(c_{\text{min}} - c_1(t_i))^2$; low values of F_2 again indicate good performance. The fitness of the system is now defined as

$$F \equiv 8 - (F_1 + F_2).$$

The constant 8 is included to ensure that the fitness values are always positive. Note that we implicitly do not only select for bi-stability, but also for a rapid convergence to the stable states.

A similar technique has been used before by Francois and Hakim (2004). The main difference with their approach is that we explicitly use an underlying model of transcription regulation. As a result, we do not fix the response functions of the genes *a priori*; instead, they are determined by the architectures of the promoters, which evolve at the level of the base-pair and amino-acid sequences. Consequently, the evolutionary algorithm can exploit a wide range of nonlinear response functions. In Francois and Hakim (2004), the response functions are fixed beforehand, and evolution acts at the level of reaction equations and rate constants. Contrary to our model, their model also includes additional reactions between gene products such as (hetero)dimerization and active degradation.

3.2.2 Logic gates

We now discuss the setup we use for the design of the logic gates. The model considers two transcription factors, TF1 and TF2, regulating one gene, *tf3*, which itself codes for a transcription factor TF3. All TFs are allowed to bind to the *cis*-regulatory region of *tf3*. We do not impose that auto-regulation should occur, but the system is free to exploit it by developing binding sites for TF3.

The models of TF binding and transcription regulation are precisely the same as in Chapter 2. However, in order to compute the output concentration of TF3, we now need to take into account the effects of auto-regulation. We do this in the following way. Below, we describe our model for the dynamics of the concentration of TF3. Given input concentrations of TF1 and TF2 (c_1 and c_2), the concentration of TF3 will converge under these dynamics towards a steady state. This steady state is considered the output of the gate.

Dynamics of concentration of TF3

Analogous to Equation 3.8, we model the dynamics of the concentration $\tilde{c}_3(\tilde{t})$ of TF3 by the following differential equation:

$$\frac{\mathrm{d}\tilde{c}_3}{\mathrm{d}\tilde{t}} = a\,p_{\mathrm{on}}(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3) - d\,\tilde{c}_3, \tag{3.10}$$

which can again be simplified using dimensionless variables $c_i(t) = (a/d)\tilde{c}_i(\tilde{t})$:

$$\frac{\mathrm{d}c_3}{\mathrm{d}t} = p_{\mathrm{on}}(c_1, c_2, c_3) - c_3. \tag{3.11}$$

Note that p_{on} is now a function of all three TF concentrations. Two of these concentrations, c_1 and c_2, are considered to be the inputs of the gate. Assuming that the system is mono-stable, each set of input concentrations (c_1, c_2) defines an equilibrium value c_3^* which is considered the output of the system. This equilibrium value obeys the condition $c_3^* = p_{\mathrm{on}}(c_1, c_2, c_3^*)$. It can be computed straightforwardly by propagating the differential equation 3.11 for any initial condition $c_3(0)$ until $c_3(t)$ converges.

Since auto-activation is allowed, the system *can* in principle become bi-stable (see Box 3.1). In bi-stable systems, the output concentration c_3^* is not uniquely defined by the input concentrations, as it depends on the initial condition $c_3(0)$. Hence, such systems do not qualify as logic gates. Our fitness function (described below) is therefore designed to penalize bi-stability. However, as the method does not exclude bi-stability for all possible input values, we also check *a posteriori* if the results are bi-stable.

Fitness function

The fitness function that we use is of the same type as the one in Chapter 2 (see 2.A), except that we now consider c_3^* to be the output of the gates in stead of p_{on}. To avoid bi-stability, we always compute the steady state value c_3^* *twice* for a given input (c_1, c_2): once using initial condition $c_3(0) = 0$, and once for $c_3(0) = 1$. In the fitness function, both resulting steady state values are compared to the goal function; if they are significantly different, at least one of them must deviate significantly from the goal, which directly results in a fitness penalty.

3.3 Results

We now turn to the results of our simulations; we start with the genetic switch and then continue with the logic gates.

3.3.1 The genetic switch

Within less than 1000 generations, a good switch is found; in the subsequent generations the result changes very little. In different runs, with various random seeds and population sizes, we always find very similar response curves and promoter architectures, irrespective of the initial DNA and TF sequences.

Network and promoter architectures

The basic architecture of the final network is that of the classical toggle switch (Cherry and Adler, 2000; Warren and ten Wolde, 2004a, 2005). It is remarkably similar to the bacteriophage λ switch. Gene *tf1* cooperatively represses gene *tf2* and vice versa. However, the promoter regions (see Fig. 3.1) show that both genes also exploit auto-activation, which stabilizes the steady states and increases the speed at which these are reached. The phage λ uses this mechanism too: it is known that the λ repressor CI not only represses Cro, but also auto-activates its own expression by recruiting RNAp to its promoter. Indeed, mutants of the phage that do not auto-activate show a strongly reduced stability of the lysogenic state (Ptashne, 2004).

The mutual repression is achieved by homo-cooperative repression modules. These modules directly overlap with the core promoter, but also completely

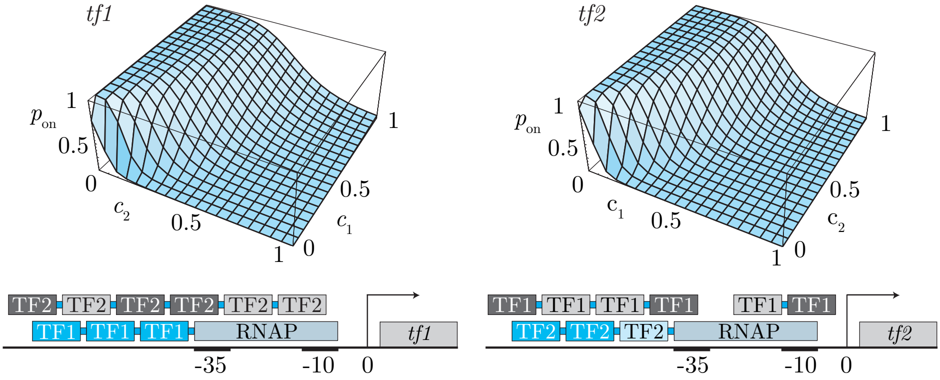


Figure 3.1: Response plots and promoters of the genetic switch emerging from the simulations(Cherry and Adler, 2000; Warren and ten Wolde, 2004a, 2005). Binding sites that have an activating effect are depicted in blue; gray sites lead to repression of transcription. Dark colors denote high affinities, light colors mean low affinites. The basic architecture of the final network is that of the classical toggle switch. Note that the two genes have nearly identical response plots, apart from the interchanged axis labels. Both the cartoons of the *cis*-regulatory regions and the response plots show that both genes auto-activates and represses each other. The cooperativity of the activation and repression modules results in steep response curves, which is required for bistability (Cherry and Adler, 2000; Warren and ten Wolde, 2004a, 2005).

overlap the auto-activation modules. The repression modules therefore act at the same time as direct repressors and as anti-activators. Again, this feature is mirrored in the phage λ. The Cro binding sites all overlap with the CI operators. Therefore Cro not only directly represses *cI*, but also prevents auto-activation by CI; Cro is both a repressor and an anti-activator of *cI*.

The response plots of both genes are remarkably similar (see Fig. 3.1). The promoter structures are also very much alike. This may be somewhat surprising, since the fitness function is not symmetric in both genes: it prescribed the dynamics of c_1 only. However, implicitly the evolutionary pressures on c_1 and c_2 are very similar. When c_2 is high, it has to repress *tf1* as strongly as possible, which implicitly means that c_2 should increase rapidly to its maximal value; when c_2 initially is low, it should not repress *tf1*, which implies that it should decay as fast as possible. Therefore, the requirements on the response of *tf2* are similar to the requirements on *tf1*. Interestingly, when we explicitly imposed these requirements on c_2 by adopting a fitness function that prescribed the dynamics of *both* genes, a functional switch was never found (within a simulation time of 20.000 generations). Instead, the system was trapped in a local fitness maximum where $c_1 \approx c_2 \approx 0.5$. This illustrates the fact that, as we demonstrate below, the evolutionary path taken in case of the asymmetric fitness function is not symmetric in both genes, even though the final state is.

Time traces

We explained that, in order to compute the fitness of a given network, we compute time traces of the system for two initial conditions. Fig. 3.2 shows three examples of such time traces. The examples are taken from three different stages of the evolutionary process: from the initial generation, from the final generation, and from an intermediate one (generation 400). In generation 0, the DNA and TF sequences are still random, and therefore none of the genes has a functional promoter[1] ($p_{\text{on}}^{(i)}(c_1, c_2) \approx 0$). Therefore the concentrations decay exponentially (Fig. 3.2(a) and (b)). The fitness value of such a network is very low: $F = 0.39$. After 400 generations, the system is still mono-stable (Fig. 3.2(c) and (d)), but the equilibrium point is reached very slowly — the explanation follows below. This leads to a much higher fitness value: $F = 7.33$. In the final result, depicted in Fig. 3.2(e) and (f), the switch is fully functional: concentration c_1 indeed converges quickly to c_{max} or c_{min} depending on the initial condition. This network is truly bi-stable ($F = 7.62$).

Phase portraits

The dynamics of these networks can better be assessed from the phase portraits of the corresponding differential equations. In Fig. 3.3, the time traces from Fig. 3.2 are depicted on top of the vector field $\vec{v}(c_1, c_2)$ defined by the differential equations:

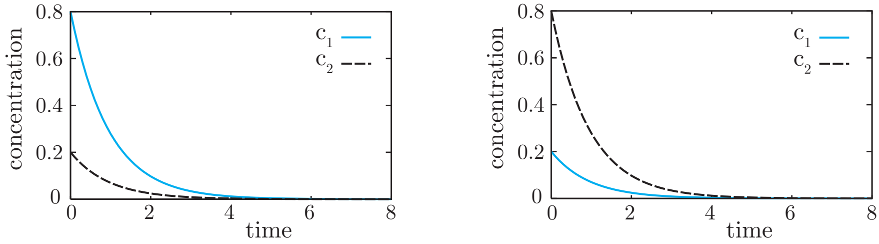
$$\vec{v}(c_1, c_2) = \begin{pmatrix} p_{\text{on}}^{(1)}(c_1, c_2) - c_1 \\ p_{\text{on}}^{(2)}(c_1, c_2) - c_2 \end{pmatrix}. \tag{3.12}$$

The phase portrait in Fig. 3.2(a), belonging to a network from generation 0, shows that all concentrations directly decay to zero irrespective of the initial conditions. This is markedly different from the phase portrait of the final solution, Fig.3.2(c). This figure clearly features two separate basins of attraction, separated roughly by the diagonal $c_1 = c_2$. The two corresponding stable equilibria are $(c_1, c_2) = (1.0, 0.0)$ and $(c_1, c_2) = (0.0, 0.92)$.

Fig. 3.2(b), the phase portrait after 400 generations, is less straightforward. It confirms that the network is still mono-stable at this stage, the equilibrium point being $(c_1, c_2) = (0.6, 0.0)$. But at low values of c_2, the concentrations change very slowly, as indicated by the short vectors in that region; as a result, it takes a long time for the system to converge to the equilibrium concentrations. This way, the network obtains a relatively high fitness value even though it is not bi-stable. The slow convergence is due to the fact that TF1 auto-activates its own production: TF1 binds cooperatively to three binding sites on the promoter of *tf1*. We can understand the consequences of the auto-activation as follows. As the basal transcription rate $p_{\text{on}}^{(1)}(0, 0)$ is non-zero, it must be that at sufficiently

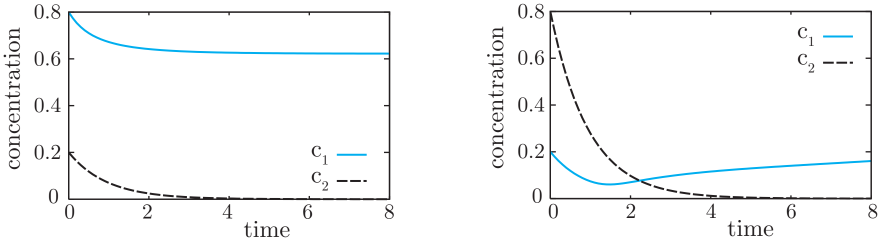
---

[1] As in this case none of the promoters is functional, the initial conditions (with reasonably high concentrations) are not realistic. However, the time traces are computed only to test if the networks are bi-stable and therefore are not required to correspond to biologically relevant scenarios.



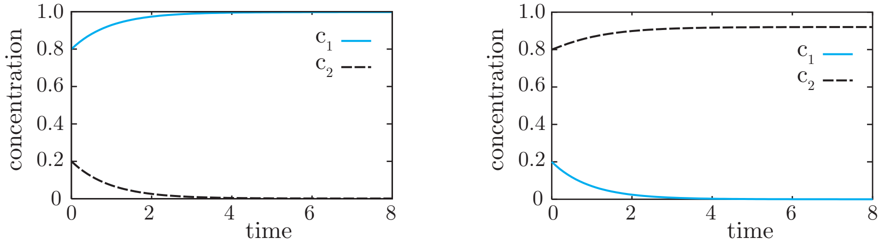
(a) Initial system. Starting condition: $c_1 = 0.8$, $c_2 = 0.2$.

(b) Initial system. Starting condition: $c_1 = 0.2$, $c_2 = 0.8$.



(c) Intermediate result after 400 generations. Starting condition: $c_1 = 0.8$, $c_2 = 0.2$.

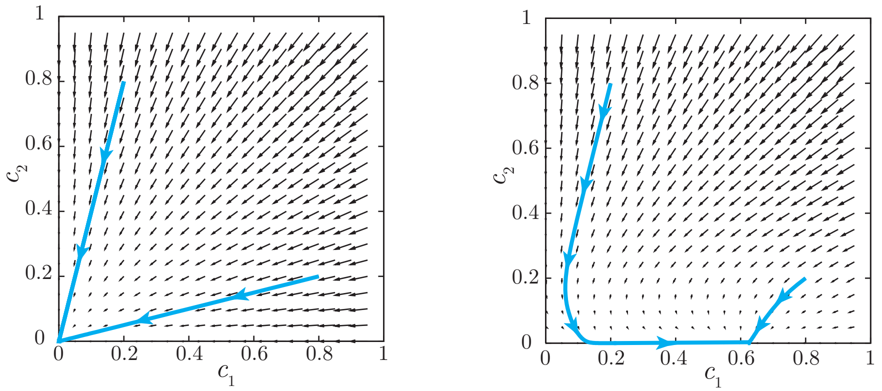
(d) Intermediate result, after 400 generations. Starting condition: $c_1 = 0.2$, $c_2 = 0.8$.



(e) Final result, after 3000 generations. Starting condition: $c_1 = 0.8$, $c_2 = 0.2$.

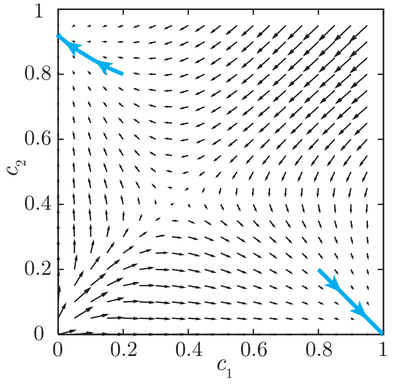
(f) Final result, after 3000 generations. Starting condition: $c_1 = 0.2$, $c_2 = 0.8$.

Figure 3.2: Time traces of the networks resulting from the simulation. In every generation of the algorithm, we test the quality of each network in the population by computing the dynamics of the concentrations as a function of time. We use two sets of initial conditions: $(c_1, c_2) = (0.8, 0.2)$ (Figs. (a), (c), (e)) and $(c_1, c_2) = (0.2, 0, 8)$ (Figs. (b), (d), (f)). Figs. (a) and (b) show time traces of a network from generation 0, which has random DNA and TFs. The genes do not have a functional promoter yet, and therefore all concentrations decay exponentially. Figs. (c) and (d) display the time traces of a network after 400 generations. Gene 2 still does not have an active promoter, but gene 1 has developed a (weak) one. The non-monotonous behavior of c_1 in Fig. (d) reflects that gene 1 represses gene 2. After the decay of concentration c_2, c_1 increases slowly, which is due to auto-activation of TF1. Finally, in Figs. (e) and (f) the switch is complete. The system now clearly is bi-stable.



(a) Phase portrait of initial system (after 0 generations).

(b) Phase portrait after 400 generations.



(c) Phase portrait after 2900 generations.

Figure 3.3: Phase portraits of the systems of differential equations corresponding to (intermediate) results of the simulations. The thick blue lines correspond to the time traces also shown in Fig. 3.2. Fig. (a) shows the typical behavior of a networks from generation 0 (before evolution). Since none of the genes has a functional promoter yet, there is only one equilibrium point: $(c_1, c_2) = (0, 0)$. Fig. (b) corresponds to an intermediate result, after 400 generations. The system still has only one equilibrium, $(c_1, c_2) \approx (0.6, 0)$, but this point is approached only very slowly from initial condition $(c_1, c_2) \approx (0.2, 0.8)$. Fig. (c) depicts the end result, which is clearly bistable; it has three equilibrium points, one of which is unstable.

low concentrations of TF1 and TF2, $\mathrm{d}c_1/\,\mathrm{d}t = p_{\mathrm{on}}^{(1)}(c_1, c_2) - c_1 > 0$. This means that if c_1 is initially in this regime, it will rise until it is equal to $p_{\mathrm{on}}^{(1)}$. However, due to the auto-activation, $p_{\mathrm{on}}^{(1)}(c_1, c_2)$ is an increasing function of c_1. For certain affinities of the TF1 binding sites, $\mathrm{d}c_1/\,\mathrm{d}t = p_{\mathrm{on}}^{(1)}(c_1, c_2) - c_1$ stays positive yet small for a considerable range of values of c_1. As a result, c_1 will increase very slowly when it is in this range.

In generation 400, the role of TF2 is still limited. However, TF2 does already repress TF1, as in the end result. Consequently, c_1 decreases while c_2 is large, but increases again after c_2 has decayed. This behavior contributes to the fitness because the repression by c_2 is most influential when the system starts from the second initial condition $(c_1, c_2) = (0.2, 0.8)$, in which case c_1 should indeed be low.

Evolutionary path

The state of the system after 400 generations illustrates the evolutionary path taken in our simulations. This path can schematically be summarized as follows:

1. *tf1* develops a functional promoter;
2. TF2 starts repressing TF1 if c_2 is high;
3. TF1 develops auto-activation to slow down the dynamics; consequently, if c_2 is initially high, it stays high for a while, whereas if it is initially low, it temporarily stays low;
4. The system becomes bi-stable due to the auto-activation of TF1;
5. TF2 develops a (low) basal expression to repress TF1 constitutively.
6. TF1 starts to repress TF2 if c_1 is high;
7. TF2 develops auto-activation to increase the speed of convergence to the steady state;

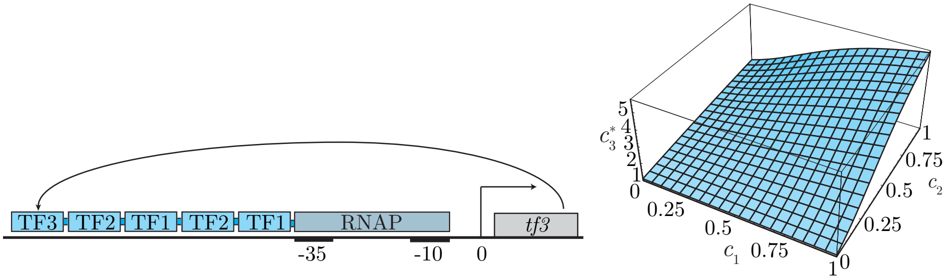
These steps occur roughly in the given order, even though some steps may be reversed or occur in overlapping time intervals, depending on the initial conditions.

Presumably, the selection pressure applied in our algorithm is not comparable to the selection pressure on real switches such as the phage λ; for that reason, we do not claim that the path taken in the simulations is similar to the evolutionary path that produced the phage λ switch.

3.3.2 Logic gates

We now turn to the logic gates. Refer to Table 2.1 for the definitions and names of the different logic gates.

For each gate, we performed simulations with several initial conditions and random seeds. Many of the logic gates resulting from our simulations indeed



(a) Promoter design (b) Response plot

Figure 3.4: AND gate using auto-activation conditional on the presence of other TFs. The regulated gene *tf3* codes for a transcription factor TF3 that binds to its own *cis*-regulatory region. The resulting auto-activation, however, depends on the presence of TF1 and TF2. Fig. (b) shows the response plot of the *cis*-regulatory region depicted in Fig. (a).

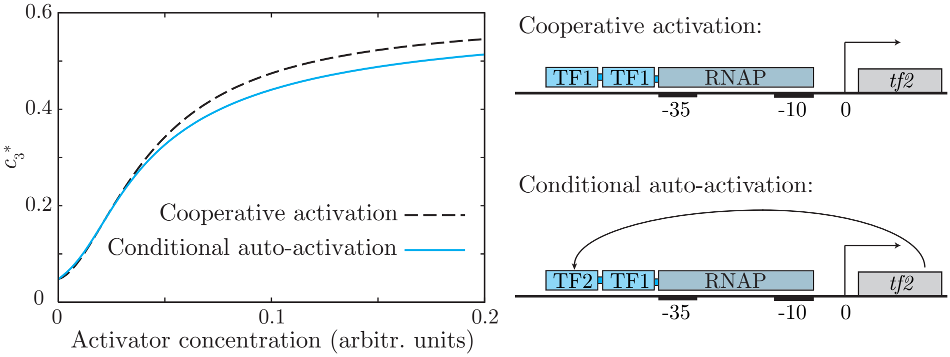
exploit auto-regulation. Several gates, such as the AND and and ANDN gates use the designs of Chapter 2 in some runs, and new designs using feedback in other ones. Other gates, notably the NOR and the NAND, always use auto-regulation in our simulations.

We identified two feedback mechanisms that are used in several gates; we describe them below.

Conditional auto-activation

The first mechanism we termed *conditional auto-activation*. This mechanism occurs regularly in the gates in which cooperative activation plays a key role, such as the AND and the ANDN gates. As an example, Fig. 3.4 shows the design of an AND gate. Overall, this AND gate is very similar to the one described in Chapter 2: a hetero-cooperative activation module is responsible for activation in the presence of both TFs. However, the hetero-cooperative module now also contains a binding site for TF3, which leads to a positive feedback loop. Importantly, TF3 bound at its binding site cannot directly recruit RNAp to the promoter; instead, it binds cooperatively with the hetero-cooperative activation module. As a result, the auto-activation is conditional on the presence of TF1 and TF2. As the concentrations of TF1 and/or TF2 increase, the auto-activation gains in strength, leading to a sharp response.

The basic mechanism can be studied in a much simpler system. In Fig. 3.5, we compare two activation mechanisms for a transcription factor TF1 regulating a gene *tf2* that codes for another transcription factor, TF2. In the first scenario, only a homo-cooperative activation module is present, consisting of two binding sites for TF1. In the second secenario, we replace the auxiliary TF1 site by a binding site for TF2, thus introducing conditional auto-activation (see Fig. 3.5(b)). We optimize the binding sites in both designs in order to maximize the steepness of response, and compare the response plots (Fig. 3.5(a)). The design using



(a) Response plots (b) Promoter designs

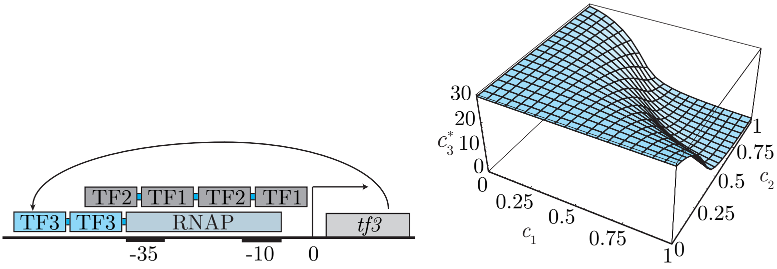
Figure 3.5: Conditional auto-activation vs. cooperative activation. The plot (a) compares the response plots corresponding to the two activation systems depicted schematically in Fig. (b). In both alternatives, TF1 activates the expression of a gene *tf2* coding for another transcription factor, TF2. In the first scenario, TF1 binds cooperatively to a pair of activator sites. In the second scenario, one TF1 site is replaced by an operator for TF2, which introduces a positive feedback loop that depends on the presence TF1. The binding affinities of all sites are chosen such as to optimize the sharpness of the response. The plots demonstrate that the responses of the two designs are comparable.

conditional auto-activation indeed produces a response that is comparable in steepness to the conventional design with two binding sites. It is, however, not better than the conventional one. This explains why cooperative activation and conditional auto-activation show up as alternatives in our simulations.

Auto-activation can sharpen repression

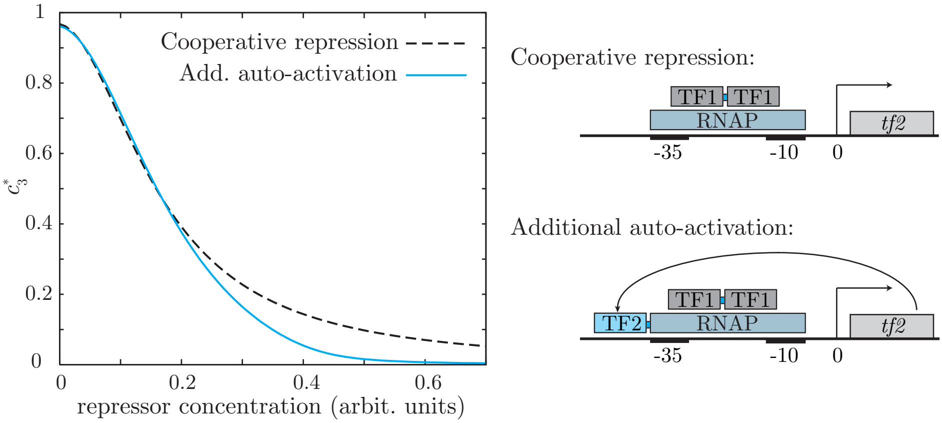
A second feedback pattern emerges in gates in which repression is important, such as the NAND, NOR and EQU gates. As it turns out, whenever steep repression is required, we also find auto-activation. Fig. 3.6 shows a representative design for the NAND gate that uses auto-activation; the resulting response plot shows an excellent NAND gate — in fact, it performs much better than the design in Chapter 2.

To explain the mechanism that is responsible for the steepness of the switch, we again analyze a less complex example. In Fig. 3.7, we compare two scenarios. In the first scenario, a transcription factor TF1 cooperatively binds to a pair of repressor sites to inhibit the gene *tf2*, which codes for another transcription factor TF2. In the second scenario we use the same configuration, but add an activator site for TF2. Thus, auto-activation competes with cooperative repression. In order to compare the two designs, we keep the basal expression level of the two scenarios fixed, and also impose the concentration c_1^{half} at which the transcription rate is half-maximal. Given these constraints, we optimize the steepness of both



(a) Promoter design (b) Response plot

Figure 3.6: NAND gate using auto-activation to sharpen the repression curve. The regulated gene *tf3* is repressed by a hetero-cooperative module consisting of binding sites for TF1 and TF2 (Fig. (a)). However, *tf3* codes for a transcription factor TF3 that binds to its own *cis*-regulatory region. At low concentrations of TF1 and TF2, the resulting auto-activation competes with the repression module; as a consequence, the response to the concentrations of TF1 and TF2 is very sharp (Fig. (b))



(a) Response plots (b) Promoter designs

Figure 3.7: Repression combined with auto-activation. Fig. (a) shows the response plots of two slightly different repression systems, shown schematically in Fig. (b). In both cases, a transcription factor TF1 represses a gene *tf2* coding for a second transcription factor by binding cooperatively to two repressor sites. In the second scenario, an activator site for TF2 is present as well, which induces auto-activation. The second scenario constitutes a more effective repression system than the first. The presence of auto-activation allows for a much weaker core promoter ($q_p \approx 1$); as a result, the displacement of RNAp from the promoter by the repressor is more effective as the concentration of TF1 increases.

designs[2]. As can clearly be seen in Fig. 3.7(a), the second scenario results in a much more effective repression. Qualitatively, this can be understood as follows. In the absence of TF1, the gene TF2 auto-activates. As a result, the bare promoter strength (RNAp affinity) does not need to be very high in order to obtain a considerable expression level. At high concentrations of c_2, the low RNAp affinity makes it easier to displace RNAp from the promoter, leading to a more complete repression.

3.4 Discussion, conclusions and outlook

We showed that our model can be used to design transcription networks that can exploit feedback. We used it to design a genetic switch. Imposing only that the system should have two stable states, a switch was obtained that is remarkably similar to the phage λ switch. The design used mutual repression and auto-activation to create two stable states.

We also used the model to design logic gates that are free to use feedback. The resulting designs shed new light on the use of auto-regulation. The result show that auto-regulation is likely to emerge in quite common functions such as AND gates. We described two mechanisms that are used in multiple gates. First, if auto-activation is conditional on the presence of other TFs, it can give rise to sharp responses that are not due to multiple binding sites of the TF. This mechanism can be an alternative to cooperative activation, for instance if the input TF is not capable of homo-cooperative binding. Second, auto-activation can strongly contribute to the sharpness and effectiveness of repression systems. Whenever sharp repression is required, auto-activation can be useful. These mechanisms may help explain the large number of auto-activators present in *E. coli*.

We showed that auto-activation can be particularly useful in repression systems. Therefore, it would be interesting to test if TFs that auto-activate are often regulated by repressors.

In none of the simulations we found auto-*repression*. On the one hand, this is rather surprising given the large number of known auto-repressors in *E. coli*. On the other hand, we mentioned in the introduction that the known functions of auto-repression are that it increases stability and can create rapid responses. In our method, such qualities are not being rewarded: as we use deterministic differential equations to model the dynamics of the gates, we ignored fluctuations, and since our selection scheme is based on the steady states of the system, the response speed was irrelevant too. In reality, these qualities may of course be important, which might explain the abundance of auto-repressors in *E. coli*. In future work, it would be interesting to use stochastic models or to select for fast responses; perhaps auto-repression shows up under those conditions. Indeed,

---

[2] In fact, the first scenario is fixed by these constraints, whereas in the second scenario one degree of freedom remains undetermined, allowing for optimization.

preliminary results suggest it does.

As we ignored response speed and stability in our approach, the resulting designs could be bad performers on these criteria. We mentioned that auto-activation is indeed known to reduce response speed in some situations and to increase the amplitude of fluctuations. Clearly, this may be a problem in some cases. On the other hand, a slow response can be turned into a positive feature as well, as it filters high-frequency noise, and fluctuations may in some cases be beneficial, for instance for stochastic switching (Kussell and Leibler, 2005). Indeed, the fact that auto-activation is found so often in *E. coli* demonstrates that the associated reduction of the response speed and the amplification of fluctuations can apparently be circumvented, tolerated or used.

Chapter 4

Chance and necessity in chromosomal gene distributions

By analyzing the spacing of genes on chromosomes, we find that transcriptional and RNA-processing regulatory sequences outside coding regions leave footprints on the distribution of intergenic distances. Using analogies between genes on chromosomes and one-dimensional gases, we constructed statistical null models to describe the data. We have used these models to estimate typical upstream and downstream regulatory sequence sizes in various species. Deviations from the models reveal bi-directional transcriptional regulatory regions in *S. cerevisiae* and bi-directional terminators in *E. coli*.

4.1 Probability distributions of intergenic distances

More and more genomes get sequenced and the locations of their open reading frames (ORFs) can be determined to a high accuracy. As a result, we now know the chromosomal locations of most genes of a variety of species. This information has been studied and used extensively for various aims. In this chapter, we step back and wonder: Can we not only characterize, but also *understand* the way genes are distributed over chromosomes?

We focus on the distributions of distances between genes on chromosomes. Presumably, these distributions are shaped in part by random processes. By various mechanisms, base pairs are regularly inserted or deleted from the genome, and sequences are inverted, shuffled and duplicated. Such processes tend to randomize the distribution of ORFs over a given chromosome. On the other hand, natural selection could lead to a bias in the actual distribution if there is a functional reason for genes to be spaced in a particular way (Warren and ten Wolde, 2004b). We use stochastic models to study which features of the distribution of intergenic distances can be explained by random processes only, and which features require an explanation in terms of functionality.

It turns out that models about the distribution of genes on a chromosome can typically be mapped to existing models of 1D gasses: particles in a one-dimensional space. This allows one to use standard methods of statistical mechanics to compute the quantities of interest.

Interestingly, we thus identify universal features that can be explained by statistical considerations only. But the *deviations* from the random models are at least as interesting. We study the distributions of genes in *Saccharomyces cerevisiae* and *Escherichia coli* in detail, and show that the deviations from our random models indeed lead to interesting predictions about transcription initiation and termination in these species.

4.2 Models

In this section we introduce three models: the “Ideal-Gas” model, the “Tonks-Gas” model, and the “Constant Force” model. Each model follows from the previous by imposing one extra constraint to the system.

4.2.1 The Ideal-Gas model

The “Ideal-Gas” (IG) model is the simplest and most naive model. It assumes that the ORFs on a chromosome are distributed at random. This assumption leads to the following predicted properties of the distance distributions.

In the first place, the probability distribution of the distances between all ORFs, measured from their closest ends, can be computed analytically and to

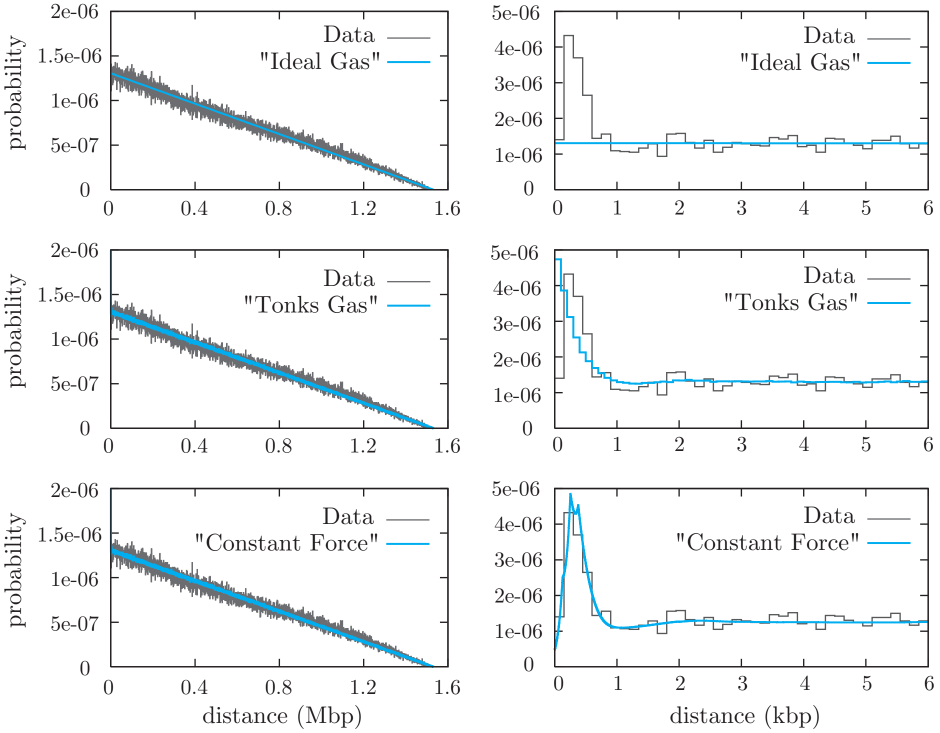


Figure 4.1: Distance distributions of *S. cerevisiae* chromosome 4. The two figures at the top compare the distance distributions to the predictions of the “Ideal-Gas” (IG) model, while the lower two show the predictions of the “Tonks-Gas” (TG) model. The left figures show that, at long distances, the IG model and the TG model are equivalent and both correctly describe the data. The figures at the right zoom in on the short-distance data. It is evident that the peak in the data at distances shorter than 1 kbp is not explained by the IG model. The TG model does show a peak (correlations) at short distances, but this peak looks qualitatively different from the data. In particular, at $n < 300$ bp, the real data show a “dip” in the distribution that is not present in the TG model.

good approximation is a straight line:

$$P_{\mathrm{IG,\ all}}(n) = \frac{2}{L}\left(1 - \frac{n}{L}\right). \tag{4.1}$$

Here L is the length of the chromosome, and n is the distance measured in base pairs. Figures 4.1(a) and (b) compare this model to the distribution of *S. cerevisiae*. As one can verify, this model applies at long distances , but at shorter distances this description breaks down: there is a clear peak in the actual distribution at short distances ($n < 1000\,\mathrm{bp}$) which is not explained by the Ideal-Gas model.

The second prediction is that the distribution of distances between *neighboring* genes (*i.e.* the lengths of intergenic regions) should be geometric (exponential) as long as $n \ll L$. The exponent is proportional to the density of genes:

$$P_{\mathrm{IG}}(n) = \rho\,\mathrm{e}^{-\rho n}, \qquad (n \ll L). \tag{4.2}$$

where the number density ρ is defined as $\rho \equiv N/L$, N being the number of genes. Clearly, an geometric distribution is a good description for distances $n > 300\,\mathrm{bp}$ in *S. cerevisiae* (see Fig. 4.1(a)), and even for shorter distances in *E. coli* (Fig. 4.1(b)), but at very short distances this behavior breaks down in both cases. Besides, the value of ρ predicted by the IG model is too low to fit the actual exponential decay.

4.2.2 The Tonks-Gas model

To account for the peak in the distance distributions at short distances, we introduce an extra component to the model. A clear error in the IG model is that it allows the genes to overlap much too often. The *S. cerevisiae* genome, for example, is very dense: about 72% of the genome is occupied by ORFs. If the genes would be distributed according to the IG model, then 78% of the genes should overlap with at least one other gene[1], whereas in reality, only about 9% do.

For simplicity, we now assume that genes do not overlap at all, but otherwise are distributed at random. This model is formally equivalent to a one-dimensional gas of polydisperse hard particles, also known as a poly-disperse Tonks Gas (Tonks, 1936). The consequence of the assumption that overlap never occurs, becomes visible if we compare the Tonks-Gas (TG) distribution with the IG distribution. Indeed, the Tonks-Gas distribution has a peak at very short distances (see Fig. 4.1(d)), but is indistinguishable from the IG model at longer distances

---

[1] If the average size of a gene is 1350 bp, then on a chromosome with length 1.5 Mbp and containing 830 genes (packing fraction 72%) the probability for a certain gene to not overlap with any other gene equals $(1 - 2 \times 1350/1\,500\,000)^{829} \approx 0.22$.

(Fig. 4.1(c)). Also, the nearest neighbor distribution changes to

$$P_{\mathrm{TG}}(n) = \left(\frac{\rho}{1-\theta}\right) \mathrm{e}^{-n\rho/(1-\theta)}, \tag{4.3}$$

where $\theta \equiv N\mu/L$ is the packing fraction and μ is the average gene length. Note that the average length of the intergenic regions decreases with a factor $(1-\theta)$ compared to the IG model.

The TG model clearly shows that "excluded-volume" interactions between the genes can explain, on the one hand, the peak in the distance distribution at short distances, and on the other hand the steepness of the exponential curve fitting the distribution of intergenic regions. However, there still is a clear discrepancy between the data and the Tonks-Gas model, especially in the *S. cerevisiae* case. This discrepancy is that for very short distances (roughly, $n < 200\,\mathrm{bp}$), the actual data show a clear *dip* (see Fig. 4.1 and 4.3(a)). Apparently, genes are not in each others close proximity as often as expected.

4.2.3 The Constant-Force model

The observation that ORFs are rarely very close together (see Fig. 4.1 for *Saccharomyces cerevisiae*) inspires the definition of the Constant-Force (CF) model, illustrated in Fig. 4.2. We hypothesize that the underrepresentation of closely spaced ORFs is caused by functional sequences directly upstream and downstream of the ORFs, which we call upstream and downstream control regions (UCRs and DCRs). UCRs include basal promoters, *cis*-regulatory regions and 5' untranslated regions (UTRs); DCRs consist of 3' UTRs, transcriptional terminators and RNA-processing signals. If ORFs approach each other closely, these regions need either to overlap or to be very short, which makes such configurations less likely. To test this hypothesis, we divide the intergenic regions into three subsets, called tandem (T), convergent (C) and divergent (D). (See Fig. 4.2.) Intergenic regions in subset T should contain one DCR and one UCR, whereas C and D intergenic regions contain two DCRs and two UCRs respectively. As UCR sequences are generally longer than DCRs, we expect that D regions are on average longer than T regions and C regions are shortest, which is indeed the case (see Figs. 4.3(a) and 4.4).

These observations inspire the following extension to the Tonks-Gas model. We assume that all ORF configurations are equally probable, except for the following constraints: (i) ORFs do not overlap; (ii) UCRs and DCRs can overlap with each other or with ORFs, but every overlapping base pair (bp) in a particular configuration makes this configuration a factor q less probable. For simplicity, we assume that in a given organism all UCRs and DCRs have a fixed length, π and τ respectively.

This model is equivalent to a one-dimensional system of hard particles with a finite-ranged, repulsive, *constant-force* interaction. Tandem, convergent and diver-

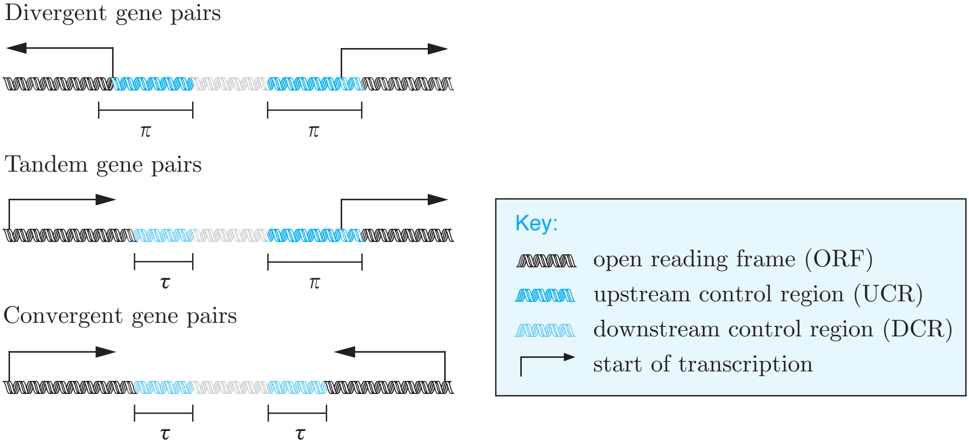


Figure 4.2: The Constant-Force model. This model assumes that a 5' UTR, a basal promoter and a *cis*-regulatory region are present upstream of every ORF. We call this the upstream control region (UCR), and assume it has a fixed size π. Downstream of each ORF, the 3' UTR and possibly a transcriptional terminator and RNA processing signals are present, to which we jointly refer as the downstream control region (DCR), assumed to have length τ. The figure shows that ORFs neighboring on the DNA can have three mutual orientations: divergent (D), tandem (T) or convergent (C). This also leads to three kinds of intergenic regions: D regions contain two UCRs, while T regions contain one UCR and one DCR, and C regions have two DCRs. In *S. cerevisiae*, the frequencies of D, T and C regions are 26.3%, 48.3% and 25.4% respectively, which is close to the random proportions 1:2:1. This holds for most fungi. In *E. coli*, T regions are more frequent due to the organization of its genes in operons (17.5%, 66.7% and 15,7%).

gent ORF pairs interact at a range $\pi + \tau$, 2τ and 2π, respectively. This mapping enables us to use the formalism of statistical physics to compute the probability distributions corresponding to this model analytically (see Appendix 4.A).

4.3 Results

We have used the CF model to study the distributions of intergenic distances in more detail. Below, we describe the results for *E. coli* and several other fungal species.

4.3.1 The CF model fits the C and T distribution of *S. cerevisiae*

The CF model fits the distributions of *S. cerevisiae* convergent and tandem intergenic distances remarkably well (see Fig. 4.3(c); we describe the fit procedure in 4.C). The fit parameters are $\tau = 61$ bp for DCRs and $\pi = 196$ bp for UCRs. These numbers provide a course estimate of the space required for the transcriptional and translational regulatory signals and RNA processing in *S. cerevisiae*.

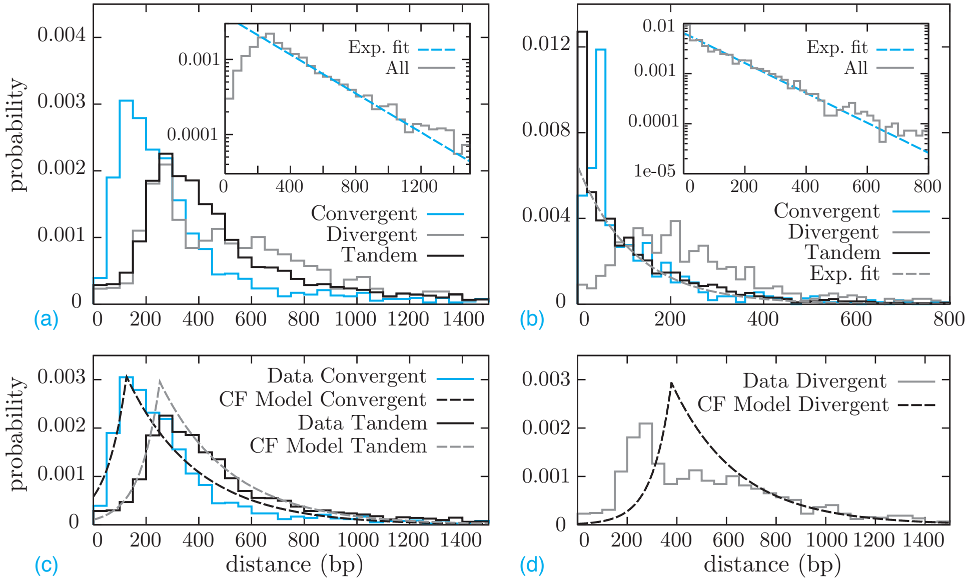


Figure 4.3: Probability distributions of intergenic distances in *S. cerevisiae* and *E. coli*. (a) Probability distributions of intergenic regions in *S. cerevisiae*. The distributions of distances between convergent (C), divergent (D) and tandem (T) gene pairs. C intergenic regions are, on average, shorter than T regions; the D regions are longest. Note also that the divergent distribution has a bimodal shape, with a peak at $n \approx 275$ bp and one at $n \approx 500$ bp. Inset: the distribution of all intergenic regions is exponential for distances larger than 300 bp (scale parameter: 335 bp), but has a "dip" at shorter distances. This dip, we argue, is a footprint of UCRs and DCRs. (b) As panel (a), but for *E. coli*. The T distribution in the main plot is exponential, except for an accumulation in the first bin. This accumulation is the result of intergenic regions inside operons, which are not separated by control regions and therefore can be arbitrarily close together (Lesnik *et al.*, 2001). The C distribution is also exponential, except for a peak at 20–60 bp, where *S. cerevisiae* has a "dip" instead. We predict that this peak is the result of bi-directional terminator sequences. The inset again shows that the distribution of all intergenic regions is largely exponential (scale parameter: 145 bp). (c) Simultaneous fit of the Constant Force (CF) model to the C and T distributions of *S. cerevisiae*. The model fits the data surprisingly well. (d) The D distribution of *S. cerevisiae* and the expected distribution according to the CF model. Clearly, the bi-modal shape of the data is not consistent with the CF model. We predict that the set of divergent intergenic regions in *S. cerevisiae* consists of two subpopulations: those containing two independent *cis*-regulatory regions, responsible for the second peak, and those containing one bi-directional *cis*-regulatory region.

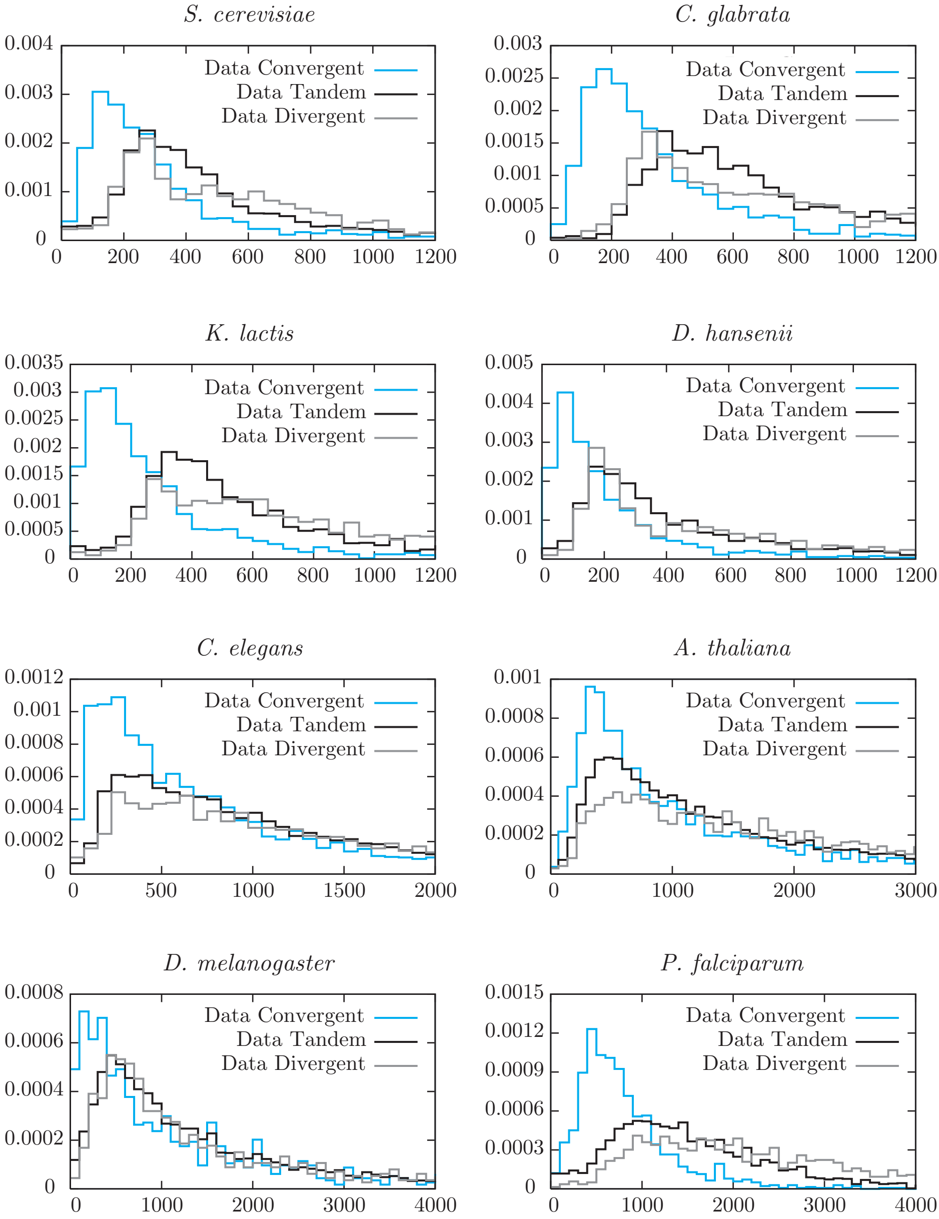


Figure 4.4: Distributions of intergenic distances, broken down into three different subsets (convergent, tandem, divergent pairs), for four fungi and four additional eukaryotes. Consistently, the convergent gene pairs are, on average, closer together than the tandem ones. The divergent genes are furthest apart. Note also that the divergent distribution is bimodal for all fungi, suggesting the presence of bi-directional promoters in all of them.

Name of organism	UCR length/bp	DCR length/bp	q
S. cerevisiae	196 ± 4	61 ± 1	0.985 ± 0.001
C. glabrata	296 ± 4	66 ± 2	0.983 ± 0.001
K. lactis	295 ± 5	38 ± 2	0.987 ± 0.001
D. hansenii	141 ± 4	28 ± 2	0.976 ± 0.001

Table 4.1: Estimates for the UCR and DCR sizes for various fungal species. The errors given in the table are the uncertainties of the fit parameters and as such should not be interpreted as variances of these quantities in the genome.

Our UCR length prediction of 196 bp is in excellent agreement with the distribution of transcription-factor-binding sites near *S. cerevisiae* start codons, which has its peak at 100–200 bp from the start codon (Harbison *et al.*, 2004). Our DCR prediction of 61 bp is supported by bioinformatics analyses of *S. cerevisiae* 3' RNA processing signals, which show that the majority of these sequences is within 20–90 bp of the stop codon (van Helden *et al.*, 2000). However, Graber *et al.* predict longer 3' UTRs (Graber *et al.*, 2002) and recent experiments show that the median of 3' UTRs lengths is ≈ 91 bp (David *et al.*, 2006), which suggests that our DCR estimate is on the low side. In Appendix 4.B we show that more refined models can provide quantitative agreement.

4.3.2 Typical UCR and DCR sizes for other fungi

We repeated this approach to estimate the UCR and DCR lengths for three additional fungi using only the ORF coordinates as input. (See Table 4.1 and Fig. 4.7). We found that UCRs are consistently longer than DCRs. The UCR and DCR lengths seem to vary independently of each other, and no dependence on gene density is apparent. Recently, it has been shown that the distribution of "rigid DNA" in *cis*-regulatory regions of fungi correlates with the position of transcription-factor-binding sites (Tirosh *et al.*, 2007). Our estimates for the UCR lengths correlate well with the position of rigid DNA in these fungi.

4.3.3 *S. cerevisiae* contains many bi-directional UCRs

We now turn to the spacing of divergent pairs in budding yeast. Interestingly, the corresponding distribution has a bimodal shape (see Fig. 4.3(d)) that is even more pronounced in other fungi (Fig. 4.6). The first, narrow peak is centered on 275 bp; the second peak is broader and is maximal around 550 bp. This shape is not consistent with the CF model.

Apparently, many divergent intergenic regions are very short: 29% are $<$ 300 bp. Because few (10%, 280 out of 2801) tandem intergenic regions, containing only one UCR, are $<$ 200 bp, it seems unlikely that two independent UCRs could fit in divergent intergenic regions with a length of the order of 275 bp. Hence we propose that the set of divergent gene pairs is composed of two sub-populations.

The first population, corresponding to the second peak, consists of pairs of genes that are regulated independently. The other sub-population consists of gene pairs that share a bi-directional *cis*-regulatory region, that is, a regulatory region containing elements such as transcription factor binding sites that regulate the expression of both flanking genes. Such a coupling could force genes to preserve their proximity, thus causing the deviation from the CF model. While bi-directional *cis*-regulatory regions are ubiquitous in *E. coli* (Warren and ten Wolde, 2004b), only a few bi-directional UCRs have been reported in *S. cerevisiae* (Bell *et al.*, 1995; Liu and Xiao, 1997; Aranda *et al.*, 2006; Ishida *et al.*, 2006). Based on Fig. 4.3(d), we predict that about 30% (426 out of 1471) of the divergent pairs are regulated by a shared *cis*-regulatory region. (A file containing the best candidates is available online[2].)

If this is true, then one would expect co-expressed divergent pairs to be overrepresented in the first peak rather than the second. We tested this using a large set of expression data; indeed, this is the case for positively correlated pairs ($p < 0.002$; see 4.D.1 for the details of the analysis). Negatively correlated pairs are typically not in the first peak. This contrasts with bi-directional UCRs in bacteria, in which dual regulators often act as a repressor for one of the genes and as an activator for the other, resulting in anti-correlated expression patterns.

We also used Gene Ontology (GO) annotations (Ashburner *et al.*, 2000) to test whether the divergent neighbors in the first peak are more often functionally related than those in the second peak. Adopting the information-theoretic measure of Resnik (1998) to quantify the similarity between GO terms, we indeed found this to be the case ($p < 9 \times 10^{-4}$ for biological process, $p < 5 \times 10^{-5}$ for cellular component; see 4.D.2).

See Box 4.1 on the BIO3–5 cluster for an illustrative example of a possibly bi-directional promoter in *S. cerevisiae*.

4.3.4 *E. coli* has many bi-directional terminators

As mentioned above, bi-directional *promoters* are well-characterized in *E. coli*. We now show that the distribution of convergent gene pairs in *E. coli* provides evidence for bi-directional transcriptional *terminators*, which are much less well described.

In accordance with the CF model, the C distribution has an exponential signature (see Fig. 4.3(b)). Convergent intergenic regions are expected to contain two DCRs. Given the typical size of Rho-independent terminators ($\approx 40\,$bp) the CF model predicts a dip at short distances ($< 80\,$bp). Instead, there is a significant excess of intergenic regions of size 20 to 60 bp ($p = 10^{-13}$; see 4.D.3). It is unlikely that two terminators would fit into such short intergenic regions.

Rho-independent terminator sequences function by stem–loop formation of the RNA transcript, and hence are largely palindromic. As the complementary strand of a palindromic sequence is a palindrome too, some terminators can function bi-directionally. Indeed, a few bi-directional terminators have been identified

---

[2] http://www.sciencedirect.com/science/journal/01689525

Box 4.1: The BIO3, BIO4, BIO5 cluster in Yeast

It is instructive to discuss one interesting example in Yeast: the BIO3, BIO4 and BIO5 cluster (Phalip *et al.*, 1999). All genes in this cluster are involved in the biotin biosynthesis pathway. BIO3 and BIO4 are transcribed in a divergent orientation from a short intergenic region of length 222 bp (which falls into the first peak in the length distribution of divergent intergenic regions) and are tightly co-expressed. The orthologs of BIO3 and BIO4 in *E. coli* are BioA and BioB; these genes are closely spaced divergent neighbors as well (87 bp), and are simultaneously repressed by BirA binding to their shared UCR. Phalip *et al.* already speculate that a similar mechanism is at work in *S. cerevisiae* (Phalip *et al.*, 1999), but the mechanism of co-expression of these genes has not been studied in detail. BIO4 and BIO5 are tandem neighbors, and are only 55 bp apart, Clearly, this cluster has many of the features that we see in our statistical analysis. We therefore suggest that detailed experimental work on the regulation of the BIO cluster might illuminate some important mechanisms that shape the distribution of genes over the *S. cerevisiae* chromosome.

experimentally (Postle and Good, 1985). Moreover, Lesnik *et al.* (2001) used an algorithm called `RNAMotif` to identify putative terminators and predicted that many of them could function bi-directionally. Given that most terminators in *E. coli* start within 60 bp downstream of their ORF (Lesnik *et al.*, 2001), genes sharing a bi-directional terminator should usually be close together; this suggests that the peak in the distribution at short distances is caused by bi-directional terminators.

To explain the data, at least 86 bi-directional terminators should be present; this would imply that as many as 23% of the operons use a bi-directional terminator (see 4.D.3).

We tested this, using the data of Lesnik *et al.* (2001)[3]. Indeed, putative terminators that `RNAMotif` classifies as bi-directional have a tendency to occur in short, convergent intergenic regions, corroborating our hypothesis ($p < 0.0003$, see 4.D.4 and 4.D.5) (Salgado *et al.*, 2000).

4.3.5 The operon structure of *E. coli* is evident from the T distribution

The T distribution of *E. coli*, like the C distribution, nicely follows the exponential distribution predicted by the CF model. However, at short distances, a clear excess of pairs is found. As we pointed out in the introduction section, this excess is due to the operon structure of the *E. coli* genome. This hypothesis can be tested conclusively by computing the distribution of distances between neighboring tandem genes *on the borders of operons* for all known operons. Fig. 4.5 shows that in this distribution the peak at short distances has completely disappeared;

---

[3] Although no statistical test is presented, a similar conclusion is reached by Yachie *et al.* (2006)

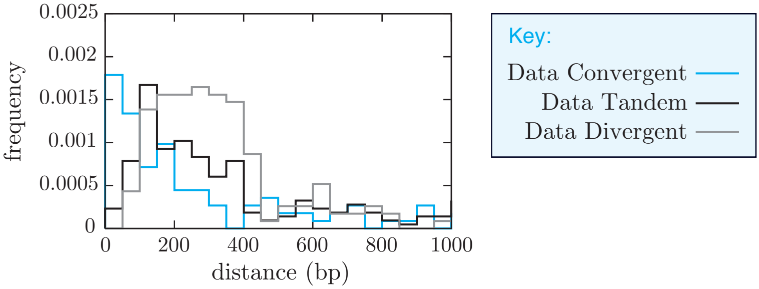


Figure 4.5: Distribution of distances between ORFs at borders of operons in *E. coli*. Note that tandem neighbors are not likely to be very close together ($n < 100$). Tandem neighbors within the same operon do not show this ”repulsion” (Fig. 4.3). This supports the idea that the repulsion between ORFs observed in all species, is indeed a consequence of regulatory regions.

this conclusively shows that this peak is due to tandem genes *within* operons. Fitting the CF model to the T distribution results in $\tau + \pi \approx 125\,\mathrm{bp}$, which is as expected.

The fact that genes within operons are much more closely packed than those at the boundaries of operons, has been recognized some time ago, and has been used for some years to predict operon boundaries (see, for instance, Salgado *et al.* (2000)).

4.4 Concluding remarks

The largest limitation of the current model is the assumption that the UCRs and DCRs have fixed sizes. Especially in higher eukaryotes, UCR and DCR lengths often have a high variance; in these cases, it is necessary to include this in the model. In the Supplement we show that this can be done and how more realistic potentials can be chosen.

That being said, the simple CF model describes many universal characteristics of the gene spacing. It not only quantitatively describes the exponential decay at large distances, but also the ”repulsion” at short distances due to UCRs and DCRs in prokaryotes and eukaryotes alike. In *E. coli* and fungi, this repulsion provides information about the typical length of UCRs and DCRs using only the ORFs coordinates as input. The model can also serve as a null model for the spacing of genes: deviations from it lead to meaningful predictions about the presence of operons, bi-directional promoters or terminators.

4.A The constant-force model

In this section we provide additional information about the Constant-Force (CF) model.

4.A.1 Assumptions

As we explained in the main text, the CF model is based on three assumptions. First, we assume that ORFs cannot overlap. Second, in a given organism, upstream control regions (UCRs) and downstream control regions (DCRs) have a fixed size (π and τ, respectively). Third, we assume that these control regions can overlap with each other and with nearby ORFs, but that such overlaps are not likely. More precisely, we assume that *whenever a base pair from such a region overlaps with another functional region, be it an ORF, a UCR or a DCR, it makes that particular configuration a factor q less probable.* For simplicity, we make no distinction between the different kinds of overlap.

A useful analogy can be drawn with a physical system. The proposed model is formally equivalent to a one-dimensional system of hard particles with finite-ranged repulsive interactions. The interaction determined by our assumptions is an interaction with a *constant force*. The range of the interaction depends on the mutual orientation of the neighboring genes. Divergent gene pairs are separated by two UCRs and therefore start interacting at a distance 2π. Convergent pairs have two DCRs in their intergenic region and therefore have an interaction range of 2τ. Lastly, intergenic regions between tandem pairs contain one DCR and one UCR, leading to an interaction range $\pi + \tau$. This analogy allows us to use the formalism of statistical physics to compute the probability distribution of the intergenic distances for this model analytically; the complete derivation follows below.

4.A.2 Derivation of the distance distributions: CF interaction with fixed range

In the CF model described above, the interaction range of the particles depends on their mutual orientation (convergent, divergent or tandem). We first derive the distance distribution for a slightly simpler system, in which the interaction range does not depend on the orientation.

We consider a one-dimensional space (representing the chromosome) of length L' containing $N-1$ particles (representing ORFs). We choose to describe the system in the micro-canonical ensemble, with fixed total energy E. The state of the system can be described by a vector $\vec{n} = (n_1, n_2, \ldots, n_N)$, where n_i is the length of the ith inter-particle space. The sum of these numbers, $L \equiv \sum_i n_i$, is the total free space in the system. The value of L is fixed and $L \gg 1$. As the particles occupy part of the total space, $L < L'$.

For now we assume that the particles interact with a finite-ranged CF potential

$U(n)$, defined as:

$$\frac{U(n)}{k_{\mathrm{B}}T} = \begin{cases} \epsilon(r-n) & \text{if } n < r, \\ 0 & \text{if } n \geq r, \end{cases} \tag{A4.1}$$

where r is the range of the interaction, and ϵ is the energy associated with an overlap of one base pair (in units of $k_{\mathrm{B}}T$); it is related to q as $\epsilon = -\ln(q)$.

In order to compute the probability distribution of intergenic distances, we divide the system into two subsystems. Subsystem 1 (S1) is a particular, but arbitrary, inter-particle space x, while subsystem 2 (S2) is the rest of the system. We will compute the probability distribution $P(n_x)$ of the length n_x of space x. We define the multiplicity function of subsystem S2, called $\Omega_2(L_2, E_2)$, as the number of states accessible for S2 given the available free length for S2, L_2, and the available energy for S2, E_2. Note that $L_2 = L - n_x$ and $E_2 = E - E_1 = E - U(n_x)$. Then the probability that inter-particle space x has length n_x is proportional to the number of states that are accessible to the rest of the system, S2, given that x has length n_x:

$$P(n_x) \propto \Omega_2(L - n_x, E - U(n_x)). \tag{A4.2}$$

By definition, the entropy $\sigma_2(L_2, E_2)$ of S2 is the logarithm of $\Omega_2(L_2, E_2)$. Therefore,

$$P(n_x) \propto \mathrm{e}^{\sigma_2(L - n_x, E - U(n_x))}. \tag{A4.3}$$

Assuming that n_x is small compared to L and that $U(n_x)$ is small compared to E, we can now expand the entropy as follows:

$$\sigma_2(L - n_x, E - U(n_x)) = \sigma_2(L, E) - n_x \frac{\partial \sigma_2(L, E)}{\partial L} - U(n_x) \frac{\partial \sigma_2(L, E)}{\partial E} + \ldots \tag{A4.4}$$

Note that by the standard Maxwell relations,

$$(\partial \sigma_2 / \partial E)_L = 1/k_{\mathrm{B}}T \qquad \text{and} \qquad (\partial \sigma_2 / \partial L)_E = p/k_{\mathrm{B}}T, \tag{A4.5}$$

where T and p are the temperature and the pressure of the system. If N is large, the higher order terms are negligible.

Now we can combine the expansion in Equation A4.4 with Equation A4.3 and obtain:

$$P(n_x) \propto \mathrm{e}^{-(n_x p + U(n_x))/k_{\mathrm{B}}T}. \tag{A4.6}$$

We can calculate this in full using the definition of the potential in Equation A4.1, arriving at:

$$P(n_x) = \begin{cases} c\,\mathrm{e}^{-n_x(\lambda - \epsilon)} & \text{if } n_x < r, \\ c\,\mathrm{e}^{-n_x \lambda + \epsilon r} & \text{if } n_x \geq r. \end{cases} \tag{A4.7}$$

Here λ is defined as $\lambda \equiv p/k_{\mathrm{B}}T$. As we picked inter-particle space x arbitrarily, this probability distribution holds for all inter-particle spaces. The number c is a normalization constant. Given r and ϵ, the value of λ is fixed if we impose the mean inter-particle distance:

$$\int n_x P(n_x)\,\mathrm{d}n_x = \frac{L}{N}. \tag{A4.8}$$

Note that, beyond the interaction range, the distribution is exponentially decreasing. Within the interaction range, the distribution is also exponential, but the sign of the exponent depends on the size of ϵ: if the repulsion is strong ($\epsilon > \lambda$), the exponent becomes positive in the interaction range. We also note that if either the range r or the repulsion ϵ is set to zero, the distance distribution simply becomes a single exponential. This is the result for the Tonks-Gas model (Tonks, 1936).

4.A.3 CF interactions with different ranges

In the previous subsection we discussed a CF model in which each particle interacts with its neighbors according to one fixed interaction range. In the relevant case, however, the interaction range depends on the mutual orientation of the particles. The interaction potentials for convergent (C), tandem (T) and divergent (D) pairs can be written as follows:

$$\begin{aligned}
\frac{U_{\mathrm{C}}(n)}{k_{\mathrm{B}}T} &= \begin{cases} \epsilon(2\tau - n) & \text{if } n \leq 2\tau, \\ 0 & \text{if } n > 2\tau, \end{cases} \\
\frac{U_{\mathrm{T}}(n)}{k_{\mathrm{B}}T} &= \begin{cases} \epsilon(\tau + \pi - n) & \text{if } n \leq \tau + \pi, \\ 0 & \text{if } n > \tau + \pi, \end{cases} \\
\frac{U_{\mathrm{D}}(n)}{k_{\mathrm{B}}T} &= \begin{cases} \epsilon(2\pi - n) & \text{if } n \leq 2\pi, \\ 0 & \text{if } n > 2\pi. \end{cases}
\end{aligned} \tag{A4.9}$$

It is rather straightforward to adjust the calculations in the previous section to this case.

We again divide the system in two parts, S1 and S2, in which S1 consists of one inter-particle region called x, and S2 is the rest of the system. The derivation in the previous section applies without alteration up to Equation A4.6, irrespective of the orientation corresponding to x (that is: C, T or D). Only in the step from Equation A4.6 to Equation A4.7, the difference in the potentials for D, T and C becomes relevant. As a result, the distributions for the C, T and D intergenic regions all have the form of Equation A4.7, except for a different range r, and a different normalization factor c:

$$P_{\mathrm{C}}(n) = \begin{cases} c_1\,\mathrm{e}^{-n(\lambda-\epsilon)} & \text{if } 0 \leq n \leq 2\tau, \\ c_1\,\mathrm{e}^{-n\lambda+2\tau\epsilon} & \text{if } n > 2\tau, \end{cases}$$

$$P_{\mathrm{T}}(n) = \begin{cases} c_2\,\mathrm{e}^{-n(\lambda-\epsilon)} & \text{if } 0 \leq n \leq \tau+\pi, \\ c_2\,\mathrm{e}^{-n\lambda+(\tau+\pi)\epsilon} & \text{if } n > \tau+\pi, \end{cases} \tag{A4.10}$$

$$P_{\mathrm{D}}(n) = \begin{cases} c_3\,\mathrm{e}^{-n(\lambda-\epsilon)} & \text{if } 0 \leq n \leq 2\pi, \\ c_3\,\mathrm{e}^{-n\lambda+2\pi\epsilon} & \text{if } n > 2\pi. \end{cases}$$

$$\tag{A4.11}$$

Here the prefactors c_1, c_2 and c_3 are defined as

$$\begin{aligned} c_1 &= \frac{\lambda(\lambda-\epsilon)}{\lambda-\epsilon\,\exp(2\tau(\epsilon-\lambda))}, \\ c_2 &= \frac{\lambda(\lambda-\epsilon)}{\lambda-\epsilon\,\exp((\tau+\pi)(\epsilon-\lambda))}, \\ c_3 &= \frac{\lambda(\lambda-\epsilon)}{\lambda-\epsilon\,\exp(2\pi(\epsilon-\lambda))}. \end{aligned} \tag{A4.12}$$

We note that the CF model has four parameters: λ, τ, π, and ϵ. However, if we impose the average length of the intergenic regions, this again leads to a constraint that eliminates one of the parameters. As the total system is a mixture of D, C and T intergenic regions in proportions $f_{\mathrm{D}} : f_{\mathrm{C}} : f_{\mathrm{T}}$ (in most genomes roughly 1:1:2), this constraint becomes:

$$\int n \frac{f_{\mathrm{D}}P_{\mathrm{D}}(n) + f_{\mathrm{C}}P_{\mathrm{C}}(n) + f_{\mathrm{T}}P_{\mathrm{T}}(n)}{f_{\mathrm{D}} + f_{\mathrm{C}} + f_{\mathrm{T}}}\,\mathrm{d}n = \frac{L}{N}. \tag{A4.13}$$

We used Monte Carlo simulations to check the validity of these equations and found excellent agreement.

4.B More detailed models

The CF model is purposely oversimplified. Such simplified models, with few parameters, provide insight into the essential ingredients of the mechanisms studied. At the same time the simplicity of the CF model leads to certain artifacts. Here we show that such artifacts can be alleviated by more detailed models. Below we discuss how one can allow for varying UCR and DCR lengths, and how alternative interaction potentials can be chosen, with distance-dependent forces.

4.B.1 Polydisperse UCRs and DCRs

The distributions of the CF model have a sharp peak; this is an artifact of our assumption that all UCRs and all DCRs have the same length. We can extend the model to describe systems with varying UCR and DCR lengths.

If UCR and DCR lengths vary, then this results in a varying interaction range r. In general, due to differences in the UCR and DCR lengths, the interaction range obeys probability distributions $Q_{\mathrm{C}}(r)$, $Q_{\mathrm{D}}(r)$ and $Q_{\mathrm{T}}(r)$ for the convergent, divergent and tandem intergenic regions respectively. Then at a given pressure λ the distributions of intergenic distances are given by

$$\begin{aligned}
P_{\mathrm{C}}(n) &= \int_0^\infty Q_{\mathrm{C}}(r) P(n|\lambda, r)\,\mathrm{d}r,\\
P_{\mathrm{D}}(n) &= \int_0^\infty Q_{\mathrm{D}}(r) P(n|\lambda, r)\,\mathrm{d}r,\\
P_{\mathrm{T}}(n) &= \int_0^\infty Q_{\mathrm{T}}(r) P(n|\lambda, r)\,\mathrm{d}r.
\end{aligned} \tag{A4.14}$$

Here $P(n|\lambda, r)$ is the probability distribution for the length n of an intergenic region, given the interaction range r and the pressure λ; it depends on the the form of the interaction potential. For instance, if the potential is that of the CF model (Equation A4.1), then $P(n|\lambda, r)$ is given by Equation A4.7. Note that we retrieve the original CF model if we insert $Q_{\mathrm{C}}(r) = \delta(r - 2\tau)$, $Q_{\mathrm{D}}(r) = \delta(r - 2\pi)$ and $Q_{\mathrm{T}}(r) = \delta(r - (\tau + \pi))$ in the above integrals.

In the case of *S. cerevisiae*, some studies (David *et al.*, 2006; van Helden *et al.*, 2000; Graber *et al.*, 2002; Perocchi *et al.*, 2007) suggest that the distribution of 3' UTRs is log-normal. We therefore assume that $Q_{\mathrm{C}}(r)$ is the distribution of the sum of two numbers drawn independently from a log-normal distribution. A sum of log-normally distributed random variables can be approximated reasonably by another log-normal distribution. We therefore assume that $Q_{\mathrm{C}}(r)$ is log-normal as well (with parameters μ and σ). The corresponding fit to the histogram of convergent intergenic distances in *S. cerevisiae* is better than the fit of the CF model and does not show the artifactual sharp peak (see Fig. 4.6). Nevertheless, the mean of the best-fitting log-normal distribution ($\mu = 4.61$, $\sigma = 0.405$, mean $= \exp(\mu + \sigma^2/2) = 109$) is rather close to the estimate resulting from the CF model ($2\tau = 122$). Below we show that a better agreement with experiment can be obtained if we allow for an alternative interaction potential.

4.B.2 Alternative potentials

In the CF model, we used the simple potential defined in Equation A4.1. This potential was convenient because of its simplicity (only one parameter) and its straightforward interpretation. It is, however, possible to generalize our approach to alternative potentials. Equation A4.6 holds for any finite-ranged potential $U(n)$; this means that Equations A4.6 and A4.14 can be used to compute the

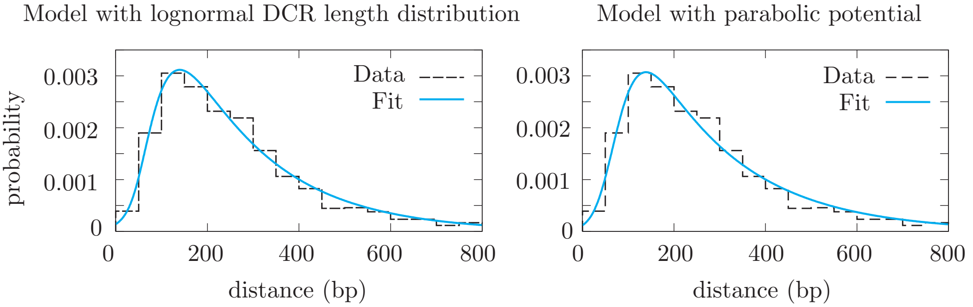


Figure 4.6: Fits of two more detailed models to the length distribution of convergent intergenic regions in *S. cerevisiae*. The fact that both models allow for nearly perfect fits shows that one needs additional, independent information to distinguish between the various compatible models. Left: CF model with log-normally distributed DCR lengths. Fit parameters for the log-normal distribution: $\mu = 4.61$, $\sigma = 0.405$, $\lambda = 5.25$, $\epsilon = 4.70 \times 10^{-2}$. Right: Log-normally distributed DCR lengths and parabolic potential. Fit parameters: $\mu = 5.01$, $\sigma = 0.314$, $\lambda = 4.95 \times 10^{-3}$, $U_0 = 4.11$.

ORF spacing for arbitrary finite-range potentials.

4.B.3 Yeast DCRs

In the main text, we mentioned that the CF model predictions for the DCR length are on the low side. Not much is known about termination sequences in *S. cerevisiae*, but most of the poly-adenylation signals seem to occur within 70 bp from the stop codon (van Helden *et al.*, 2000). Estimates for the median 3' UTR length in *S. cerevisiae* range from 80 to about 100 bp (David *et al.*, 2006; Graber *et al.*, 2002). These numbers are indeed a bit higher than our result of 61 bp in Table 4.1. This suggests that, even though the CF model does predict the qualitative features of the distributions in Yeast, such as the exponential tail of the distribution, in order to get accurate quantitative agreement with the DCR lengths found in recent experiments, the assumptions of the CF model are too crude. Using the above techniques, we can refine the model and get better agreement.

First, recent studies suggest that the 3' UTRs in Yeast can be approximated by a log-normal distribution; we therefore now choose $Q_\mathrm{C}(r)$ to be log-normal (with parameters μ and σ). Second, recent experiments strongly suggest that many 3' UTRs are long and that they often overlap considerably (David *et al.*, 2006; Perocchi *et al.*, 2007); nevertheless, the ORFs hardly ever get closer together than 120 bp. This suggests a model in which the force is not constant; instead, the repulsion seems to be high at short distances, but low at longer distances. One way to model this is to use the following quadratic potential instead of

Equation A4.1:

$$\frac{U(n)}{k_\mathrm{B}T} = \begin{cases} U_0\left(1-\frac{n}{r}\right)^2 & n < r, \\ 0 & n \geq r. \end{cases} \tag{A4.15}$$

The fit of this model (with μ, σ, U_0 and λ as parameters, but a given mean distance) to the convergent data is excellent (see Fig. 4.6); also, the resulting log-normal distribution for $Q_\mathrm{C}(r)$ has a mean $\exp(\mu+\sigma^2/2) = 160$, which leads to a mean DCR length of about 80 bp. This is in good agreement with the experimental results.

In the study of van Helden *et al.* (2000), poly-adenylation signals were found at about 35 bp and 55 bp downstream of the stop codon of ORFs. It is tempting to speculate that these sequences are responsible for the strong repulsion starting at a distance of about 120 bp in convergent intergenic regions.

4.B.4 Higher eukaryotes

In Fig. 4.4 and 4.7 the intergenic distance distributions for various different organisms are shown. Strikingly, the simple CF model can very well describe the qualitative features of all these model organisms, such as the exponential tail of the distributions and the dependence of the distributions on orientation.

In complex, multicellular eukaryotes, control regions typically are very long and exhibit a high variance (*e.g.* Hajarnavis *et al.* (2004)). As the lengths and variances increase, the assumptions of the constant force model become less justified. Above we have shown that the CF model can be extended to incorporate alternative potentials and polydisperse interaction ranges. This allows us to produce excellent fits to the data for all organisms. Nevertheless, when it comes to *predicting* the length distributions of UCRs and DCRs for higher organisms, the results depend too sensitively on the choice of the potential to produce meaningful predictions. Therefore we refrain from using the fit parameters for *D. melanogaster*, *A. thaliana*, *C. elegans* and *P. falciparum* as predictions for the DCR and UCR lengths.

4.C Fitting procedure

Fig. 4.4 shows the distributions of intergenic distances for four different fungi and four additional eukaryotes, broken down into three different subsets (convergent, tandem, divergent). Here we describe the procedure we used to fit the model to these data. Since the divergent set does not fit the data — we argued that this is due to bi-directional promoters — we only fit the C and T data. This is sufficient to obtain estimates for π and τ.

The C and T distributions are also displayed in Fig. 4.7 in log-linear scale, combined with fits of the CF model. We used these fits to estimate UCR and DCR sizes in these species. The fit parameters are given in Table 1.

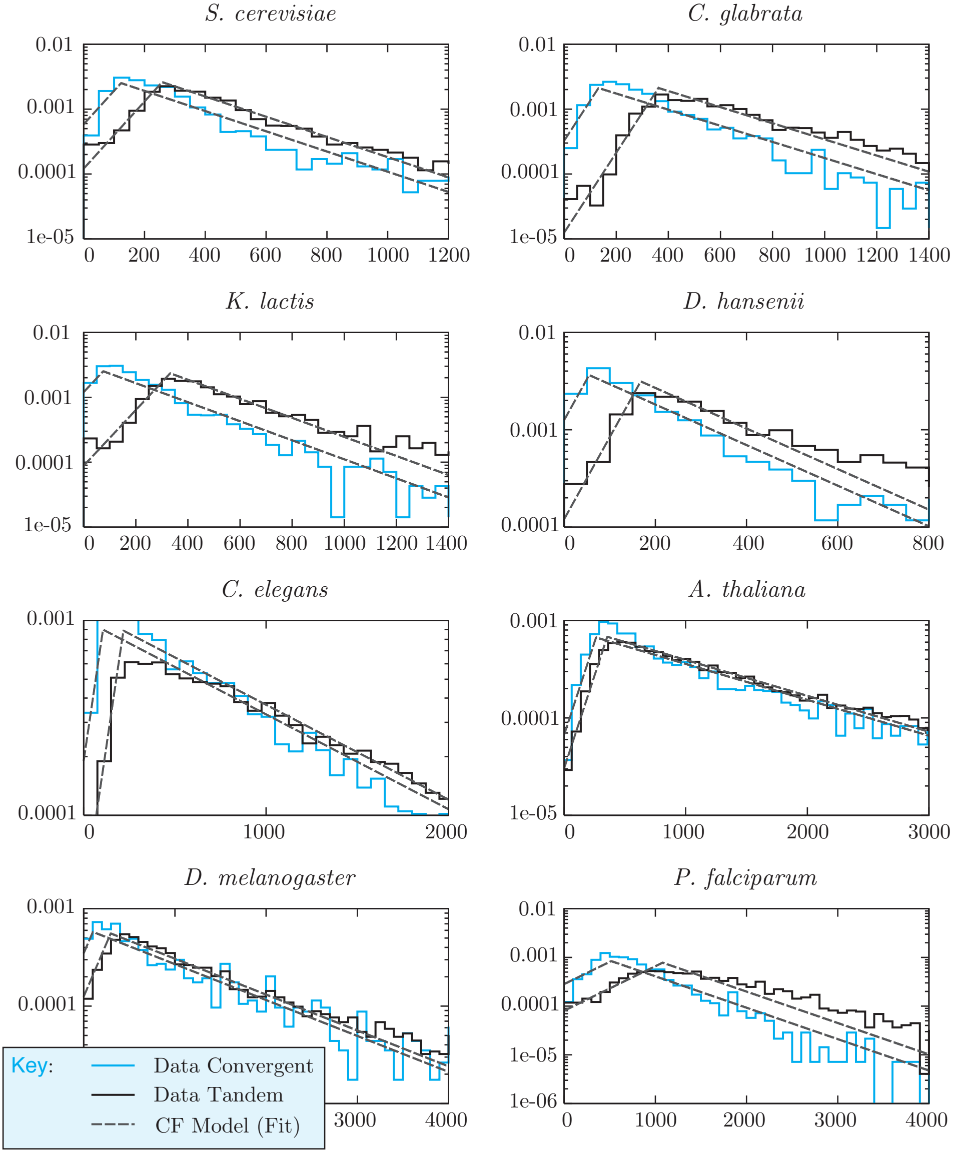


Figure 4.7: Maximum likelihood fits of the CF model to the distance distributions of four fungi and four additional eukaryotes. Despite the simplicity of the CF model, it does capture the qualitative features of each of the genomes, such as the dependence of the ORF spacing on relative orientation, and the exponential tails. The fits are used to estimate the sizes of UCRs and DCRs for the fungi; see Table 4.1. Such quantitative estimates are probably not reliable for the higher eukaryotes, for which the model assumptions may be too crude.

We adopt a maximum likelihood method to fit our model to the data and to determine the errors in the fit parameters. For a given set of observed intergenic distances ($\{n_{\mathrm{C}}\}$ and $\{n_{\mathrm{T}}\}$ for convergent and tandem pairs, respectively), Bayes' rule states that the likelihood of a set of fit parameters obeys

$$P(\pi,\tau,\lambda,\epsilon\,|\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}) = P(\,\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}|\,\pi,\tau,\lambda,\epsilon)\frac{P(\pi,\tau,\lambda,\epsilon)}{P(\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\})}.$$

Here $P(\pi,\tau,\lambda,\epsilon)$ is the prior probability distribution, which we take to be uniform. In that case

$$P(\pi,\tau,\lambda,\epsilon|\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}) \propto P(\,\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}|\,\pi,\tau,\lambda,\epsilon).$$

The parameter values with maximal likelihood are therefore those that maximize $P(\,\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}|\,\pi,\tau,\lambda,\epsilon)$.

In practice, it is more convenient to work with the logarithm of the likelihood, as

$$\begin{aligned}\log\left(P(\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}|\pi,\tau,\lambda,\epsilon)\right) &= \log\left(\prod_{n\in\{n_{\mathrm{C}}\}} P_{\mathrm{C}}(n) \prod_{n'\in\{n_{\mathrm{T}}\}} P_{\mathrm{T}}(n')\right)\\ &= \sum_{n\in\{n_{\mathrm{C}}\}} \log\left(P_{\mathrm{C}}(n)\right) + \sum_{n'\in\{n_{\mathrm{T}}\}} \log\left(P_{\mathrm{T}}(n')\right).\end{aligned}$$

The functions $P_{\mathrm{C}}(n)$ and $P_{\mathrm{T}}(n)$ were given in Equation A4.11. If we define

$$X_{\mathrm{C}}^{\leq} \equiv \sum_{\substack{n\in\{n_{\mathrm{C}}\}\\ n\leq 2\tau}} n, \qquad X_{\mathrm{C}}^{>} \equiv \sum_{\substack{n\in\{n_{\mathrm{C}}\}\\ n\;>2\tau}} n, \qquad X_{\mathrm{T}}^{\leq} \equiv \sum_{\substack{n\in\{n_{\mathrm{T}}\}\\ n\leq\tau+\pi}} n, \qquad X_{\mathrm{T}}^{>} \equiv \sum_{\substack{n\in\{n_{\mathrm{T}}\}\\ n>\tau+\pi}} n,$$

and call the total number of convergent and tandem pairs n_{C} and n_{T}, this reduces to

$$\begin{aligned}\log\left(P(\{n_{\mathrm{C}}\},\{n_{\mathrm{T}}\}|\pi,\tau,\lambda,\epsilon)\right) = n_{\mathrm{C}}\log(c_1) + n_{\mathrm{T}}\log(c_2) - (\lambda-\epsilon)\left(X_{\mathrm{C}}^{\leq} + X_{\mathrm{T}}^{\leq}\right)\\ - \lambda\left(X_{\mathrm{C}}^{>} + X_{\mathrm{T}}^{>}\right) + 2n_{\mathrm{C}}\epsilon\tau + n_{\mathrm{T}}\epsilon(\tau+\pi),\end{aligned} \tag{A4.16}$$

which can be maximized straightforwardly. To avoid possible influences of rare outliers, we only used values of n that fall in the domain that is plotted Fig. 4.7. This is correct if we modify c_1 and c_2 in Equation A4.16 such that, given the domain D,

$$\int_D P_{\mathrm{C}}(n)\,\mathrm{d}n = \int_D P_{\mathrm{T}}(n)\,\mathrm{d}n = 1. \tag{A4.17}$$

If we plot the likelihood as a function of one of the parameters, while keeping the other parameters at their maximum likelihood value, the plot can very well be approximated by a Gaussian. We use the standard deviation of this Gaussian as the error in the maximum likelihood parameter values.

In the main text, we discussed the values of τ and π, but not of q. The probability that two randomly chosen base pairs are the same and could therefore overlap is 1/4. The fact that q is much higher than 1/4 shows that "overlap" is much easier than expected based on this argument. This could reflect the density of functional elements, but also the flexibility of functional sequences, and the fact that regulatory regions are not mono-disperse.

4.D Statistical tests

In this section we describe the statistical tests that are mentioned in the main text. The tests

4.D.1 Co-expressed divergent pairs in *S. cerevisiae* are closer together than expected

In the main text, we state that co-expressed divergent gene pairs in *S. cerevisiae* have a tendency to have short intergenic regions. We tested this hypothesis as follows.

We used the expression data compiled by Dr. Andre Boorsma and Prof. Harmen J. Bussemaker to compute, for each neighboring pair of genes, the Pearson's correlation coefficient of their expression in about 900 experiments[4]. These coefficients were usually low (< 0.3). We then split the set of divergent pairs into two subsets: those with a low correlation coefficient (< 0.3), called set 1, and those with a high one (≥ 0.3), called set 2. Next, we used a rank sum test to check whether the intergenic regions in set 2 are indeed shorter than expected at random.

The rank sum test was performed as follows. We first ranked the intergenic regions according to their length. Then, we computed the sum of the ranks of the intergenic regions in set 2; we call it R. Next, we randomized the ranks in the data set 10^7 times, each time computing the rank sum of set 2 in the randomized data. Finally, we counted the number of times R was smaller or equal to the rank sums obtained from the randomized data. The results show that the ORF pairs with a high correlation coefficient are significantly closer together than the ones with a low one ($p < 0.002$).

We checked that the observed signal is not due to paralogous gene pairs by excluding them from the set and repeating the test; this did not change the result. As a control experiment, we tested whether the same signal is also present in

---

[4] See the following references for the original publications: Boer *et al.* (2003); Boorsma *et al.* (2004); Bro *et al.* (2003); Chu *et al.* (1998); Daran-Lapujade *et al.* (2004); Devaux *et al.* (2001); Fleming *et al.* (2002); Gasch *et al.* (2000, 2001); Harris *et al.* (2001); Hughes *et al.* (2000); Lagorce *et al.* (2003); McCammon *et al.* (2003); Mnaimneh *et al.* (2004); Murata *et al.* (2003); Sahara *et al.* (2002); Spellman *et al.* (1998); Tai *et al.* (2005); Yoshimoto *et al.* (2002)

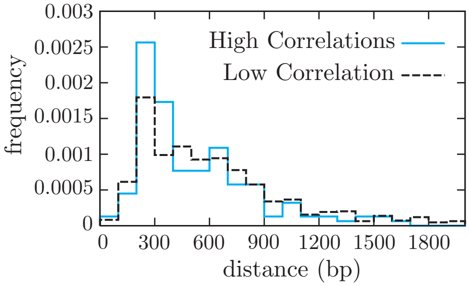


Figure 4.8: Distance distributions for divergent, neighboring gene pairs with high or low correlation in their expression, in *S. cerevisiae*. The figure shows that pairs with a high correlations coefficient are more likely to be close together than the ones with low correlation coefficients. This fact supports the hypothesis that the pairs in the first peak have a shared, bidirectional *cis*-regulatory region.

the set of *tandem* neighbors. This is not the case (rank sum test: $p = 0.56$). We note, however, that a similar signal *was* present in the set of convergent pairs ($p < 2 \times 10^{-4}$); we have no satisfactory explanation for this fact.

The mean intergenic distance in set 1 is 721 bp; in set 2, it is 558 bp. The difference in the distance distributions of both sets is visually apparent (see Fig. 4.8).

4.D.2 Divergent genes in *S. cerevisiae* are more likely to be associated with the same process or component if they are close together

We studied whether there is an association between intergenic distance and functional similarity in divergent gene pairs in *S. cerevisiae*. In order to do this, we need to be able to quantify the functional similarity between two given genes. For this purpose we used the GO annotations of the GO Consortium (version 5.463 of Aug. 22 2007, see Ashburner *et al.* (2000)) and the information-theoretic measure for semantic similarity proposed by Resnik (1998).

The Gene Ontology is a hierarchical vocabulary of terms that can be assigned to genes. It falls apart into three independent taxonomies, each defined to describe one aspect of genes: the *biological process* they are involved in, the *molecular function* they perform and the *cellular component* they are associated with. Pairs of GO terms can be linked by five types of relations, such as the “is a” or “is part of” relations. For instance, as the nuclear membrane is part of the nucleus, the GO term “nuclear membrane” is related to the term “nucleus” through an “is part of” relation. Through such relations, the GO terms form a hierarchy.

Note that if a gene is assigned the term “nuclear membrane”, meaning that it is associated to the nuclear membrane, it is implicitly also assigned the term “nucleus” owing to the “is part of” relation. For convenience we will use the following notation: if the assignment of the term b implies the assignment of term a through “is a” or “is part of” relations, we write $b \Rightarrow a$.

The semantic similarity measure defined by Resnik (1998) was first proposed for and applied to the processing of natural language. Applied to the GO ontology it provides, for each pair of GO terms a and b, a similarity score. For this, it uses the information content of each term, which is a function of the probability value $p(a)$ that is assigned to each term a. We define $p(a)$ as the fraction of the genes in the given organism that has either been assigned a or any other term b for which $b \Rightarrow a$ holds. The information content is then defined as $I(a) \equiv -\log(p(a))$.

The similarity measure $s(a, b)$ is defined in terms of the information content:

$$s(a,b) \equiv \max_{c \in C} I(c) \quad \text{in which} \quad C = \{c \,|\, a \Rightarrow c \text{ and } b \Rightarrow c\}. \tag{A4.18}$$

If, for a given aspect α (biological process, molecular function or cellular component), gene A has been assigned GO terms $a_1 \ldots a_n$, and gene B has been assigned term $b_1 \ldots b_m$, then we define the similarity between genes A and B on this aspect as:

$$S(A, B, \alpha) \equiv \max_{i,j} s(a_i, b_j). \tag{A4.19}$$

Thus we can compute similarity values for each gene pair and for each aspect of the GO ontology.

If a certain gene did not have any assignment for some aspect, then the similarity score with any other gene was considered undefined on this aspect, and the gene was excluded from the analysis corresponding to this aspect.

We performed the following statistical tests. First, we divided the data set in two subsets, set 1 and set 2. The first set consisted of all divergent pairs that were in the "first" peak and set 2 contained all divergent neighbors that were further apart. As the bordering value we chose $d = 2\pi = 375\,\text{bp}$, since intergenic regions that are longer than that value can easily accommodate two independent promoters. We applied a Wilcoxon–Mann–Whitney rank-sum test to challenge the null hypothesis that the similarity scores of the pairs in set 1 and set 2 are drawn from the same distribution. We repeated this test for each aspect of the GO. The test results are $p = 8.8 \times 10^{-4}$, $p = 0.06$ and $p = 4.8 \times 10^{-5}$ for biological process, molecular function and cellular component respectively. We also ran a Spearman rank correlation test on the data, which resulted in the values $p = 2.0 \times 10^{-5}$, $p = 0.24$ and $p = 1.3 \times 10^{-5}$.

We conclude that the similarity scores belonging to the aspects biological process and cellular component are associated with intergenic distance. We do not, however, find a significant association between molecular function and intergenic distance. This is not very surprising, as proteins with a similar molecular function (*e.g.* "DNA binding" proteins), can act in very different processes and cellular components, so that there is no clear a priori reason to co-regulate them using a shared UCR.

We repeated this analysis for the convergent and tandem gene pairs. For the convergent pairs, none of the statistics were significant. However, the tandem pairs showed a similar pattern as the divergent ones; the Spearman rank correlation test resulted in $p = 4.6 \times 10^{-5}$, $p = 0.41$ and $p = 1.8 \times 10^{-5}$ for biological process, molecular function and cellular component respectively.

4.D.3 Greater than expected number of convergent intergenic regions with length 20–60 bp in *E. coli*

Here we show that the number of convergent intergenic regions with a length in the range 20–60 bp, is significantly larger than expected in *E. coli*.

In the calculation below, we estimate the significance of the peak in a very conservative way. We take the best-fitting exponential probability distribution as our null distribution (the fit is shown in Fig. 1(b) in the main text; its scale factor equals 145 bp). This way, we *underestimate* the statistical significance of the peak as we ignore the fact that the CF model actually predicts a *dip* in the distribution at the place of the peak.

Given the exponential null distribution, the fraction of the sample that is expected in the domain 20–60 bp is 0.21. Since the total number of convergent pairs is 543, the number of pairs in this domain is a random variable X that is distributed binomially with $p = 0.21$ and $N = 543$. The observed number of pairs in this domain in *E. coli* is 198; the probability for this to happen given the null distribution is $P(x \geq 198; p = 0.21, n = 543) < 10^{-13}$.

Based on the numbers above we should have expected $0.21 \times 543 = 114$ pairs in the domain 20–60 bp. The actual observed number is 198; this means that we need about 86 bi-directional terminators to explain the data. If we assume that *E. coli* has 750 operons, we estimate that at least $(2 \times 86/750) \times 100\% = 23\%$ of the operons is terminated by a bi-directional terminator.

4.D.4 In *E. coli*, putative terminators in C regions are more often bi-directional than those in T regions

Here we show that the fraction of putative terminators that is classified as bi-directional by `RNAMotif` software (Lesnik *et al.*, 2001), is larger in C regions than in T regions.

The statistical test was performed as follows. Our null hypothesis is that the terminators in the C region are a random sample from the total set of terminators in C or T regions. In total, the C and T regions together contain 1198 putative terminators, of which 222 are classified as bi-directional by `RNAMotif`. The C regions contain 378 putative terminators, of which 104 are bi-directional according to `RNAMotif`. If the 378 are chosen at random from the total set of 1198 terminators, then the number of terminators in the sample that are classified as bi-directional is a hypergeometric random variable. The probability to observe at least 104 bi-directional terminators in a random sample of 378 terminators, taken from a set of 1198 terminators containing 222 bi-directional ones, equals 1×10^{-7}.

4.D.5 Putative bi-directional terminators in C regions tend to occur in short regions

The fraction of the putative terminators in convergent intergenic regions that could be bi-directional (according to the `RNAMotif` algorithm) is significantly larger in short intergenic regions ($< 100\,\mathrm{bp}$) than in long ones. We used the same statistical test as in the previous subsection. In total, the C regions contain 378 putative terminators; of these, 104 are classified as bi-directional. The short convergent intergenic regions contain 158 putative terminators, of which 62 are bi-directional according to `RNAMotif`. The probability to observe at least 62 bi-directional terminators in a random sample of 158 terminators, taken from a set of 378 terminators containing 104 bi-directional ones, is 2×10^{-5}.

4.D.6 Convergent operons that are close together are not more often functionally related

We tested whether the convergent operons that are close together ($< 70\,\mathrm{bp}$) are more likely to be active in the same biological process or cellular component than ones that are further apart. For this we used the GO annotations from the GOA Database (Camon *et al.*, 2004)(version date: September 9. 2007). The same method was used as in Section 4.D.2, except that we now had to perform the analysis on the level of operons rather than genes. In order to compute the similarity between two operons, we compared the GO assignments for each gene in the first operon with each gene in second; the maximum of these scores was used as a similarity measure for the operons. We did not find a significant signal for any of the aspects of the Gene Ontology (molecular function: $p = 0.20$; biological process: $p = 0.25$; cellular component: $p = 0.27$).

Chapter 5

The role of terminator loss in the evolution of genomes

Any genomic sequence is continuously challenged by mutations such as nucleotide substitutions, insertions, deletions and inversions. Those sequences that are not under sufficient purifying selection are therefore likely to be destroyed sooner or later. Here we study the implications of this use-it-or-lose-it principle for the destiny of transcriptional terminator sequences in prokaryotic genomes. Terminators of genes that are not expressed for an appreciable period of time experience a significantly reduced purifying selection. This introduces the risk of terminator loss. We argue that the loss of a terminator can directly lead to the emergence of operons and the utilization of bi-directional terminators. To prove the concept, we present a simple model of genome evolution and developed a novel simulation scheme based on population genetics. In this model, operons and shared terminators indeed emerge spontaneously. Moreover, the model reproduces the spacing of genes in the model prokaryotes *Escherichia coli* and *Bacillus subtilis*, including the characteristic close spacing of genes in operons and the differences in spacing between convergent, divergent and tandem gene pairs.

5.1 Introduction

In evolution, the maxim “Use it or lose it” holds quite generally. Any genomic sequence that is not under sufficient purifying selection is destined to be washed away by a continuous stream of mutations. For example, transcription factor binding sites that are not being used most of the time are likely to be lost. It has been suggested that, for this reason, genes that are not being used most of the time are typically repressed when they are not necessary, rather than activated when they are needed — only thus a constant selective force protects the transcription factor binding site against deleterious mutations (Savageau, 1977). Here we ask what this use-it-or-lose-it principle implies about the fate of transcriptional terminator sequences in prokaryotic genomes (see Box 5.1). We argue that the loss of terminators by random mutations can directly lead to the formation of operons and the emergence of bi-directional terminators. Terminator loss can thus play a critical role in shaping the spatial and orientational distribution of genes along the genome.

An extended literature is devoted to the question why genes in many (mainly prokaryotic) genomes are organized in operons. Two main lines of thought have been brought forward. The first is based on the fact that genes within one operon are naturally co-regulated and co-expressed (Jacob and Monod, 1961b; Price *et al.*, 2005b, 2006). Operon formation could therefore have evolved as a means to co-express genes. Indeed, genes in one operon are often — but not always — functionally related (de Daruvar *et al.*, 2002). However, genes that are not in the same operon can also be co-regulated by independent but similar promoters. Assuming that from a functional perspective these two alternative arrangements are equally valuable, the question rises which of these configurations is more likely to emerge in the course of evolution. Some have argued that, in an evolutionary context, two independent promoters are easier to develop than one operon, in particular because in the latter case the two genes have to get in close proximity by chance (Lawrence and Roth, 1996; Lawrence, 1999). Others have claimed the opposite: that re-arrangements are frequent and that operons form more rapidly than multiple promoters would, especially if complex transcriptional regulation is required (Price *et al.*, 2005b).

The second line of thought argues that clustering related genes does not necessarily provide a fitness advantage to the organism *per se*, but instead allows groups of genes that are jointly required for a selectable phenotype to spread to other clades efficiently by horizontal gene transfer (HGT), thereby contributing to the reproductive succes of these genes. This model is called the “selfish operon model” (Lawrence and Roth, 1996). It provides a mechanism that gradually drives together genes that are functionally interdependent. Thus, it also explains that genes in a given operon are often functionally related. On the other hand, the model predicts that essential genes are typically not in operons, which is contradicted by empirical studies (Pal and Hurst, 2004; Price *et al.*, 2005b). Moreover, many operons contain genes that are not functionally related and new

Box 5.1: Rho-dependent and Rho-independent transcriptional termination

In bacteria and bacterial phages, transcription termination is achieved by two mechanisms: Rho-independent (also called “intrinsic”) or Rho-dependent termination.

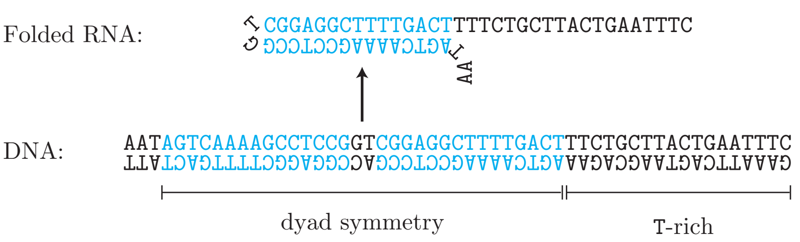
Rho-dependent terminators rely on the binding of the protein Rho to the transcript. Rho is a homo-hexameric protein; the hexamer forms a ring around the transcript. When bound to the message, Rho can translocate in a 5’ to 3’ direction along the transcript. This process is ATP-dependent; Rho acts as an ATPase. Even though many details are not yet known, it is believed that Rho can catch up with the RNAp when RNAp pauses at one of its pause sites. The subsequent interaction between the Rho hexamer and RNAp leads to the termination of the transcript.

Rho-dependent termination is often highly regulated. Rho interacts with several termination factors, such as NusA, NusB and NusG. This allows for a regulated choice between alternative termination sites. (See Ciampi (2006) for a review on Rho-dependent termination and further references.)

Rho-independent termination does not rely on the action of a particular protein. Instead, the termination is triggered by the folding of the RNA transcript into a stem-loop structure (see the Example below). The terminator generally consists of two parts: a thymidine rich sequence of about twelve nucleotides, preceded by a palindromic sequence that, as RNA, folds in a “hairpin”-like structure due to ordinary Watson–Crick base pairing. How exactly these sequences force termination of the transcript, is not fully understood (Lesnik *et al.*, 2001; Alberts *et al.*, 1994).

A Rho-independent terminator can be bi-directional, meaning that it is also functional on the opposite strand. This results from the fact that the complementary strand of the palindromic part of the terminator is automatically a palindrome too. This means that the addition of a `T`-rich sequence on the complementary strand may often be sufficient to turn a uni-directional terminator into a bi-directional one.

Example: The P14-tonB transcriptional terminator


Folded RNA:
CGGAGGCTTTTGACTTTTCTGCTTACTGAATTTC
DNA:
AATAGTCAAAAGCCTCCGGTCGGAGGCTTTTGACTTTCTGCTTACTGAATTTC
dyad symmetry
T-rich

operons form regularly from genes that have not been obtained by horizontal transfer (Pal and Hurst, 2004; Price *et al.*, 2005b). Also, even though the selfish operon model does explain why interdependent gene sets would cluster on the genome, it does not readily elucidate why the transcription units of these genes would merge into one poly-cistronic transcription unit (TU). This final step requires that co-regulation of the genes involved should at the very least not impede their function, showing that co-regulation plays an important role in the selfish operon model as well. Clearly, the selfish operon model and the co-regulation model are not mutually exclusive and both mechanims might contribute to some extent to the formation of operons.

In the discussions of both models, it is implicitly or explicitly assumed that operons would not exist in the absence of any selective pressure to create them. Here we suggest quite the opposite: even if operons do not have any selective advantage (neither at the level of organisms, nor at the level of clusters of genes), operons are expected to emerge. The reason is that the terminators that define the borders of TUs are continually challenged by myriad mutations. On evolutionary time scales they will survive only if they are under constant and sufficient purifying selection. Whenever this is not the case, the terminator will be lost, and operons form. As a proof of principle, we have developed a minimal model of genome evolution and a novel simulation scheme based on Kimura–Ohta population genetics. We show that, in this model, terminator loss directly leads to poly-cistronic TUs. We therefore claim: neighboring tandem TUs are likely to merge into operons unless there is sufficient selective pressure to express them independently.

A related question is what are the consecutive mutational events that most likely lead to the merging of mono-cistronic TUs into poly-cistronic ones. Neither the co-regulation model nor the selfish operon model offer or imply such detailed mechanisms and therefore most scenarios are compatible with both. We argue that terminator loss is a likely first step in the merging of nearby TUs. We show that a weak selection pressure on the lengths of RNA transcripts is sufficient to explain the subsequent shortening of the intergenic region between the genes in the newly-formed operon leading to the characteristic close spacing of genes in operons. As it turns out, the same selection pressure also explains the fact that both promoters and terminators tend to tightly flank the genes they belong to.

The sequences called Rho-independent or intrinsic terminators function by virtue of their tendency to fold, when transcribed into RNA, into stable hair-pin structures (Farnham and Platt, 1981). This requires that these sequences are largely palindromic. By definition, the complementary strand of a palindrome is palindromic as well, which explains why some terminators are bi-directional (Postle and Good, 1985; Carlomagno *et al.*, 1985). Such terminators can operate on both strands of the DNA and can therefore be shared by two convergently transcribed TUs. We argue that while terminator loss between *tandem* TUs leads to poly-cistronic transcripts, terminator loss between *convergent* TUs can

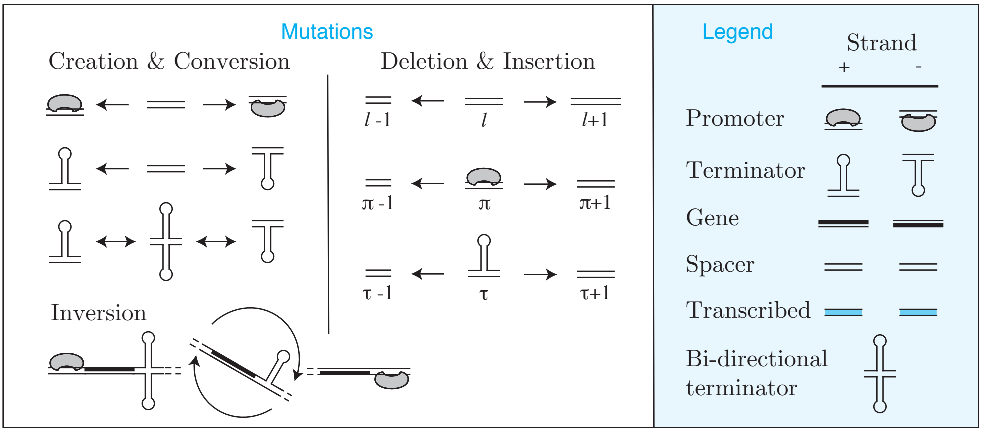


Figure 5.1: Cartoon illustrating the mutations allowed in our model. First, promoters and terminators can be created from spacers. Second, uni-directional terminators can be converted into bi-directional ones and vice versa. Third, promoters and terminators can be destroyed by insertions and deletions, which turns them into regular spacers. Insertions and deletions also result in changes in spacer lengths. Fourth, random pieces of DNA can be inverted.

lead to the exploitation of bi-directional terminators. The proposed selection pressure on the lengths of transcripts predicts that such convergent pairs should be very closely spaced, which is supported by the distribution of distances between convergent gene pairs in *E. coli* and *B. subtilis*.

5.2 Model

Below we describe a minimal computational model that leads to the spontaneous emergence of operons and shared bi-directional terminators due to terminator loss; it also reproduces the spacing of genes and operons in model prokaryotes such as *E. coli* and *B. subtilis*.

5.2.1 A daisy-chain genome

We represent the genome as a circular daisy chain consisting of objects of four kinds: spacers (stretches of non-coding DNA), genes, promoters and terminators. Each of these objects is endowed with several properties. First, each has a length expressed in base pairs. Terminators, promoters and genes have fixed lengths, whereas spacers can have any length. Second, promoters, terminators and genes have an orientation: they are either on the positive strand or on the negative strand. In addition, terminators can be bi-directional. Finally, genes have a property that we call “color”. The colors of genes symbolize their expression requirements. We impose that genes with different colors should be regulated separately and therefore cannot be in the same operon.

5.2.2 Four elementary types of mutations

To keep the model as transparent as possible, we allow for only four kinds of mutations, represented schematically in Fig. 5.1:

1. insertion or deletion of a single base pair. If an insertion or deletion takes place in a spacer, the length of the spacer is adjusted accordingly. If it occurs in a promoter, a terminator or a gene, we assume that this destroys it; as a result, the object is converted into a spacer with the correct length.
2. conversion of a uni-directional terminator into a bi-directional one, or vice versa.
3. inversion of a randomly chosen piece of the chromosome. If one of the end-points of the inverted piece is inside a promoter, a terminator or a gene, this object is destroyed. The object then splits in two pieces that each become a spacer. We note that multiple occurrences of random inversions result in translocation of chromosomal segments; we therefore do not need to explicitly include translocations in the model.
4. creation of a promoter or terminator: a piece of spacer is converted to a promoter or a terminator with a random orientation.

5.2.3 Kimura–Ohta population genetics

Our simulations mimic the population genetics of a population of prokaryotes with a population size N. We assume that various types of mutations (*e.g.* insertions, deletions, inversions) occur in each of the individuals in this population. According to Kimura–Ohta theory, eventually all mutants either go extinct or get fixed in the population (Kimura, 1962; Kimura and Ohta, 1969). The probability for a particular mutation i to become fixed depends on the selection coefficient s_i associated with the mutation[1] in the following way:

$$P_{\mathrm{K}}(s_i) = \frac{1-\mathrm{e}^{-2s_i}}{1-\mathrm{e}^{-2Ns_i}}. \tag{5.1}$$

The derivation of this equation was given in Section 1.2. In this derivation it is assumed that the mutation rates are low enough to assume that individual mutation events can be treated independently. Although this is generally not correct, Equation 5.1 has been used successfully in several studies (*e.g.* Mustonen and Lassig (2005)).

Since after each fixation event the population is clonal, one can represent the population by a single copy of the genome. Writing the rate of mutations of type m as μ_m, each particular mutation i of type m occurs in the population at

---

[1] If the fitness of the mutant is F_{mut} and the fitness of the rest of the population is F, then the selection coefficient s is defined as $s \equiv F_{\mathrm{mut}}/F - 1$.

the rate $\mu_m N$ and gets fixed with an effective rate $u_i \equiv \mu_m N P_{\mathrm{K}}(s_i)$. We have developed a novel and efficient simulation scheme that ensures that each mutation indeed occurs at this effective mutation-fixation rate (see 5.A).

5.2.4 Selection on genome length and gratuitous transcription

The fitness of the chromosome is defined in terms of the total length of the genome and the total amount of gratuitous mRNA produced by the genome. We assume that each promoter acts as the start of a TU. The TU ends with the first terminator that is found downstream of this promoter and resides on the same strand (or is bi-directional). This completely specifies the transcriptome of the model chromosome. Note that by this definition, TUs can partly overlap. Apart from genes, TUs may contain spacers, promoters and terminators (on the complementary strand); such pieces of DNA lead to gratuitous transcription. The total length of the gratuitous parts of TUs, summed over all TUs, is called S. The selection coefficient s associated with a certain mutation is then determined by the following rules:

- ✓ We assume that all genes are essential; this means that all genes need to stay intact and that they need to reside in some TU. If this is not the case, the organism is not viable ($s = -\infty$, $P_{\mathrm{K}}(s) = 0$). As a consequence of this rule, the sizes of genes are irrelevant in our model.
- ✓ Genes in one operon should have the same color (else: $s = -\infty$, $P_{\mathrm{K}}(s) = 0$);
- ✓ In all other cases, the selection coefficient is given by

$$s = -\alpha \Delta L - \beta \Delta S, \tag{5.2}$$

where ΔL and ΔS are the changes due to the mutation in, respectively, the length of the genome L and the gratuitous part of the transcriptome S. One can interpret α as the normalized fitness cost of increasing the length of the genome with one base pair, and β represents the normalized fitness cost of transcription per base pair.

5.3 Results

Fig. 5.2 displays the distance distributions for convergent, tandem and divergent gene pairs resulting from the simulations, juxtaposed with the real data from *E. coli* and *B. subtilis*. The similarities between data and theory are striking. Importantly, all distributions have an exponential tail; moreover, the convergent and the tandem distributions have a clear peak, both in our simulations and in the real data. This is discussed below.

The simulation results are a snapshot of the genome after 5.9×10^{10} generations, which at a generation time of 1 hour corresponds to approximately 7 million years. We note however that the dynamics of the simulation are sensitive to

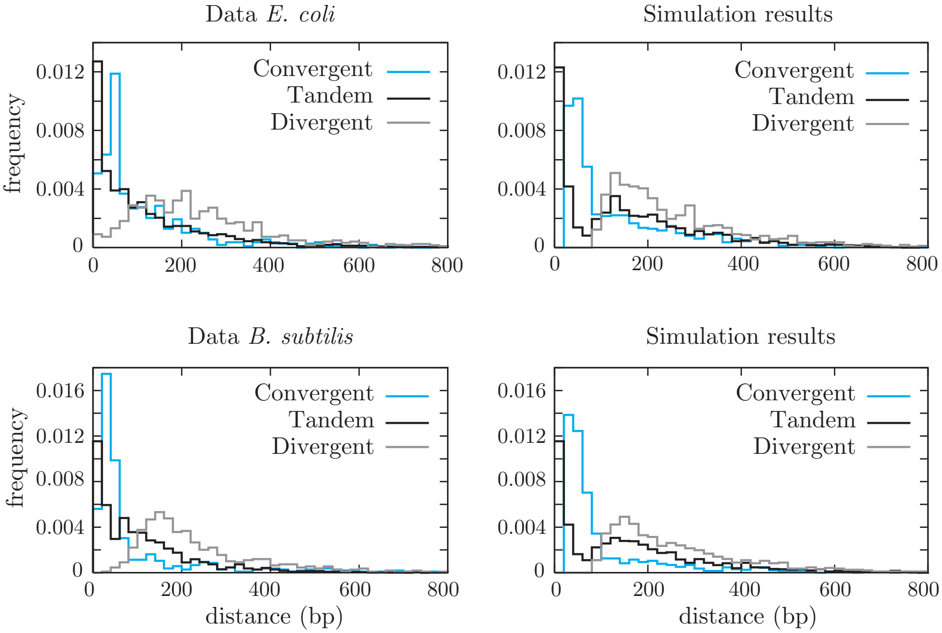


Figure 5.2: Simulation results compared to real data. The left figures show the frequency distribution of intergenic distances in *E. coli* and *B. subtilis* for convergent, divergent and tandem gene pairs. The right figures shows the same data for the simulated genomes after a simulation time of 5.9×10^{10} generations, which at a generation time of 1 hour corresponds to approximately 7 million years. The parameters of these simulations are discussed in 5.A and summarized in Table 5.1. The peak in the Tandem plot shows that operons have formed. Both the real data and the simulations show a peak in the convergent distribution at 20 to 60 bp. In the simulations, this peak is caused by bi-directional terminators. This supports our earlier hypothesis (Hermsen *et al.*, 2008)) that this is also the case in the real prokaryotes.

the choice of parameters. For instance, a tenfold increase in the number of gene colors would slow down the formation of operons considerably, whereas a higher value of α leads to a speed-up of this process, as it decreases the mean fitness cost of terminator loss. The evolutionary time should therefore be interpreted as an order of magnitude estimate.

5.3.1 Spontaneous formation of operons

The distance distribution of tandem gene pairs resulting from the simulations is bi-modal. This reveals that operons have formed: intergenic regions *within* operons are, on average, much shorter than those *between* operons, resulting in the bi-modal distribution. In the *B. subtilis* data this bi-modality is directly visible as well; in *E. coli*, a peak at short distances is clearly present, but the bi-modality is not as apparent. This may be a consequence of the fact that many operons in *E. coli* have internal regulatory sequences, resulting in a broader distribution of

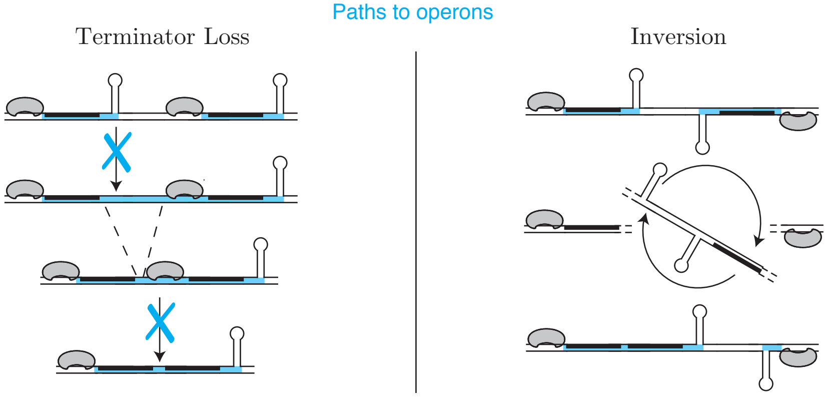


Figure 5.3: Paths to operons (see the legend in Fig. 5.1). In simulations of the model, operons emerge spontaneously. The most frequent pathway is through terminator loss (left). First a terminator between tandem genes is lost, leading to a poly-cistronic transcription unit (TU). The intergenic region is now being transcribed, resulting in an altered selection pressure on its length. Consequently, it is likely to shrink gradually. Finally, the internal promoter is lost. A second pathway is mediated by an inversion, which directly brings a gene under the control of another promoter (right).

distances between genes inside operons (Price *et al.*, 2006).

We have identified two important paths of operon formation occuring in our simulations (Fig. 5.3). In most cases, a terminator between two tandem genes is first destroyed by a mutation. The probability for this mutation to become fixed in the population depends on the length of the intergenic spacer; as a result, most newly-formed poly-cistronic TUs have relatively short intergenic spacers. Nevertheless, these lengths are typically larger than the average intergenic regions inside operons. However, after the terminator loss has occurred, the intergenic region is inside a TU. The selection pressure against gratuitous mRNA subsequently alters the balance between effective insertion and deletion rates, and tends to shorten the spacers further. Also, the internal promoter has become unnecessary; any mutation destroying it leads to a slight fitness increase. As a result, the promoter is removed and a genuine operon results. We stress that the initial loss of the terminator is a rare event; afterwards, the process proceeds “downhill”.

A second pathway is mediated by the inversion of a piece of DNA. As Fig. 5.3 shows, an inversion of a genome fragment can effectively move a set of genes to a different operon.

5.3.2 Bi-directional terminators

The distribution of the convergent pairs clearly shows a peak at distances around 40–60 bp, as in the real data of *E. coli* and *B. subtilis*. In the simulations, this is indicative of convergent pairs sharing a bi-directional terminator. The most

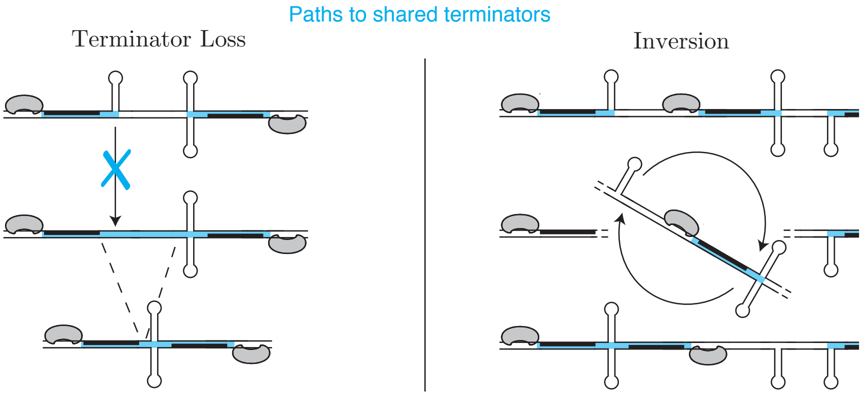


Figure 5.4: Paths to shared terminators (see the legend in Fig. 5.1). The paths leading to shared terminators in our simulations are similar to those leading to operons (see Fig. 5.3). In the most frequent pathway (left), terminator loss occurs first; next, transcription of the intergenic region leads to an altered selection pressure on the length of the intergenic region, which tends to shorten it gradually. An inversion can directly lead to a similar configuration (right).

likely process leading to this configuration is the following (see Fig. 5.4). The first step is again that one of the terminators in the intergenic region between a convergent pair is destroyed by a mutation. This happens very often, but usually such mutants are not fixed in the population. In some cases, however, the other terminator in the intergenic region can function bi-directionally. In this case, the only fitness cost of this mutation is due to the gratuitous transcription of the spacer that was originally between the terminators. If this spacer is not too long, the fitness burden is small and the mutation has a chance to become fixed in the population. After this, the insertion–deletion balance in the intergenic region is biased towards deletions and drives the two genes towards each other, effectively binding them together. Again, losing the first terminator is a rare event, but afterwards the process continues downhill. This is strong support for our earlier suggestion that the peak in the convergent distribution in *E. coli* is caused by bi-directional terminators (Hermsen *et al.*, 2008).

A second scenario is presented in Fig. 5.4; here, an inversion directly puts a gene, including a promoter, in front of a bi-directional terminator.

In *B. subtilis* the peak in the convergent distribution is higher than in *E. coli*. This may be explained by the fact that in *B. subtilis* Rho-dependent termination is hardly used (de Hoon *et al.*, 2005). Possibly, it therefore contains a higher fraction of intrinsic terminators, and hence has a higher tendency to form bi-directional terminators.

5.3.3 Spacing of genomic elements

Both in the real data and in the simulation results, the tail of the probability distributions of intergenic distances is exponential. In the model, this is due to the fact that insertions and deletions of single base pairs are by far the most frequent mutations. The lengths of intergenic regions are therefore largely determined by the equilibrium statistics induced by these moves only. This raises a picture in which big moves, such as inversions, promoter or terminator loss, translations and gene duplications can be considered rare events; in between such incidents the distances equilibrate with respect to the dynamics of the small mutations.

Assuming that this separation of time scales indeed holds, the following detailed balance condition applies to the equilibrium probability distribution $P_{\mathrm{u}}(l)$ of untranscribed spacers lengths:

$$\frac{P_{\mathrm{u}}(l)}{P_{\mathrm{u}}(l+1)} = \frac{\mu_{\mathrm{del}} P_{\mathrm{K}}(\alpha)}{\mu_{\mathrm{in}} P_{\mathrm{K}}(-\alpha)}. \tag{5.3}$$

Here μ_{in} and μ_{del} are the insertion and deletion rates per base pair, and P_{K} and α were introduced in Equations 5.1 and 5.2. Equation 5.3 shows that the equilibrium distribution is geometric and that its scale factor is determined by α, μ_{in} and μ_{del}. In a similar fashion, the equilibrium length distribution P_{t} of spacers that are being transcribed (from a single promoter) is given by

$$\frac{P_{\mathrm{t}}(l)}{P_{\mathrm{t}}(l+1)} = \frac{\mu_{\mathrm{del}} P_{\mathrm{K}}(\alpha+\beta)}{\mu_{\mathrm{in}} P_{\mathrm{K}}(-\alpha-\beta)}. \tag{5.4}$$

This equation holds for intergenic regions inside operons, but also for $5'$ and $3'$ untranslated regions (*i.e.* spacers that are located between a promoter and the first downstream gene and spacers between a terminator and the first gene upstream). The model therefore shows that a very simple fitness function that only takes into account the cost of transcription and genome size can unify a number of observations that are usually taken for granted: the facts that genes in operons are generally closer together than genes on the border of operons and that both promoters and terminators typically tightly flank the genes they belong to. As we discussed, the same mechanism also provides a plausible explanation for the observed peak in the distribution of convergent gene pairs.

If the mean length of un-transcribed spacers is $\bar{l}$ and the mean of the transcribed ones is $\bar{m}$, then Equations 5.1, 5.3 and 5.4 imply the following relations with the fitness costs (see Appendix 5.B):

$$\alpha = \frac{1}{2N}\left(\frac{1}{\bar{l}} + \ln\left(\frac{\mu_{\mathrm{in}}}{\mu_{\mathrm{del}}}\right)\right), \tag{5.5}$$

$$\alpha+\beta = \frac{1}{2N}\left(\frac{1}{\bar{m}} + \ln\left(\frac{\mu_{\mathrm{in}}}{\mu_{\mathrm{del}}}\right)\right). \tag{5.6}$$

These equations demonstrate that the value of α cannot be determined from the mean distance $\bar{l}$ without knowledge of μ_{del} and μ_{in}. In order to fit the model value of $\bar{l}$ to the genomic data, one can choose any value of $\alpha \geq 0$ and then adjust $\mu_{\mathrm{in}}/\mu_{\mathrm{del}}$ or vice versa. In our simulations, we chose $\mu_{\mathrm{in}}/\mu_{\mathrm{del}} = 1$ and adjusted α to fit the value of $\bar{l}$ of *E. coli*.

The value of β can be obtained by subtracting Equations 5.5 and 5.6. Estimating that $\bar{m} \approx 13\,\mathrm{bp}$ and $\bar{l} \approx 125\,\mathrm{bp}$ in both *E. coli* and *B. subtilis*,

$$\beta = \frac{1}{2N}\left(\frac{1}{\bar{m}} - \frac{1}{\bar{l}}\right) \approx \frac{1}{25N}. \tag{5.7}$$

This provides an estimate for the fitness cost of gratuitous transcription. The observed distance distributions are consistent with a weak to moderate selection on transcript lengths. Assuming a population size of at least $N = 10^5$, β is of the order of 10^{-6}. Unfortunately, this is too small to be measured directly in competition experiments. We do note that the estimate of β is sensitive to some simplifications of our model, such as our assumption that insertions and deletions always span 1 bp.

The idea that the close spacing of genes in operons is due to the cost of producing gratuitous mRNA leads to the prediction that highly-transcribed operons have a closer spacing. This was studied by Price *et al.* (2006) and turns out to be incorrect for *E. coli* due to the presence of internal promoters in highly-expressed TUs. As we mentioned, internal promoters are never beneficial in our simplified model; therefore, they are quickly removed by random mutations. In reality, internal promoters clearly do play a role in regulatory fine-tuning.

In *E. coli* several gene pairs inside operons overlap slightly (Fukuda *et al.*, 2003). Often the third position of the stop codon of the first gene (`UAA` or `UGA` in these cases) overlaps with the first position of the start codon of the second gene (`AUG`). This configuration might have a functional reason due to translational coupling (Fukuda *et al.*, 2003). For simplicity, we did not include such advanced mechanisms in our model. Similarly, the data show that the divergent pairs in *E. coli* and *B. subtilis* can be very close together ($< 80\,\mathrm{bp}$); this is not possible in our model, since divergent pairs are separated by at least twice the size of a core promoter. In reality, the promoters of these gene pairs may in rare cases overlap with each other or with the upstream open reading frame (Warren and ten Wolde, 2004b); for instance, in *E. coli* the promoter relBp overlaps with ORF *ydfV*, allowing for an unusually close spacing between these divergent genes (see: EcoCyc database (Keseler *et al.*, 2005)).

5.4 Discussion

We acknowledge that our simple model can not account for all known operon histories. For instance, newly formed operons are strongly enriched in so-called ORFan genes (Price *et al.*, 2005b). ORFan genes are genes that lack identifiable homologs outside of a group of closely related bacteria. This shows that innovations introduced for instance by phages probably play an essential role in many cases of operon genesis. Also, it has been shown that seemingly unlikely displacements of genes in operons by horizontal gene transfer do occur (Omelchenko *et al.*, 2003). These and other observations suggest that many different mechanisms contribute to some extent to the formation of operons. That being said, terminator loss is a plausible first step in the fusion of two *neighboring* TUs.

Our model assumes that at least some terminators are under weak selection; is this likely? Upon terminator loss, read-through leads to costly elongated transcripts. If the first downstream TU is on the opposite strand, it can also give rise to harmful antisense transcripts. Otherwise, if the downstream TU is transcribed from the same strand, read-through is likely to alter its expression profile due to co-transcription. Moreover, in both cases transcriptional interference should be anticipated (Shearwin *et al.*, 2005). It is therefore likely that many terminators are under strong selective pressure and will not readily be lost. However, several scenarios can be envisioned in which the selection pressure on a terminator is strongly relaxed. First, in environments where the TU is repressed the terminator is not being used, which should lead to a strongly reduced purifying selection. Second, in some cases the sole effect of the co-transcription of a tandem downstream gene is an increase in the expression level of that gene; the resulting fitness effect then depends strongly on the current environment and can be weak. Moreover, such an effect can be transient, as subsequent mutations can rapidly compensate for the altered expression level. Third, if directly downstream of the TU a terminator is present that could function bi-directionally, this could alleviate the selection pressure on its terminator. Such scenarios indicate that the mechanism proposed here can indeed occur in real systems. Terminator loss is not necessarily terminal.

In this light, one might be surprised about the scarcity of operons in most eukaryotic genomes. One could argue that in eukaryotic genomes, the loss of a terminator sequence would also result in poly-cistronic RNAs; why does this not lead to operons in eukaryotes? As it turns out, the process of *translation* provides a natural barrier (Lawrence, 1999). In prokaryotes, ORFs are typically preceded by a ribosomal binding site (the Shine–Dalgarno sequence, after Shine and Dalgarno (1975)). The presence of this sequence directly upstream of a start codon is sufficient to induce the assembly of a ribosome on the start codon, irrespective of the location of the ORF on the mRNA (Alberts *et al.*, 1994). This means that, if two TUs merge as a result of terminator loss, the standard translation machinery is directly able to translate all ORFs on the resulting poly-cistronic mRNA. In contrast, in eukaryotes only the first ORF is translated by default (Blumenthal, 2004). This is dictated by the standard translation mechanism. Initially, the small ribosomal subunit binds to the $5'$ end of the mRNA; it then starts moving in the $5'$ to $3'$ direction until it finds the first start codon. There the ribosome assembles and starts the translation process. When it encounters the stop codon, the ribosome disassembles and thus ignores the other ORFs on the mRNA. Poly-cistronic mRNAs therefore require specific elements, called internal ribosome entry sites (IRES), in order to enable the translation of the consecutive genes (Blumenthal, 2004). Alternatively, poly-cistronic pre-mRNA can be trans-spliced to create mono-cistronic mRNAs, as is common in *Caenorhabditis elegans* (Blumenthal *et al.*, 2002). Again specific sequences are required to guide this process (Graber *et al.*, 2007). The bottom line is that, in eukaryotes, loss-of-function mutations in terminator sequences typically do not result in functional poly-cistronic mRNAs, whereas in prokayotes they do.

Our model assumes that a small fitness cost is associated with gratuitous DNA and gratuitous RNA. The showed that cost of gratuitous DNA can not be computed from the distance distributions since the ratio $\mu_{\mathrm{in}}/\mu_{\mathrm{del}}$ is unknown, but the cost of gratuitous RNA *can* be deduced within the context of the model. We estimated that the fitness cost per base pair is very small. Nevertheless, it may be possible to measure an effect on the bacterial growth speed in experimental conditions if a large stretch of nucleotides is inserted in a transcribed region.

Evidence for the scenario of operon formation proposed here has been described by Price *et al.* (2006), who have compared the gene order of *E. coli* to its relatives to identify recently destroyed or formed operons. Loss of a terminator will result in the merging of TUs that were already adjacent. Indeed, Price et al. conclude that new operons often comprise functionally unrelated genes that were already in proximity before the operon formed. Also, they find that modifications of existing operons are often the result of merging, appending or prepending processes; insertions of genes *inside* existing operons are more rare. This is consistent with the terminator loss mechanism.

To obtain additional evidence, more direct phylogenetic tests of the terminator loss scenario will have to be done. In Chapter 6 we will try to do so. If terminator loss occurs at a sufficient rate, the model predicts the existence of pairs of neighboring, tandem TUs that are independently transcribed in several related clades or species but have merged in one (or several) of them. Currently it is not straightforward to find such examples, as very little independent information is available about the extent of TUs in prokaryotic species other than *E. coli*. Hopefully, such obstacles will be resolved in the near future by additional genomic information.

5.A Methods

5.A.1 Simulating population genetics

According to the Kimura–Ohta theory, a mutation that occurs with a rate μ in each individual in the population, becomes fixed in the population at a rate $\mu N P_{\mathrm{K}}(s)$, where $P_{\mathrm{K}}(s)$ is the mutation-fixation probability if a mutation with selection coefficient s, and N is the population size (see Section 1.2.4). Our goal is to construct a simulation method that ensures that all possible mutations occur at the correct mutation-fixation rate.

Our simulation scheme is an unusual Monte Carlo method. In Monte Carlo schemes, trial moves (mutations) are performed and subsequently accepted or rejected according to some acceptance rule that typically depends on the energy change due to this move. The acceptance probability P_{acc} is usually based on the Bolzmann factors of the old and the new state. In our scheme, we also perform trial mutations, but now the acceptance probability depends on the selection coefficient s corresponding to the mutation.

A straightforward way to arrive at the correct effective mutation-fixation rates would be to try each mutation i of type m at a rate $\mu_m N$ (the rate at which the mutation occurs in the total population) and then accept it with probability $P_{\mathrm{acc}}(s_i) = P_{\mathrm{K}}(s_i)$ (the probability of fixation). This would indeed lead to the correct mutation-fixation rate $\mu_m N P_{\mathrm{K}}(s_i)$. This method, however, would be extremely inefficient for the following reason. Most mutations in the model are either nearly neutral or strongly deleterious ($s \lesssim 1/N$), as is also expected for real molecular systems (Kimura, 1979; Gillespie, 1991). For such mutations, the acceptance probability $P_{\mathrm{K}}(s)$ is very small ($\lesssim 1/N$) so that practically all trial moves would be rejected.

The following scheme solves that problem. For every type of mutation m, we selected a value $s_{\max,m}$ such that for mutations i of type m the probability that $s_i > s_{\max,m}$ is negligible. Next, we assigned a *decreased* trial rate $\mu_m N P_{\mathrm{K}}(s_{\max,m})$ to trial moves of type m and accepted them with an *increased* probability

$$P_{\mathrm{acc},m}(s_i) \equiv \frac{P_{\mathrm{K}}(s_i)}{P_{\mathrm{K}}(s_{\max,m})}. \tag{A5.1}$$

This results in the correct effective rate for each mutation i of type m with selection coefficient s_i, since

$$\mu_m N P_{\mathrm{K}}(s_{\max,m}) \times P_{\mathrm{acc},m}(s_i) = \mu_m N P_{\mathrm{K}}(s_i). \tag{A5.2}$$

By choosing the values $s_{\max,m}$ as low as possible, a speed-up factor of $\approx 10^4$ was obtained. During the simulation, we monitored that the requirement $s_i \leq s_{\max,m}$ was never violated.

To perform our trial mutations, we use an event-driven algorithm. This means that, in each step of the algorithm, we make an inventory of all possible mutations

quantity	symbol	value
promoter length	π	40 bp
terminator length	τ	23 bp
number of genes	n	3454
number of colors	c	3
population size	N	10^5
rate insertion	μ_{in}	$1 \times 10^{-9}\,\mathrm{bp}^{-1}$
rate deletion	μ_{del}	$1 \times 10^{-9}\,\mathrm{bp}^{-1}$
rate creation terminator	μ_{ct}	1×10^{-12} per site
rate creation promoter	μ_{cp}	1×10^{-12} per site
rate uni- to bi-directional	μ_{ub}	5×10^{-11} per terminator
rate bi- to uni-directional	μ_{bu}	2.45×10^{-9} per terminator
rate inversion	μ_{iv}	$1 \times 10^{-12}\,\mathrm{bp}^{-1}$
fitness cost DNA length	α	$4 \times 10^{-8}\,\mathrm{bp}^{-1}$
fitness cost transcription	β	$3.6 \times 10^{-7}\,\mathrm{bp}^{-1}$

Table 5.1: Parameter settings for the simulations. Shown are the parameters belonging to the simulation results for *E. coli* presented in Fig. 5.2. The same parameters were also used to fit the *B. subtilis* data; the only difference is that in these simulations μ_{ub} is chosen $\approx 12\times$ higher, leading to 10 times more bi-directional terminators.

and their rates. The sum of these rates determines the (Poissonian) probability distribution of next-mutation times, which we use to stochastically determine when the next mutation occurs in the population. Next we decide *which* particular mutation to perform; the probability for a particular mutation to be chosen is proportional to its rate.

5.A.2 Parameters

The parameters for the results presented in Fig. 5.2 are listed in Table 5.1 and were based on the following considerations. Effective population sizes estimates range from 10^5 to 10^8 (Berg, 1996); we used 10^5. The promoter length was chosen to be typical for the core promoter in *E. coli*: 40 bp. We used EcoCyc database (Keseler *et al.*, 2005) to compute the mean length of known Rho-independent terminator sequences in *E. coli*, and used the result ($\tau = 23$ bp) for our simulations. The insertion and deletion rates are estimated to be comparable to base pair substitution rates (Berg and Kurland, 2002) which in *E. coli* is of the order of $\approx 1 \times 10^{-9}$ per replication (Drake *et al.*, 1998). The creation of terminators has to be a rare event, since otherwise intergenic regions would become littered with gratuitous terminators. We therefore use a considerably lower rate for terminator creation. For simplicity, the promoter creation rate is assumed to be equal to the terminator creation rate. We assume that only a small fraction (2%) of the terminators that are used uni-directionally could actually function bi-directionally. Therefore the conversion from uni- to bi-directional is much slower than the reverse process. Presumably, the latter could result from mutations in several places in the terminator, which suggests that this rate should be a few times higher than the point mutation rate. Although inversions and translocations have occurred regularly in the evolution of bacteria (Itoh *et al.*, 1999; Zivanovic *et al.*, 2002), we expect these to happen considerably less often than small insertions and deletions. As we explained in Section 5.4, the values of α and β tune the average length of intergenic regions; we chose these values such as to fit the data of *E. coli*. To fit the plots for *B. subtilis*, the same parameter values were used as for *E. coli*, except for an increased fraction of promoters that can function bi-directionally: 20%.

5.A.3 Initialization

Initially, the chromosome was prepared as follows. The genome contained 3553 genes in random orientations. Each gene has its own promoter and its own terminator (no operons). The lengths of the spacers are prepared according to the equilibrium probability distribution that would apply if the dynamics would be completely dominated by insertions and deletions only (see Section 5.4). In this equilibrium, the distribution of spacer lengths is shorter for spacers that are transcribed than for spacers that are not, even though both distributions are geometric.

5.B Derivation of Equation 5.5

The mean length of the spacer regions is denoted by $\bar{l}$. In order to achieve this mean value in our model, the probability distribution $P(l)$ of the length of intergenic regions should obey:

$$\frac{P(l)}{P(l+1)} = \mathrm{e}^{1/\bar{l}}. \tag{A5.3}$$

At the same time, equilibrium dictates that

$$P(l)\mu_{\mathrm{in}}(l+1)P_{\mathrm{K}}(-\alpha) = P(l+1)\mu_{\mathrm{del}}(l+1)P_{\mathrm{K}}(\alpha). \tag{A5.4}$$

Here we used that, in a spacer of length l, there are $l+1$ places where one can insert a base pair, and that the fitness cost of an insertion or a deletion are $s = \alpha$ and $s = -\alpha$ respectively. The rates μ_{del} and μ_{in} are the deletion and insertion rates. It follows that:

$$\frac{P(l)}{P(l+1)} = \frac{\mu_{\mathrm{del}}P_{\mathrm{K}}(\alpha)}{\mu_{\mathrm{in}}P_{\mathrm{K}}(-\alpha)}. \tag{A5.5}$$

We now use Equation 1.45 and anticipate that $\alpha \ll 1$ (which we can check at the end of the derivation), and obtain

$$-\frac{\mu_{\mathrm{in}}}{\mu_{\mathrm{del}}}\mathrm{e}^{1/\bar{l}} = \frac{1-\mathrm{e}^{-2N\alpha}}{1-\mathrm{e}^{2N\alpha}}. \tag{A5.6}$$

If we define $x \equiv \mathrm{e}^{2N\alpha}$, the previous equation becomes quadratic in x; solving this equation leads to

$$\alpha = \frac{1}{2N}\left(\frac{1}{\bar{l}} + \ln\left(\frac{\mu_{\mathrm{del}}}{\mu_{\mathrm{in}}}\right)\right). \tag{A5.7}$$

This shows that, if we fix the mean $\bar{l}$, the fitness cost of one base pair, α, depends on the population size. This is not surprising, as we saw that the consequence of fitness differences is proportional to the population size. We can also see that α depends on the ratio of the insertion and deletion rate. As we do not know this a priori ratio, we choose them equal in our model; in this case

$$\alpha = \frac{1}{2N\bar{l}}. \tag{A5.8}$$

In reality, $\bar{l} \approx 125\,\mathrm{bp}$ and $N \approx 10^5$. This confirms that, indeed, α is a very small fitness cost, in retrospect justifying our assumption earlier on.

Chapter 6

The evolutionary dynamics of intergenic distances

Intergenic distances change due to insertions and deletions. As these mutations occur stochastically, the lengths of intergenic regions carry out a "random walk" on long time scales. Here, we study this "diffusive" behavior by comparing the distances between neighboring genes in *E. coli* with the distances between their orthologs in related species. These data can be compared to a formal model based on a Master equation. We show that the divergence of the lengths of intergenic regions in *Escherichia coli* and *Salmonella Typhi* is compatible with our model, but that insertions and deletions larger than one base pair cannot be ignored, as they contribute strongly to the speed of the divergence.

We also use the model to identify operons that may recently have formed due to the merging of two transcription units.

6.1 Introduction

On evolutionary time scales, the lengths of intergenic regions constantly change. The static distributions of intergenic distances that we studied in Chapter 4 are a consequence of these evolutionary dynamics. Indeed, in Chapter 5 we demonstrated that the distance distributions of bacteria such as *Escherichia coli* and *Bacillus subtilis* can be reproduced by a strongly simplified model of genome evolution. Here we take a closer look at the evolution of the lengths of intergenic regions. We compare the distances between genes in *E. coli* to the distances between their orthologs in two close relatives: *Shigella dysenteriae* and *Salmonella enterica subsp. enterica serovar Typhi* (abbreviated below as *Salmonella Typhi* or *S. Typhi*). This allows us to study how intergenic distances of related species diverge in the course of time.

Intergenic regions grow and shrink due to insertions and deletions (jointly referred to as indels). In intergenic regions, these indels are typically only a few base pairs long[1]. As the occurrence of mutations is a stochastic process, one would expect that the lengths of intergenic regions perform a "random walk" on evolutionary time scales. We propose a stochastic model for the evolutionary "diffusion" of intergenic regions and compare it to the divergence between *E. coli* and *S. Typhi*.

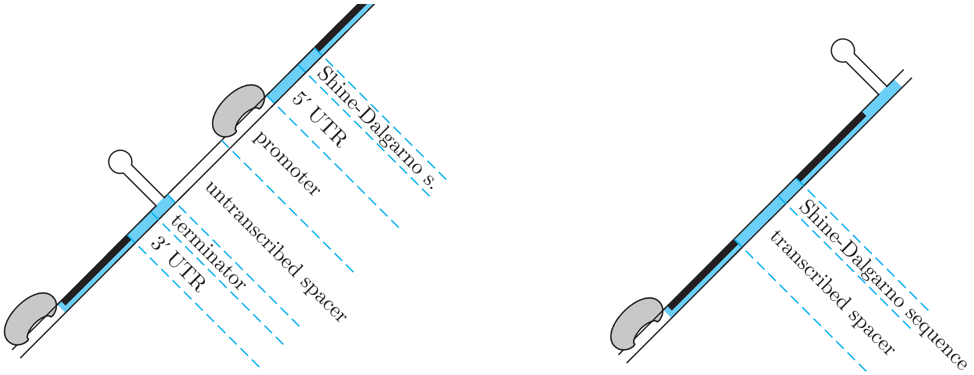
As a first project, we focus on intergenic regions between tandem genes. This has several reasons. First, tandem intergenic regions are particularly interesting since they come in two flavors: those *inside* transcription units (TUs) and those *between* them[2]. In Chapter 4 we demonstrated that the tandem intergenic regions inside TUs (called *intra*-operon regions) are typically much shorter than the ones between TUs (*inter*-operon regions). These differences in the mean length have to reflect differences in effective mutation-fixation rates of indels. Therefore the evolutionary "diffusion" of the lengths of intergenic regions should also be different for the two types.

A second reason to focus on tandem intergenic regions has to do with our predictions in Chapter 5. There, we suggested that tandem neighboring operons could merge due to the loss of a terminator sequence. If this happens, the intergenic region between the operons changes from type "inter" to type "intra". Conversely, if an operon would split in two parts, an intergenic region has to switch from type "intra" to type "inter". In both cases, the mode of diffusion changes after switching. We model the influence of switching on the evolution of the length of intergenic regions and compare the results with the data of *E. coli* and *S. Typhi* to find operons that may have formed recently by a merging or splitting event.

---

[1] The analysis of Messer and Arndt (2007) of indels in primates suggests that about one half of the indels is only one nucleotide long; yet, rates can differ strongly between different organisms.

[2] We acknowledge that the border between these two categories is fuzzy: in the presence of internal promoters and weak terminators some intergenic regions may be hard to classify. In the present analysis we ignore these exceptions.



(a) Intergenic region between transcription units (type “inter”).

(b) Intergenic region inside a transcription unit (type “intra”).

Figure 6.1: Intergenic regions between tandem genes exist in two types: those *between* transcription units (a) and those *inside* a transcription unit (b). Intergenic regions of different types have different compositions; for instance, inter-operon regions contain a promoter and a terminator, whereas intra-operon regions do not. The blue parts of the DNA denote parts that are being transcribed; see the legend in Fig. 5.1 for the meaning of the other symbols.

6.2 Short description of the model

In this section, we give a short description of our model. The model describes how the lengths of intergenic regions evolve due to random insertions and deletions. We cannot *directly* compare these dynamics to real data of *E. coli*, because we know only the current lengths of its intergenic regions and not how long these were in the past. This problem can be circumvented by comparing the lengths of the intergenic regions in *E. coli* with those in related species, such as *S. Typhi*. The idea is to use our model to predict the joint probability that a given intergenic region has length n in *E. coli* and length m in *S. Typhi*. This probability distribution *can* be compared to the data.

The probability that a given intergenic region has length n in *E. coli* and length m in *S. Typhi* is different for the different types of intergenic regions (intra- or inter-operon ones). Moreover, it is again different for intergenic regions in which a merging or splitting event has taken place in one of the species. The total joint probability function $\mathrm{J}_{mn}^{\mathrm{tot}}$ can therefore be written as

$$\mathrm{J}_{mn}^{\mathrm{tot}} = w_1 \mathrm{J}_{mn}^{\mathrm{intra}} + w_2 \mathrm{J}_{mn}^{\mathrm{inter}} + w_3 \mathrm{J}_{mn}^{\mathrm{merge}} + w_4 \mathrm{J}_{mn}^{\mathrm{split}}, \tag{6.1}$$

in which $\sum_i w_i = 1$. The four terms correspond to the contributions of intra-operon intergenic regions, inter-operon regions, those in which a merging event has occurred, and those in which a splitting event has taken place. In writing Equation 6.1, we implicitly assumed that the rates at which merging and splitting events take place are low, so that we can neglect histories in which, in a given intergenic region, more than one such event has occurred since the divergence of the two species from their common ancestor.

Below, we briefly describe our model and how we compute the quantities in 6.1; we refer to Appendix 6.B for more detailed derivations. The spirit of the calculations is similar to the method presented by Mustonen and Lassig (2005), who studied the evolution of the binding energies of transcription factor binding sites; however, the dynamics of the two systems is clearly different.

6.2.1 Model assumptions

As we mentioned, we distinguish two types of tandem intergenic regions, referred to as types “intra” and “inter”. These intergenic regions consist of various parts, such as promoters or untranslated regions (UTRs). Intergenic regions of different types have different compositions; this is illustrated in Fig. 6.1. Intergenic regions of type “inter” contain a terminator and a promoter. Apart from these objects, three “spacers” are present: one 3’ UTR, one 5’ UTR, and an untranscribed spacer. Also, the downstream gene is preceded by a Shine–Dalgarno sequence or ribosomal binding site. (Note that, in our definition, the 5’ UTR does *not* include the Shine–Dalgarno sequence.) Intergenic regions of type “intra”, on the other hand, do not contain terminators or promoters; we assume that they consist of a transcribed spacer and a Shine–Dalgarno sequence.

An important (but reasonable) assumption is that the lengths of the promoter, terminator and Shine–Dalgarno sequences are fixed. This means that only the spacer lengths evolve — with this we mean the UTRs, the transcribed spacers and the untranscribed spacers. The evolution of an intergenic region is then given by the independent evolution of its spacers. Therefore it is sufficient for our model to describe the dynamics of the different spacers.

The dynamics of the three kinds of spacers — transcribed spacers, untranscribed spacers and UTRs — are not the same. This is clear from the fact that their average lengths are different: in *E. coli*, the average length of untranscribed spacers is of the order of 125 bp, whereas UTRs and transcribed spacers are about an order of magnitude shorter. This means that the balance between insertions and deletions must be different for the various kinds of spacers. It is important to appreciate that the relevant indel rates are the combined result of mutation *and* selection. In fact, it is plausible that the mutation rates *before* selection are equal in each type of spacer, and that the length differences reflect the dissimilarities in the selection pressures acting on the different spacers. (See Section 1.2.4 of the Introduction, Section 5.3.3 of Chapter 5 and Appendix 6.C.)

Even though the typical lengths of the different types of spacers are not the same, we use one and the same model to describe their dynamics, be it with different parameter values. The model is discussed below, but can be summarized as follows. First: as most indels are very short, we make the convenient assumption that all insertions and deletion have a length of 1 bp only. In the Results section we examine the consequences of this assumption in some detail. Second: we assume that the rates of insertions and deletions in a given spacer are equal at all possible sites. These two assumptions lead to the theory described below.

6.2.2 Master equation for spacers

In a spacer of length n, a deletion of length 1 bp can occur at n places; therefore the total deletion rate should be proportional to n. Insertions can occur at $n+1$ places, so that the total insertion rate should be proportional to $n+1$. Consequently, longer intergenic regions are expected to "move" faster. These considerations can be formalized in the following Master equation for the probability $P(n,t)$ that a given spacer has length n at time t:

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = n\,a\,P(n-1,t) + (n+1)b\,P(n+1,t) - ((n+1)a + nb)P(n,t). \tag{6.2}$$

Here a is the insertion rate per possible insertion point in the intergenic region, and b is the deletion rate for any given base pair. We stress that a and b are mutation-fixation rates — with this we mean rates at which mutations are fixed in the population (see Section 1.2.4). The dynamics described by this Master equation obey detailed balance, which is important in the derivations below.

If $b > a$, the equation has the equilibrium distribution

$$P_{\mathrm{eq}}(n) = \left(1 - \frac{a}{b}\right) \mathrm{e}^{-\ln(b/a)n}. \tag{6.3}$$

The mean of this geometric distribution is $(b/a-1)^{-1}$ and is fixed by the ratio b/a. Therefore b/a should be chosen appropriately for the three different kinds of spacers (see Appendix 6.C).

6.2.3 The joint probabilities

From Equation 6.2, it is possible to compute the probability $\mathrm{G}_{mo}(t)$ that a spacer of initial length o evolves to a length m in a time t. This amounts to solving the Master equation (numerically) for the initial condition $P(n,0) = \delta_{no}$ (where δ_{no} is the Kronecker delta function).

We assume that the two species had their last common ancestor a time t_{a} ago. Suppose that, in this common ancestor, a particular spacer had an (unknown) length n_{a}. The probability that this spacer now has length n in species 1 and length m in species 2 is then given by the product $\mathrm{G}_{nn_{\mathrm{a}}}(t_{\mathrm{a}})\mathrm{G}_{mn_{\mathrm{a}}}(t_{\mathrm{a}})$. As the length n_{a} is unknown, the joint probability for a randomly selected spacer to have length n in species 1 and m in species 2, $\mathrm{J}_{nm}(t_{\mathrm{a}})$, can be computed by summing over all possible values of n_{a}:

$$\mathrm{J}_{nm}(t_{\mathrm{a}}) = \sum_{n_{\mathrm{a}}} \mathrm{G}_{nn_{\mathrm{a}}}(t_{\mathrm{a}})\mathrm{G}_{mn_{\mathrm{a}}}(t_{\mathrm{a}})P_{\mathrm{a}}(n_{\mathrm{a}}). \tag{6.4}$$

Here $P_{\mathrm{a}}(n_{\mathrm{a}})$ is the probability that the spacer had length n_{a} in the ancestor. If we make the assumption that the length distribution of spacers in the common ancestor was in equilibrium, then it follows that $P_{\mathrm{a}}(n_{\mathrm{a}}) = P_{\mathrm{eq}}(n_{\mathrm{a}})$. Remember that $P_{\mathrm{eq}}(n_{\mathrm{a}})$ was defined in Equation 6.3. This enables us to simplify Equation 6.4 and to compute $\mathrm{J}_{nm}(t_{\mathrm{a}})$ explicitly for given values of a, b and t_{a}.

Once we have computed the joint probabilities $\mathrm{J}_{nm}(t_{\mathrm{a}})$ for each type of *spacer*, we can also compute the probability for a *complete intergenic region* to have a length n_1 in species 1 and a length n_2 in species 2. The result is different for the different types of intergenic regions ($\mathrm{J}_{mn}^{\mathrm{intra}}$ for type "intra" and $\mathrm{J}_{mn}^{\mathrm{inter}}$ for type "inter"; see Equation 6.1). The calculation involves a suitable convolution of the probabilities for the different spacers in the intergenic region; in addition, the "static" objects such as the promoters and terminators should be taken into account. We refer to Appendix 6.B for the details of these calculations.

Up to now, we have implicitly assumed that no merging or splitting event occurred during the evolution of the intergenic region. We now turn to the treatment of the merging and splitting processes.

6.2.4 Merging and splitting

If two neighboring operons merge, the inter-operon intergenic region between the operons converts to type "intra". If an operon splits in two parts, the opposite happens. As we indicated, we assume that the rate at which merging and splitting events occur is low, so that we can ignore the possibility that more than one of these events has happened in a given intergenic region since the divergence of the two species from their common ancestor. Also, for a given intergenic region, we only consider histories in which a merging or splitting event occurred in only one of the species — not in both.

When computing the joint probability for the splitting and merging modes (*i.e.* $\mathrm{J}_{mn}^{\mathrm{merge}}$ and $\mathrm{J}_{mn}^{\mathrm{split}}$), we have to sum over the probabilities of all possible histories. If an intergenic region is of type "intra" in species 1 and of type "inter" in species 2, this can be explained by two scenarios: either a splitting event has occurred in species 2, or a merging has happened in species 1. Therefore, both scenarios contribute to the total probability. Also, the splitting or merging could have happened at any time between the divergence from the common ancestor (a time t_{a} ago) and the present. Therefore, we also have to integrate over all possible splitting or merging times. For our model, this integral can be computed explicitly (see Appendix 6.B).

A fundamental question is what happens to the lengths of the spacers at the moment of the splitting or merging event. Many scenarios can be envisioned. In case of operon merging, we proposed in Chapter 5 that the loss of a terminator is a likely first step. Later, the promoter may be lost (or not: many operons have internal promoters). Alternatively, a large deletion could remove both the terminator and the promoter at the same time. Our model assumes that, at the moment of merging, both the promoter and the terminator are destroyed, but their length is preserved. In other words, the total length of an intergenic region does not change at the moment of a merging event. Of course, the mutation-fixation rates of insertions and deletions do change upon merging; the model takes this into account.

In case of operon splitting, again several scenarios can be imagined. One possibility is that a promoter and terminator emerge from pre-existing sequences due to base pair substitutions and other small mutations. Another scenario would be that a promoter and/or a terminator is inserted as a whole, for instance due to a recombination event with other parts of the DNA. In the first scenario, the length of the intergenic region directly after splitting is equal to the length just before it. This requires that the length of the intergenic region before splitting was long enough to accommodate a terminator and a promoter. In the second scenario, the length of the intergenic region makes a sudden jump at the moment of splitting, as a promoter and terminator are inserted.

In fact, many hybrid scenarios can be envisioned. Maybe, first internal regulation evolves. New promoters or terminators may initially overlap with one of the genes; both may initially be weak. Clearly, our model is too coarse-grained to describe all these options. In the results presented below, we assume that at the moment of splitting a promoter and a terminator are inserted in the intergenic region; the length of the region then increases instantaneously.

6.3 Results

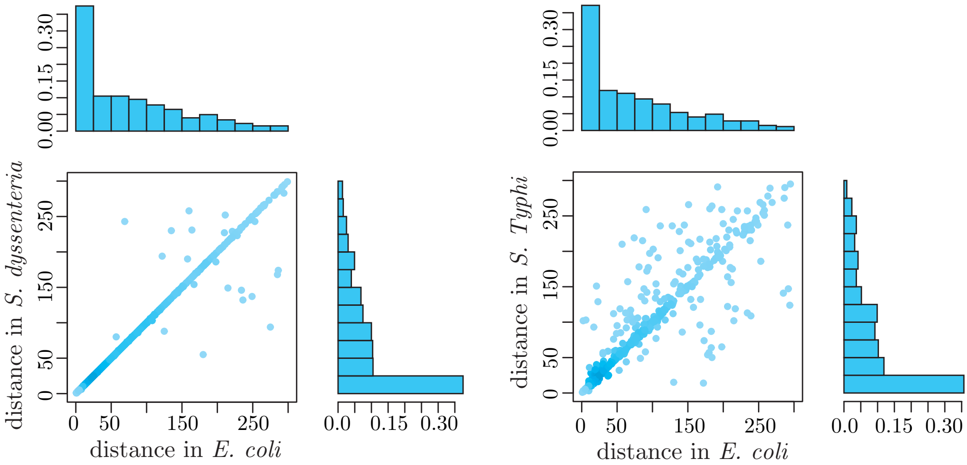
We now present the results of our analysis.

6.3.1 The data: *E. coli* versus *S. dysenteriae* and *S. Typhi*

We compare the lengths of intergenic regions in *E. coli* to the lengths of their orthologous intergenic region in *S. dysenteriae* and *S. Typhi* (see Appendix 6.A for details on the selection of the orthologs). In Fig. 6.2, the resulting data are represented in scatterplots.

Fig. 6.2(a) shows the data for *Escherichia coli* and *Shigella dysenteriae*. The bacterium *S. dysenteriae* is very closely related to *E. coli*; many authors even consider the genus Shigella as a part of *E. coli* (Pupo *et al.*, 1997). One would therefore expect that most intergenic distances in *E. coli* should be very similar to those in *S. dysenteriae*. Indeed, almost all points fall on the main diagonal of the scatterplot (the plot contains 832 points, so the outliers really are exceptions).

The species *Salmonella enterica subsp. enterica serovar Typhi* and *E. coli* both belong to the family of enterobacteriaceae, but not to the same genus. This means that the species are quite related, but much less so than *E. coli* and *S. dysenteriae*. *S. Typhi* diverged from the *E. coli* lineage about 100 million years ago (Lawrence and Ochman, 1998). Fig. 6.2(b) shows the scatterplot for *S. Typhi* vs. *E. coli* (651 points). As expected, this plot shows much less correlation between the distances than the plot for *S. dysenteriae*. Due to insertions and deletions in the intergenic



(a) *Shigella dyssenteria* vs. *E. coli* (b) *Salmonella Typhi* vs. *E. coli*

Figure 6.2: Scatterplots of the length of tandem intergenic regions in *E. coli* compared to their length in *S. dysenteriae* (Fig. (a)) and *S. Typhi* (Fig. (b)). As *S. dysenteriae* is very closely related to *E. coli*, almost all intergenic regions have a similar length in both organisms, as apparent from the concentration of points on the diagonal of the plot. *S. Typhi* is a more distant relative of *E. coli*; clearly, their conserved intergenic regions are less similar in length.

regions of both species, the plot shows a much wider scatter. Below, we study if the characteristics of this scatterplot are compatible with the model described above.

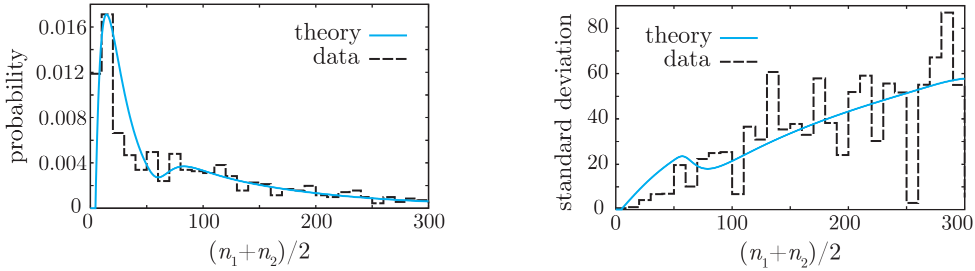
6.3.2 The model can fit the data

As the density of points in the scatterplots is very low in most regions, it is hard to directly compare the scatterplots to the probability distributions computed from our model. Therefore, we study two key signatures of the distribution and compare these to our model. To this end, we make use of a coordinate transformation. Given the lengths n_1 and n_2 of a certain intergenic region in *E. coli* and *S. Typhi*, we define:

$$x \equiv \frac{n_1 + n_2}{2}, \tag{6.5}$$

$$y \equiv \frac{n_1 - n_2}{2}. \tag{6.6}$$

The coordinate x is the mean length of the intergenic region in the two species, and y is a measure of the distance between the point (n_1, n_2) and the diagonal $n_1 = n_2$ of the scatterplot. In terms of these coordinates we can define two functions characterizing the scatterplots: the probability distribution $\rho(x)$, and the function $\sigma_y(x)$, defined as the standard deviation of y for points with a given



(a) Probability density as a function of $(n_1 + n_2)/2$.

(b) Standard deviation of $(n_1 - n_2)/2$ as a function of $(n_1 + n_2)/2$.

Figure 6.3: Fits of the model to key aspects of the scatterplots: the probability density $\rho(x)$, where $x = (n_1 + n_2)/2$, and the standard deviation $\sigma_y(x)$ in $y \equiv (n_1 - n_2)/2$ as a function of x. The model can fit the data reasonably well; yet, the fit parameter t_a (the time that passed since the divergence of *E. coli* and *S. Typhi* from their common ancestor) is unrealistically high: $t_\mathrm{a} = 16/b^\bullet$ (see text). The size of $\sigma_y(x)$ can therefore not be explained by insertions as deletions of single nucleotides.

value of x.

The functions $\rho(x)$ and $\sigma_y(x)$ can be interpreted as follows. Shortly after divergence, the distances in both species are (almost) equal. This means that $x \approx n_1 \approx n_2$. Therefore, $\rho(x) \approx P(n_1) \approx P(n_2)$, where $P(n_i)$ denotes the length distribution of intergenic regions in species i. At this point, $y \approx 0$ for all intergenic regions, so that $\sigma_y(x) \approx 0$ for all x. As time progresses, the values of n_1 and n_2 de-correlate. Hence, $\sigma_y(x)$ increases with time for all x. At the same time, $\rho(x)$ evolves too; for instance, the variance of $\rho(x)$ decreases as n_1 and n_2 de-correlate.

We computed the functions $\rho(x)$ and $\sigma_y(x)$ both for the data in Fig. 6.2(b) and for our model, using t_a (the divergence time) as a fitting parameter. The remaining parameters (the indel rates, the fraction of the intergenic regions that are of type "intra", and the lengths of promoters, terminators and Shine–Dalgarno sequences) are chosen based on the *static* length distributions of intergenic regions (see Appendix 6.C). The results are presented in Fig. 6.3. Both plots demonstrate that our model fits the data quite well. The plot 6.3(b) shows that, in accordance with our model, longer intergenic regions "diffuse" faster than short ones.

Only at small x, where the distributions are dominated by intergenic regions of type "intra", our model predicts a diffusion that is slightly too fast relative to the diffusion at larger x; this is evident from the peak in the model (at ≈ 60 bp) that is not observed in the data. In our model, this peak is a consequence of the following. Consider an intergenic region of length ≈ 70 bp of the type "inter". This intergenic region must have very short spacers, since a large fraction of it is occupied by a promoter and a terminator. Hence, the rate of change of this intergenic regions is very low. Intergenic regions of the type "intra" with a similar length contain quite a long spacer, and therefore have a much higher total insertion and deletion rate.

We could produce a better fit by lowering the insertion and deletion rates for transcribed spacers as compared to those of the untranscribed spacers; this can remove the peak (see Appendix 6.C). However, we do not have a clear explanation why the diffusion of transcribed spacers would be much slower than that of the untranscribed ones. One possibility is that, in reality, long transcribed spacers are likely to contain internal promoters or other sites for internal regulation. This could impose evolutionary length constraints that are not included in the model.

6.3.3 Required number of mutations far too high

Even though the fits shown in Fig. 6.3 are quite good, it turns out that the fit parameter t_a is not in the correct range: $t_\mathrm{a} \approx 16/b^\bullet$, where we use $1/b^\bullet$ as our unit of time, $b^\bullet$ being the deletion rate in the UTRs (the highest rate in the system). This means that the measured standard deviation $\sigma_y(x)$ is expected to occur only after about 16 deletions and a similar number of insertions have occurred on average at every site in the spacers. If this would be correct, no homology whatsoever should be expected between intergenic regions of *E. coli* and *S. Typhi*, except for sites that are under selection. This is not consistent with the high level of sequence similarity observed (*e.g.* Mustonen and Lassig (2005)). We conclude that the differences in the lengths of intergenic regions between *E. coli* and *S. Typhi* cannot be attributed to indels of single base pairs only.

Our initial assumption that all insertions and deletions have a length of 1 bp is apparently incorrect; yet, the fits in Fig. 6.3 are reasonable. This raises two questions. The first is: What is the influence of indels larger than 1 bp on the evolution of the lengths of intergenic regions? The second question is: Under which conditions can our simple model be used as an approximation of models that *do* include larger indels, by interpreting the parameters a and b as “effective” insertion and deletion rates that represent the total effect of indels of all sizes? We now address these questions within a class of models that do include indels larger than 1 bp.

6.3.4 Larger insertions and deletions

We replace our initial model (that we call the “restricted” model) by a more general one. In our extended model, indels can be larger than 1 bp, but are still assumed to be short: we assume that indels of length d or larger can be ignored. The Master equation describing the evolution of spacers can then be written as:

$$\begin{aligned}\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = {} & \sum_{0<i<d} b_i(n+1)P(n+i,t) + \sum_{0<i<d} a_i(n-i+1)P(n-i,t) \\ & - \sum_{0<i<d} b_i(n-i+1)P(n,t) - \sum_{0<i<d} a_i(n+1)P(n,t).\end{aligned} \tag{6.7}$$

Here a_i is the rate of insertions of length i per insertion site and b_i is the rate of deletions of length i per possible locus. A subtle point is that in a spacer of length n, insertions of any size can occur in $n+1$ places, whereas deletions of length i can occur in only $n-i+1$ places. For $d=2$ we retrieve our restricted model; the description becomes more general if a larger d is chosen.

The dynamics described by the Master equation 6.7 depend on all parameters a_i and b_i. However, different combinations of parameter values may result in a similar behavior. To study what are the relevant (combinations of) parameters, we now approximate the Master equation by a Fokker–Planck equation (van Kampen, 1992). This requires that we from now on consider n as a continuous quantity and $P(n,t)$ as a probability density function over the positive real numbers. We assume that d is small compared to the length scale at which $P(n,t)$ varies; under this assumption, we can replace $P(n+i,t)$ and $(n-i+1)P(n-i,t)$ by their second order Taylor expansions around $i=0$:

$$P(n+i,t) \approx P(n,t) + i\frac{\partial P(n,t)}{\partial n} + \frac{1}{2}i^2\frac{\partial^2 P(n,t)}{\partial n^2},$$

$$(n-i+1)P(n-i,t) \approx \left((n-i+1) - i(n-i+1)\frac{\partial}{\partial n} - \frac{i^2(n+1)}{2}\frac{\partial^2}{\partial n^2}\right)P(n,t).$$

Thus we arrive at the following Fokker–Planck equation:

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \left(-\frac{\partial}{\partial n}\left(B - A(n+1)\right) + \frac{1}{2}\frac{\partial^2}{\partial n^2}D(n+1)\right)P(n,t), \qquad (6.8)$$

in which the constants A, B and D are defined as

$$A \equiv \sum_{0<i<d} i(b_i - a_i), \qquad B \equiv \sum_{0<i<d} i^2 b_i, \qquad D \equiv \sum_{0<i<d} i^2(b_i + a_i). \qquad (6.9)$$

This is a diffusion equation with a drift term, in which both the diffusion coefficient and the drift coefficient are linear in n.

Apparently, in the regime where we can use the Fokker–Planck approximation, the evolution of the system is fixed by the three lumped parameters D, A and B. The parameter D sets the time scale of the diffusive part of the dynamics, whereas A and B scale the speed and length dependence of the drift term. In the restricted model ($d=2$), the parameters reduce to $A=b-a$, $B=b$, and $D=b+a$.

The Equations 6.9 show that D and B are sums of the indel rates, weighted by their length *squared*. This means that indels of length $i>1$ contribute strongly to these values, even if the rates a_i and b_i decrease with i. This holds to a lesser extent for A, in which the rates are weighted *linearly* by their length. The restricted model, which does not take into account indels with $i>1$, is therefore likely to strongly underestimate the speed of the dynamics, and in particular of the diffusion part; this explains that it needs unreasonably high indel numbers to fit the data.

This effect can easily lead to an underestimation of the diffusion speed of two orders of magnitude. The following toy model illustrates this. Suppose that the rates a_i and b_i decays linearly as a function of i, according to

$$a_i = \begin{cases} a_1(11-i)/10 & \text{if } 0 < i \leq 10, \\ 0 & \text{else,} \end{cases} \tag{6.10}$$

and

$$b_i = \begin{cases} b_1(11-i)/10 & \text{if } 0 < i \leq 10, \\ 0 & \text{else,} \end{cases} \tag{6.11}$$

In this case, $D = 121(a_1 + b_1)$. Ignoring the indels with $i > 1$ in this case indeed leads to an underestimation of D by a factor 121.

Apart from the overall speed of the dynamics, what else changes when we allow for larger indels? By writing the Fokker–Planck equation as

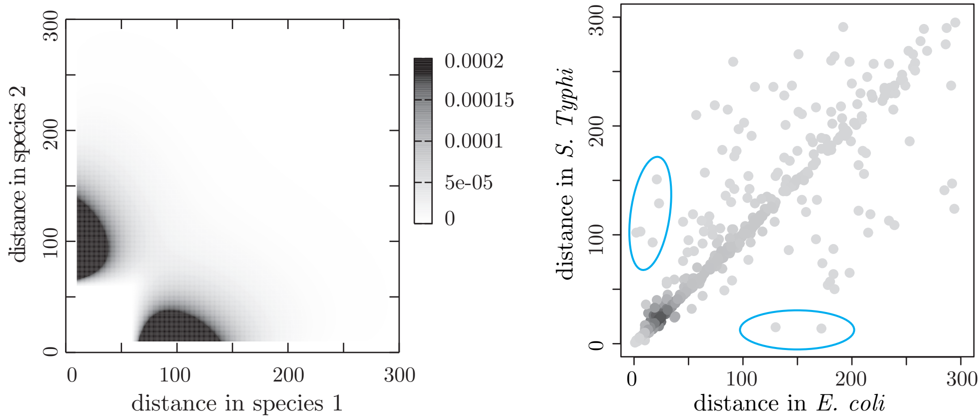
$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = D\left(-\frac{\partial}{\partial n}\left(\frac{B}{D} - \frac{A}{D}(n+1)\right) + \frac{1}{2}\frac{\partial^2}{\partial n^2}(n+1)\right)P(n,t), \tag{6.12}$$

it is clear that the solution of the equation is fixed by the parameters $\tilde{B} \equiv B/D$ and $\tilde{A} \equiv A/D$ except for an overall scaling of the time. It follows that the steady-state solution of Equation 6.12, $P_{\mathrm{st}}(n)$, can be written in terms of the two parameters $\tilde{A}$ and $\tilde{B}$ only; indeed,

$$P_{\mathrm{st}}(n) \propto (1+n)^{2\tilde{B}-1}\mathrm{e}^{-2\tilde{A}n}. \tag{6.13}$$

This is the product of an exponential factor depending on $\tilde{A}$ and a second factor depending on $\tilde{B}$. In contrast, the steady-state solutions of our *restricted* model were strictly exponential (see Equation 6.3). This means that our restricted model can certainly not be used as an approximation for the extended model, unless $\tilde{B} \approx 1/2$. Conversely, if we do assume that $\tilde{B} \approx 1/2$, then our restricted model can approximate the dynamics of the extended model with any set of parameter values D and A (provided $D \geq A$, which is true by definition: see Equation 6.9) by choosing effective rates a and b such that $D = b + a$ and $A = b - a$.

To summarize, assuming that indels are typically short so that the system is approximated well by the Fokker–Planck equation, the restricted model should be able to fit the data for some values a and b, provided that $\tilde{B} \approx 1/2$. From the definition of $\tilde{B}$ this means that $\sum_i i^2 b_i \approx \sum_i i^2 a_i$. The fact that the restricted model works reasonably well suggests that this relation holds approximately.



(a) Expected probability density for the merging and splitting modes.

(b) Scatterplot of intergenic distances: *E. coli* vs. *S. Typhi*

Figure 6.4: Probability density of the merging and splitting modes resulting from our model ($t_a = 16/b^\bullet$) compared with the real data. Two small groups of point, indicated in Fig. (b), occur in the domains of the scatterplot where intergenic regions are expected that recently underwent a merging or splitting event.

6.3.5 Outliers are candidates for merging or splitting

In Fig. 6.4 we compare the results of our calculations of the merging/splitting pathways to the data. The total joint probability density for the merging and splitting modes as resulting from our model calculations is shown in Fig. 6.4(a). The results shown correspond to the time point $t_a = 16/b^\bullet$ (the value of the fit parameter in Fig. 6.3). The intergenic regions in which a merging or splitting event has recently taken place are expected to appear in particular regions of the graph, where one of the distances is rather small, and the other distance is rather large. These results should be compared with the data in Fig. 6.4(b). We identified two small clusters of points that are located in the correct part of the scatterplot. The positions of these points suggest that they correspond to intergenic regions in which a merging or splitting event has recently taken place.

The two clusters contain 5 and 2 points. However, the intergenic regions included in the plot were selected using a very strict criterium on the conservation of the flanking genes (see Appendix 6.A). Using a slightly less strict criterium plus a manual check, we could select several more examples; the final set contained 14 intergenic regions that are long in *S. Typhi* and short in *E. coli* and 4 in which the opposite is the case. These intergenic regions and their sizes are listed in Table 6.1.

We studied the intergenic regions in Table 6.1 in detail. For several reasons, it is hard to decide if merging or splitting event has indeed occurred in a given region. To start with, in many cases there is no experimental evidence proving that a

E. coli			*S. Typhi*		
gene pairs		n_1	gene pairs		n_2
caiD	*caiE*	6	*caiD*	*caiE*	112
glpG	*glpR*	19	*glpG*	*glpR*	93
hisG	*hisD*	6	*hisG*	*hisD*	103
hypA	*hypB*	4	*hypA*	*hypB*	70
lexA	*dinF*	19	*lexA*	*dinF*	179
panB	*panC*	12	*panB*	*panC*	126
radA	*nadR*	21	*radA*	*nadR*	151
rpsG	*fusA*	28	*rpsG*	*fusA*	97
sdaB	*ygdG*	22	*sdaB*	*exo*	111
ubiH	*visC*	23	*visB*	*visC*	129
wcaJ	*wzxC*	2	*wcaJ*	*wzxC*	102
yafD	*yafE*	4	*yafD*	*yafE*	78
ybhS	*ybhR*	11	*t2072*	*t2073*	116
ygiF	*glnE*	23	*t3122*	*glnE*	118

(a) $n_1 < n_2$

E. coli			*S. Typhi*		
gene pairs		n_1	gene pairs		n_2
dnaG	*rpoD*	195	*dnaG*	*rpoD*	15
flgF	*flgG*	172	*flgF*	*flgG*	14
uup	*pqiA*	130	*t1858*	*pqiA*	15
yfhK	*yfhG*	165	*t0292*	*yfhG*	2

(b) $n_1 > n_2$

Table 6.1: List of intergenic regions that lie in those areas of the scatter plot (Fig. 6.4(b)) that are associated with merging or splitting processes. The intergenic regions in *E. coli* and their ortholog in *S. Typhi* are specified by the pairs of genes flanking them. The numbers n_1 and n_2 are the lengths of the regions in *E. coli* and *S. Typhi* respectively.

given intergenic region is of type "inter" or "intra". Algorithms predicting operons in *E. coli* are largely based on the lengths of the intergenic regions and therefore do not offer completely independent information. In *S. Typhi*, the extents of the majority of transcription units have not been determined experimentally and predictions are typically based on knowledge about *E. coli*. Through sequence analysis, promoter and terminator predictions have been made (*e.g.* Huerta and Collado-Vides (2003); Lesnik *et al.* (2001); Kingsford *et al.* (2007)), but these have a limited accuracy and scope (for instance, terminator predictions are limited to Rho-independent terminators).

Nevertheless, some patterns can be identified in the data. We discuss them now.

Intergenic regions similar to *uup–pqiA*: overlapping internal promoters

Four cases are very similar: *dnaG–rpoD*, *hypA–hypB*, *uup–pqiA* and *rpsG–fusA*. We use the region *uup–pqiA* as an example.

The intergenic region *uup–pqiA* is long in *E. coli* (130 bp) and short in *S. Typhi* (15 bp). We do not know of any direct experimental evidence showing if *uup* and *pqiA* are co-transcribed in *E. coli*. Given the length of the intergenic region in *E. coli* and the fact that the two genes are not obviously functionally related, one would expect that the *uup–pqiA* region is between two operons in *E. coli*. This is also predicted by Price *et al.* (2005a). In *S. Typhi*, the region is very short

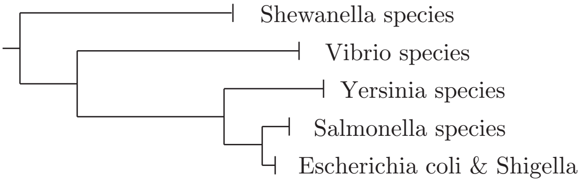


Figure 6.5: Phylogenetic tree of the species in the text. The information for this tree was taken from Price *et al.* (2006).

and therefore more likely to be an intra-operon region. As the length of the orthologous region in *Yersinia pestis*, a closely related out-group species (see the phylogenetic tree in Fig. 6.5), is 129 bp — very similar to the length in *E. coli* — this suggests that the operons have merged in *S. Typhi*. This scenario is supported by the predictions of both Lesnik *et al.* (2001) and Kingsford *et al.* (2007) (using the software `RNAMotif` and `TransTermHP` respectively) of a Rho-independent terminator in the *uup–pqiA* intergenic region in *E. coli*, but not in *S. Typhi*.

However, some subtleties cast doubt on this scenario. Experimentally, two promoters, pqiAp1 and pqiAp2, have been found inside the open reading frame *uup* in *E. coli* (Koh and Roe, 1995). From these promoters, *pqiA* is transcribed. In this light, it is hard to imagine that the terminator predicted by `RNAMotif` and `TransTermHP` actually exists: in that case one would expect the transcripts from both promoters pqiAp1 and pqiAp2 to terminate on this terminator and therefore never to reach *pqiA*. Hence an alternative scenario should be considered. In this scenario, the genes *uup* and *pqiA* are in the same operon, both in *E. coli* and in *S. Typhi*. The promoters pqiAp1 and pqiAp2 would then be internal promoters allowing for regulatory fine-tuning of the expression of *pqiA*. The deletion in *S. Typhi* of part of the intergenic region could in that case be without severe consequences, because it contains neither a promoter nor a terminator. On the other hand, this scenario does not readily explain the apparent conservation of the unusual intergenic length in *E. coli* and *Yersinia pestis*.

The three cases *dnaG–rpoD*, *hypA–hypB* and *rpsG–fusA* are very similar to the example *uup–pqiA*. In all cases, the intergenic region is short in one species and long in the other, suggesting a merging or splitting event; but in each case internal promoters have been found overlapping with the gene directly upstream of the intergenic region (*i.e.* in *uup*, *dnaG*, *hypA* and *rpsG*), suggesting the alternative scenario.

Intergenic regions similar to *glnE–ygiF*: recent deletion of intervening genes

Several intergenic regions are similar to *glnE–ygiF*. We present this region as an example.

The region *glnE–ygiF* is rather short in *E. coli* (23 bp). This short spacing suggests that the flanking genes are in the same operon — which is also predicted by Price *et al.* (2005a) — but there is no direct experimental evidence for this. In

S. Typhi the region is quite long: 118 bp. Given that the spacing is even wider in *Yersinia pestis* (144 bp) it seems likely that part of the intergenic region was deleted in *E. coli* and the flanking operons merged as a result of this.

An alternative scenario presents itself if we study the same region in *Vibrio* and *Shewanella* species. These bacteria are even more distantly related to *E. coli* and *S. Typhi* than the *Yersinia*—see the phylogenetic tree in Fig. 6.5. In both *Vibrio* and *Shewanella*, the two genes *glnE* and *ygiF* are separated on the chromosome by one to six other genes, depending on the species. Some of these genes are coded on the same DNA strand as *glnE* and *ygiF*, and some on the opposite strand; this strongly suggests that in (at least some of) these species, the genes *glnE* and *ygiF* are not in the same operon. Apparently, the intervening genes have gradually been deleted or translocated in the recent course of evolution. Therefore, the length differences of the intergenic region in *E. coli*, *S. Typhi* and *Yersinia* can be interpreted in two ways. The first interpretation is that the intergenic region is of type "intra" in *E. coli* and of type "inter" in *S. Typhi* and *Yersinia*; the second interpretation argues that the length difference could be the result of an incomplete deletion of one or more intervening genes followed by a process of erosion that happens to have advanced more in *E. coli*. In the latter case, the genes *glnE* and *ygiF* could be in the same operon in all three species.

In our data set, the intergenic regions *lexA–dinF* and *wcaJ–wzxC* are also of this type: in other organisms these gene pairs are separated by one or several genes, suggesting a recent deletion of intervening genes. Deletion of intervening genes was also identified by Price *et al.* (2006) as one of the important mechanisms of operon formation; they present several examples.

Intergenic regions similar to *sdaB–ygdG*: recent insertion or translocation of genes

The intergenic region *sdaB–ygdG* is short in *E. coli* (22 bp) but rather long in *S. Typhi* (111 bp). In both species, it is not known if the genes are in the same transcription unit. It is relevant here to consider the upstream neighborhood of this locus. Both in *E. coli* and in *S. Typhi* the gene order is: *ygdH–sdaC–sdaB–ygdG*. The proteins SdaC and SdaB are clearly functionally related: SdaC is a serine transporter and SdaB a serine deaminase. The other genes, *ygdG* and *ygdH*, are not obviously related.

Interestingly, in several less closely related bacteria, such as the *Yersinia*, *Shewanella* and *Vibrio* species, *sdaC* and *sdaB* are also next to each other, but at a very different position on the chromosome. In these species, *ygdH* and *ygdG* are neighbors. This strongly suggests that a stretch of DNA containing *sdaC* and *sdaB* has been translocated after divergence of *E. coli* and *S. Typhi* from *Yersinia*, and that it has been inserted between *ygdH* and *ygdG*.

Again, at least two scenarios could explain the length difference between *E. coli* and *S. Typhi*. First, it is possible that the operon *sdaC–sdaB* was first inserted including a terminator. In that case, the short intergenic distance between *sdaB* and *ygdG* in *E. coli* suggests that the *sdaC–sdaB* operon subsequently merged with the transcription unit of *ygdG*, possibly through the loss of a terminator.

Second, the operon *sdaC–sdaB* may have been inserted without a terminator. In this case, *ygdG* became part of the operon from the start. Directly after the insertion, the intergenic distance *sdaB–ygdG* may have been long, which it still is in *S. Typhi*; in *E. coli*, deletions could have removed parts of it.

Several other intergenic regions show signs of recent insertion of one of the flanking genes. Sometimes these genes seem to have been translocated from other parts of the chromosome, and sometimes they may have been acquired through horizontal gene transfer. Examples are: *caiD–caiE*, *radA–nadR* and *yafD–yafE*. Price *et al.* (2006) also mention horizontal transfer and re-arrangement as an important mechanism for operon formation.

Remaining intergenic regions

In some of the remaining cases, the intergenic region under consideration is short in all related organisms; the long intergenic distance in *E. coli* or *S. Typhi* is the exception. This holds for *flgF–flgG*, *hisG–hisD*, *ybhS–ybhR* and *yfhK–ythG*. In all these cases, the extent of the transcription units is unknown. In *yfhK–ythG* a promoter is predicted by Huerta and Collado-Vides (2003).

The lengths of the remaining regions, *panB–panC* and *ubiH–visC*, vary extensively between organisms. For instance, *panB* and *panC* overlap with three nucleotides in the *Pseudomonas*, *Shewanella* and *Shigella* species — this shows that they are in the same operon in these species. The intergenic region is quite short in *Vibrio* (11 bp), *Buchnera* (15 bp) and *E. coli* (12 bp), but rather long in *Yersinia* (88 bp), *Shewanella* (69 bp) and *S. Typhi* (129). It is unclear what this variability conveys.

6.4 Conclusions

We presented a stochastic model to describe the evolution of the lengths of intergenic regions. By comparing the lengths of intergenic regions in *E. coli* to those of their orthologs in *S. dysenteriae* and *S. Typhi*, we could fit our model to real data. The data show that the distances perform a diffusion-type motion with drift. The data are compatible with a model in which both the diffusion rate and the drift are linear functions of the length of the intergenic regions, as was predicted by the model.

However, our initial assumption that insertions and deletions are all 1 bp long leads to an unreasonable value of a fit parameter: unrealistically high numbers of mutations are required to fit the data. This shows that the differences between the lengths of intergenic regions in related species cannot be explained by indels of 1 bp only. Indeed, by analyzing a more general model, we demonstrated that larger indels can contribute strongly to the time scales of the diffusion. Yet, the initial, restricted model can be used as an approximation of the more general one, provided the indels are small and $\sum_i i^2 a_i \approx \sum_i i^2 b_i$. In this case, the insertion and deletion rates in the initial model should be re-interpreted as *effective* rates that reflect the effect of the indels of all sizes together.

More detailed models could be constructed if more information becomes available about the distributions of insertion and deletion lengths in bacterial genomes. Such information could be obtained through careful analysis of alignments of intergenic regions of several bacteria. This is, however, not a straightforward endeavor. In order to distinguish between insertions and deletions, the sequences of at least three closely related species should be aligned, and the typical artifacts of alignment algorithms such as *gap attraction* should be avoided.

Our calculations also address the scenarios in which, during the divergence from the common ancestor, two operons merged or one operon was split in one of the lineages. These computations lead to the prediction of intergenic regions in which a merging or splitting event may recently have taken place. These intergenic regions are long in one of the species (*E. coli* or *S. Typhi*), and short in the other. In most cases, the extents of the relevant transcription units in *E. coli* and/or *S. Typhi* have not been determined experimentally, which makes it hard to conclude if a splitting or merging event has indeed taken place. Yet, a close examination of the candidates suggests that other mechanisms than merging or splitting events may also lead to intergenic regions with very different lengths in two related species. These include: (i) internal promoters, which in some cases (partly) overlap with one of the open reading frames; (ii) recent deletions of genes, and (iii) recent insertions or translocations. It would be interesting to see if expression patterns could shed more light on the candidate regions and their recent evolutionary history.

6.A The data for the scatterplots

In order to create the scatterplots in Fig. 6.2, we first identified orthologous gene pairs between *E. coli* and each of the other organisms by aligning them using the NCBI `blastp` program. We used amino acid sequences downloaded from NCBI Genbank[3]. As the (apparent) lengths of some intergenic regions can be (and in fact are) affected by errors in the open reading frame predictions of the flanking genes, we used a very strict criterium for inclusion in our data set: genes were considered orthologs only if their sequence identity was at least 40% and the length of the alignment was at least 95% of the longest of the two genes.

Next, we used the selected orthologs to list all pairs of tandem neighbors in *E. coli* that were preserved as tandem neighbors in the other species. The intergenic regions between these gene pairs were included in the data set.

6.B Details of the model

Here, we discuss the derivations and computational aspects of the model in more detail. The calculations are similar in spirit to the method described in (Mustonen and Lassig, 2005), dealing with the evolution of the binding energies of transcription factor binding sites.

6.B.1 The evolution of spacers

We assume that the evolution of the intergenic regions is determined by the evolution of the spacers that they contain. These spacers come in three different kinds: UTRs, untranscribed spacers and transcribed spacers. The basic model of the evolution of the distances is the same for each of these, except for different insertion and deletion rates a and b. The Master equation and steady-state probability distribution for the spacers are given in Equations 6.2 and 6.3.

We can numerically approximate the time evolution of this system by imposing a maximal length N. Given such a maximal length, the Master equation 6.2 reduces to a system of N linear differential equations; in vector and matrix notation, it can be written as:

$$\frac{\mathrm{d}\vec{p}(t)}{\mathrm{d}t} = \mathrm{M}\vec{p}(t), \tag{A6.1}$$

---

[3] http://www.ncbi.nlm.nih.gov/

where

$$\mathrm{M}_{nm} = \begin{cases} (n+1)b & \text{if } n+1=m, \\ na & \text{if } n=m+1, \\ -(n+1)a-nb & \text{if } n=m<N, \\ -nb & \text{if } n=m=N, \\ 0 & \text{else.} \end{cases} \tag{A6.2}$$

Here $\vec{p}(t)$ is defined by $p_n(t) \equiv P(n,t)$. This set of equations can easily be solved numerically. Defining E as the matrix having the eigenvectors $\vec{v}_1 \ldots \vec{v}_N$ of M as columns, we can define

$$\mathrm{D} \equiv \mathrm{E}^{-1}\mathrm{M}\mathrm{E}, \tag{A6.3}$$

where D is a diagonal matrix. The solution of Equation A6.1 is then given by

$$\vec{p}(t_2) = \mathrm{E}e^{\mathrm{D}(t_2-t_1)}\mathrm{E}^{-1}\vec{p}(t_1), \tag{A6.4}$$

which can be computed straightforwardly for given t_1, t_2 and initial condition $\vec{p}(t_1)$. We also define the matrix $\mathrm{G}_{nm}(t_2,t_1)$ as the probability to evolve, given length m at time t_1, to length n at time t_2:

$$\mathrm{G}(t_2,t_1) \equiv \mathrm{E}e^{\mathrm{D}(t_2-t_1)}\mathrm{E}^{-1}. \tag{A6.5}$$

Evidently, $\mathrm{G}(t_2,t_1)$ only depends on the *difference* between t_2 and t_1; it is therefore useful to also define

$$\mathrm{G}(t) \equiv \mathrm{E}e^{\mathrm{D}t}\mathrm{E}^{-1}. \tag{A6.6}$$

Here we are interested in the situation where two organisms evolved independently after their divergence from a common ancestor, a time t_a ago. We set out to compute the probability $\mathrm{J}_{nm}(t)$ of finding a spacer of length n in species 1 that has a length m in species 2. This distribution can be computed given the distance distribution in the common ancestor, $P_\mathrm{a}(n)$:

$$\mathrm{J}_{nm}(t_\mathrm{a}) = \sum_o P_\mathrm{a}(o)\mathrm{G}_{no}(t_\mathrm{a})\mathrm{G}_{mo}(t_\mathrm{a}). \tag{A6.7}$$

Note that at $t_\mathrm{a}=0$,

$$\mathrm{J}_{nm}(0) = \sum_o P_\mathrm{a}(o)\delta_{no}\delta_{mo} = \delta_{nm}P_\mathrm{a}(n), \tag{A6.8}$$

confirming that, at time $t_\mathrm{a}=0$, the organisms are identical so that $\mathrm{J}_{nm}(0)=0$ if $m \neq n$.

After a long period of evolution, the length distribution of spacers in an organism should converge to the equilibrium distribution given in Equation 6.3. In the following, we assume that this convergence had already taken place in the common ancestor of the two species[4]. This means that $P_{\mathrm{a}}(n) = P_{\mathrm{eq}}(n)$. We indicated before that detailed balance holds for the dynamics described in the Master equation 6.2. Therefore, we can exploit the following detailed-balance relation,

$$P_{\mathrm{eq}}(m)\mathrm{G}_{nm}(t) = P_{\mathrm{eq}}(n)\mathrm{G}_{mn}(t), \tag{A6.9}$$

to simplify expression A6.7:

$$\begin{aligned}\mathrm{J}_{nm}(t_{\mathrm{a}}) &= \sum_o \mathrm{G}_{no}(t_{\mathrm{a}})\mathrm{G}_{mo}(t_{\mathrm{a}})P_{\mathrm{a}}(o) = \sum_o \mathrm{G}_{no}(t_{\mathrm{a}})\mathrm{G}_{mo}(t_{\mathrm{a}})P_{\mathrm{eq}}(o) \\ &= \sum_o \mathrm{G}_{no}(t_{\mathrm{a}})\mathrm{G}_{om}(t_{\mathrm{a}})P_{\mathrm{eq}}(m) = \mathrm{G}_{nm}(2t_{\mathrm{a}})P_{\mathrm{eq}}(m).\end{aligned} \tag{A6.10}$$

This shows that the independent evolution of species 1 and species 2 from an unknown ancestor over a time t_{a} is equivalent to the evolution of species 2 from species 1 in a time $2t_{\mathrm{a}}$. This time reversal of the evolutionary process is possible only because of detailed balance. In the second line of Equation A6.10 we also used that

$$\sum_o \mathrm{G}_{no}(t_1)\mathrm{G}_{om}(t_2) = \mathrm{G}_{nm}(t_1 + t_2). \tag{A6.11}$$

As we mentioned, the above theory applies to all three types of spacers, except that the parameters a and b are different for the different types (see Appendix 6.C). To distinguish quantities such as a, $P(n)$ and G_{nm} for the transcribed and the untranscribed spacers in the remaining sections, we use the following notation. We use the label * to indicate that a quantity is a property of the transcribed spacers (*e.g.* $\mathrm{J}^\star_{nm}(t_{\mathrm{a}})$). To show that a quantity belongs to the untranscribed spacers, we use the label $^\circ$ (*e.g.* $\mathrm{G}^\circ_{nm}(t)$). The notation for the UTRs is introduced below; it uses the label $^\bullet$.

6.B.2 The evolution of the UTRs

Intergenic regions of type "inter", always contain two UTRs: one 3' UTR and one 5' UTR (see Fig. 6.1). As we assume that these evolve under the same dynamics (have equal insertion and deletion rates $a^\bullet$ and $b^\bullet$), it is convenient to describe the evolution of the total length of the two UTRs with a single Master equation,

---

[4] It may be illuminating to point out that, in this model, the length distributions of the two species are in equilibrium at all times. Both organisms start with the distribution of their common ancestor, which we assume has equilibrated during the many millions of years of evolution before the speciation event. Next, each of the organisms evolves independently, but stays in equilibrium. However, directly after the speciation, the lengths in both species are perfectly correlated. This correlation does change in the course of evolution, and our method is aimed at describing this de-correlation process.

given by:

$$\frac{\mathrm{d}P^{\bullet}(n,t)}{\mathrm{d}t} = (n+1)a^{\bullet}P^{\bullet}(n-1,t) + (n+1)b^{\bullet}P^{\bullet}(n+1,t) - ((n+2)a^{\bullet} + nb^{\bullet})P^{\bullet}(n,t). \tag{A6.12}$$

Note the difference between this equation and Equation 6.2: given the total length n of the two UTRs together, the number of places where an insertion can occur is $n+2$. In the models for single spacers, this is $n+1$.

For the equilibrium probability distribution $P^{\bullet}_{\mathrm{eq}}(n)$, the detailed balance relation

$$a^{\bullet}(n+1)P^{\bullet}_{\mathrm{eq}}(n-1) = b^{\bullet}nP^{\bullet}_{\mathrm{eq}}(n) \tag{A6.13}$$

holds; from this the equilibrium solution follows:

$$P^{\bullet}_{\mathrm{eq}}(n) \propto (n+1)\exp\left(-n\ln\left(\frac{b^{\bullet}}{a^{\bullet}}\right)\right). \tag{A6.14}$$

Again, the system can be approximated numerically by imposing that $n \leq N$, leading to the following system of linear differential equations:

$$\frac{\mathrm{d}\vec{p}(t)}{\mathrm{d}t} = \mathrm{M}^{\bullet}\vec{p}(t), \tag{A6.15}$$

where

$$\mathrm{M}^{\bullet}_{nm} = \begin{cases} (n+1)b^{\bullet} & \text{if } n+1=m, \\ (n+1)a^{\bullet} & \text{if } n=m+1, \\ -(n+1)a^{\bullet} - nb^{\bullet} & \text{if } n=m<N, \\ -nb^{\bullet} & \text{if } n=m=N, \\ 0 & \text{else.} \end{cases} \tag{A6.16}$$

As for Equation A6.4, the solution follows from solving the eigensystem of this matrix. Also, the correlation matrix $\mathrm{G}^{\bullet}(t)$ and the joint probability matrix $\mathrm{J}^{\bullet}(t_{\mathrm{a}})$ can be defined as in Equation A6.6 and A6.10.

6.B.3 Combining the evolution of the UTRs and the untranscribed spacer

In the end, we are interested in the evolution of the length of the complete intergenic regions instead of the dynamics of the spacers separately. In the intergenic regions of type “intra”, only the transcribed spacer evolves, so that the evolution of this type of regions is directly given by the evolution of these spacers (see Fig. 6.1). This is not the case for the intergenic regions of type “inter”: remember that intergenic regions of this type consist of one untranscribed spacer and a pair of UTRs. Here we combine the results from the previous sections to construct the dynamics for the intergenic regions of the "inter" type.

We define the correlation function $\mathrm{G}^{\bullet\circ}(l, l', t|l'', l''')$ as the probability that an intergenic region with total UTR length l'' and untranscribed spacer length l''' evolves, after a time t, to the state in which the UTR length is l and the spacer length l'. As the evolution of the spacer and the UTRs are independent, we can write:

$$\mathrm{G}^{\bullet\circ}(l, l', t|l'', l''') = \mathrm{G}^{\bullet}_{l''l}(t)\mathrm{G}^{\circ}_{l'''l'}(t). \tag{A6.17}$$

The joint probability $\mathrm{J}^{\bullet\circ}_{nm}(t)$ of the total length of the two spacers is less obvious. We have to sum over the distance distribution of the ancestor, $P_{\mathrm{a}}(l, l')$, and over all ways n and m could be a sum of the lengths of the untranscribed spacer and the UTRs:

$$\mathrm{J}^{\bullet\circ}_{nm}(t) = \sum_{l,l'} \sum_{l''\leq n} \sum_{l'''\leq m} P_{\mathrm{a}}(l, l')\mathrm{G}^{\bullet\circ}(l''', m - l''', t|l, l')\mathrm{G}^{\bullet\circ}(l'', n - l'', t|l, l'). \tag{A6.18}$$

As we assumed that the distance distributions in the ancestor were in equilibrium and that the different spacers evolve independently,

$$P_{\mathrm{a}}(l, l') = P^{\bullet}_{\mathrm{eq}}(l)P^{\circ}_{\mathrm{eq}}(l'), \tag{A6.19}$$

so that

$$\begin{aligned}
\mathrm{J}^{\bullet\circ}_{nm}(t) &= \sum_{l,l'} \sum_{l''\leq n} \sum_{l'''\leq m} P^{\bullet}_{\mathrm{eq}}(l)P^{\circ}_{\mathrm{eq}}(l')\mathrm{G}^{\bullet\circ}(l''', m - l''', t|l, l')\mathrm{G}^{\bullet\circ}(l'', n - l''|l, l') \\
&= \sum_{l,l'} \sum_{l''\leq n} \sum_{l'''\leq m} P^{\bullet}_{\mathrm{eq}}(l)P^{\circ}_{\mathrm{eq}}(l')\mathrm{G}^{\bullet}_{l'''l}(t)\mathrm{G}^{\circ}_{m-l''',l'}(t)\mathrm{G}^{\bullet}_{l''l}(t)\mathrm{G}^{\circ}_{n-l'',l'}(t) \\
&= \sum_{l\leq m} \sum_{l'\leq n} P^{\bullet}_{\mathrm{eq}}(l)P^{\circ}_{\mathrm{eq}}(m - l)\mathrm{G}^{\bullet}_{l'l}(2t)\mathrm{G}^{\circ}_{n-l',m-l}(2t).
\end{aligned} \tag{A6.20}$$

Here we again used detailed balance (Equation A6.9) and Equation A6.11 to eliminate the summation over the ancestral distribution.

Splitting and merging of operons

In the previous section, we described the evolution of intergenic regions inside operons and between operons. We implicitly assumed that operons would not split or merge. If they do, this means that an intergenic region that used to be *between* two operons, is suddenly *inside* an operon. As a result, it will start to evolve differently. Similarly, if an operon splits, an intergenic region that used to be *inside* an operons, is suddenly *between* two operons. Here we derive the probabilities to find an intergenic region that has length n in one species and length m in the other species, given that a splitting event or a merging event has occurred in one of the species.

We assume that the rates at which operons merge or split, v_{m} and v_{s}, are very low. In that case, the switching events hardly disturb the equilibrium length distributions of the spacers and UTRs, and the results of the previous sections still hold to good approximation. Under this condition, we can also neglect scenario's in which, in a given intergenic region, more than one switching or merging event has taken place since the divergence of the two species. We also ignore the option that, in a given intergenic region, a merging or splitting event could have taken place in *both* organisms.

Merging

We first consider the scenario in which a merging of operons takes place in one of the species — say, in species 2. The merging could have taken place at any time since the common ancestor, which means that we have to integrate over time. In that case the joint probability that the total length of the spacers is m in species 1 (in which the intergenic region is of type "inter"), and n in species 2 (in which it is of type "intra"), can be written as:

$$\mathrm{J}^{\mathrm{m}}_{mn} = \frac{1}{t_{\mathrm{a}}}\int_0^{t_{\mathrm{a}}} \mathrm{d}t \sum_{i=0}^{m}\sum_{j=0}^{N}\sum_{k=0}^{N-j-\pi-\tau} P^{\bullet}(i)P^{\circ}(m-i)\mathrm{G}^{\bullet}_{ji}(t_a+t)\mathrm{G}^{\circ}_{k,m-i}(t_a+t)\mathrm{G}^{\star}_{n,j+k+\pi+\tau}(t_a-t). \quad \text{(A6.21)}$$

As we explained in Section 6.2.4, here we assume that, at the moment of merging, the promoter and terminator are destroyed, but their total length $\pi+\tau$ is preserved and therefore added to the total spacer length. It should also be noted that the length of the Shine–Dalgarno sequence is preserved in the merging process and therefore not relevant in these calculations; it should however be added to m and n to arrive at the total length of the intergenic regions.

With some effort, the time integration can be worked out explicitly. In order to do this, we have to use the definitions (Equation A6.6) of the correlation functions $\mathrm{G}^{\bullet}$, G°, and $\mathrm{G}^{\star}$:

$$\mathrm{G}^{\bullet}(t) \equiv \mathrm{E}^{\bullet}\mathrm{X}^{\bullet}(t)\left(\mathrm{E}^{\bullet}\right)^{-1}, \quad \text{(A6.22)}$$

$$\mathrm{G}^{\circ}(t) \equiv \mathrm{E}^{\circ}\mathrm{X}^{\circ}(t)\left(\mathrm{E}^{\circ}\right)^{-1}, \quad \text{(A6.23)}$$

$$\mathrm{G}^{\star}(t) \equiv \mathrm{E}^{\star}\mathrm{X}^{\star}(t)\left(\mathrm{E}^{\star}\right)^{-1}, \quad \text{(A6.24)}$$

in which

$$\mathrm{X}^{\bullet}(t) \equiv \mathrm{e}^{\mathrm{D}^{\bullet}t}, \quad \mathrm{X}^{\circ}(t) \equiv \mathrm{e}^{\mathrm{D}^{\circ}t}, \quad \mathrm{X}^{\star}(t) \equiv \mathrm{e}^{\mathrm{D}^{\star}t}. \quad \text{(A6.25)}$$

Clearly, the time dependency is only in the (diagonal) matrices $\mathrm{X}^{\bullet}(t)$, $\mathrm{X}^{\circ}(t)$ and

$\mathrm{X}^\star(t)$. This means that we can write Equation A6.21 as

$$\mathrm{J}^{\mathrm{m}}_{mn} = \frac{1}{t_\mathrm{a}} \sum_{i=0}^{m} \sum_{j,o,p,q=0}^{N} \sum_{k=0}^{N-j-\pi-\tau} P^\bullet(i) P^\circ(m-i) \mathrm{E}^\bullet_{jo} (\mathrm{E}^\bullet)^{-1}_{oi} \mathrm{E}^\circ_{kp} (\mathrm{E}^\circ)^{-1}_{p,m-i} \mathrm{E}^\star_{nq} (\mathrm{E}^\star)^{-1}_{q,j+k+\pi+\tau} \int_0^{t_\mathrm{a}} \mathrm{d}t \mathrm{X}^\bullet_{oo}(t_\mathrm{a}+t) \mathrm{X}^\circ_{pp}(t_\mathrm{a}+t) \mathrm{X}^\star_{qq}(t_\mathrm{a}-t). \tag{A6.26}$$

Next we use that $\mathrm{X}^\bullet_{oo}(t) = \exp(e^\bullet_o t)$, where $e^\bullet_o$ is the oth eigenvalue of the matrix $\mathrm{M}^\bullet$; similar expressions hold for X° and $\mathrm{X}^\star$. Therefore the integration can be performed explicitly:

$$\int_0^{t_\mathrm{a}} \mathrm{d}t \mathrm{X}^\bullet_{oo}(t_\mathrm{a}+t) \mathrm{X}^\circ_{pp}(t_\mathrm{a}+t) \mathrm{X}^\star_{qq}(t_\mathrm{a}-t) = (e^\bullet_o + e^\circ_p - e^\star_q)^{-1} \left(\mathrm{e}^{2(e^\bullet_o + e^\circ_p) t_\mathrm{a}} - \mathrm{e}^{(e^\bullet_o + e^\circ_p + e^\star_q) t_\mathrm{a}} \right). \tag{A6.27}$$

This means that:

$$\mathrm{J}^{\mathrm{m}}_{mn} = \frac{1}{t_\mathrm{a}} \sum_{i=0}^{m} \sum_{j,o,p,q=0}^{N} \sum_{k=0}^{N-j-\pi-\tau} P^\bullet(i) P^\circ(m-i) \mathrm{E}^\bullet_{jo} (\mathrm{E}^\bullet)^{-1}_{oi} \mathrm{E}^\circ_{kp} (\mathrm{E}^\circ)^{-1}_{pm-i} \mathrm{E}^\star_{nq} (\mathrm{E}^\star)^{-1}_{qj+k+\pi+\tau} (\mathrm{e}^\bullet_o + e^\circ_p - e^\star_q)^{-1} \left(\mathrm{e}^{2(e^\bullet_o + e^\circ_p) t_\mathrm{a}} - \mathrm{e}^{(e^\bullet_o + e^\circ_p + e^\star_q) t_\mathrm{a}} \right). \tag{A6.28}$$

This summation can now be carried out numerically[5].

In practice, a brute force summation requires a sum over six indices (i, j, k, o, p and q) for each value of m and n. This would result in a calculation time scaling as $\mathcal{O}(N^8)$, which is out of the question for $N \approx 400$. A more efficient factorization can reduce the required number of steps to $\mathcal{O}(N^4)$.

Splitting

To describe the histories in which an operon splits, we use a similar method. Here, however, we have to decide what happens to the length of the intergenic region at the moment the splitting occurs. Before the splitting, the intergenic region only contains a transcribed spacer and a Shine–Dalgarno sequence; after the splitting

---

[5] The factor $(e^\bullet_o + e^\circ_p - e^\star_q)^{-1}$ in Equation A6.28 can become very large if $e^\bullet_o + e^\circ_p - e^\star_q$ happens to be very small. This is likely to occur for some values of (o, p, q), and in these cases the numerical evaluation may fail. Therefore, it is safer to use an algorithm that detects such cases and then uses a Taylor approximation of the integrand in Equation A6.27 for $\epsilon \equiv e^\bullet_o + e^\circ_p - e^\star_q \ll 1/t_\mathrm{a}$:

$$\int_0^{t_\mathrm{a}} \mathrm{d}t \mathrm{X}^\bullet_{oo}(t_\mathrm{a}+t) \mathrm{X}^\circ_{pp}(t_\mathrm{a}+t) \mathrm{X}^\star_{qq}(t_\mathrm{a}-t) \approx (t_\mathrm{a} + \frac{1}{2}\epsilon t_\mathrm{a}^2 + \ldots) \exp((e^\bullet_o + e^\circ_p + e^\star_q) t_\mathrm{a})$$

has occurred, we have two UTRs, a promoter, a terminator and an untranscribed spacer. The question rises whether the promoter and terminator evolve from preexisting sequences or are *added* to the sequence. The same question can be raised about the UTRs.

In reality, both processes may occur. It is imaginable that some promoters develop from scratch, whereas others may arise from duplications of other sequences, for instance in recombination events. As far as we are aware, there is no compelling evidence for either scenario, or for other alternatives. We therefore make a convenient choice: at the moment of splitting, a promoter is inserted including the 5' UTR, and a terminator is inserted including the 3' UTR. Now, we still need to choose the total length of the UTRs. We choose to draw the inserted UTRs from the equilibrium distribution of the UTR lengths, which was given in Equation A6.14.

Given these model choises, the probability to find a total spacer length n in species 1 and a total spacer length m in species 2, given that a splitting event occurred in species 2, can be written as:

$$\mathrm{J}^{\mathrm{s}}_{mn} = \frac{1}{t_{\mathrm{a}}} \int_0^{t_{\mathrm{a}}} \sum_{j=0}^{N} \sum_{i=0}^{m} P^{\star}(n) \mathrm{G}^{\star}_{jn}(t_{\mathrm{a}} + t) \mathrm{G}^{\circ}_{ij}(t_{\mathrm{a}} - t) P^{\bullet}(m - i). \tag{A6.29}$$

The time integration can again be done explicitly and leads to

$$\mathrm{J}^{\mathrm{s}}_{mn} = \frac{1}{t_{\mathrm{a}}} \sum_{j,o,p=0}^{N} \sum_{i=0}^{m} P^{\star}(n) P^{\bullet}(m-i) \mathrm{E}^{\star}_{jo} (\mathrm{E}^{\star})^{-1}_{on} \mathrm{E}^{\circ}_{ip} (\mathrm{E}^{\circ})^{-1}_{pj} \\ (e^{\star}_{o} - e^{\circ}_{p})^{-1} \left(\mathrm{e}^{2(e^{\star}_{o} - e^{\circ}_{p}) t_{\mathrm{a}}} - \mathrm{e}^{(e^{\star}_{o} - e^{\circ}_{p}) t_{\mathrm{a}}} \right). \tag{A6.30}$$

6.C Parameter choices

The parameters of the model presented in this chapter are: the total length of a promoter plus a transcriptional terminator ($\pi + \tau$; only the sum is relevant), the size of a Shine–Dalgarno sequence (s), the fraction of intergenic regions that is of type "intra" ($\lambda = w_1 + (w_3 + w_4)/2$; see Equation 6.1) and the insertion and deletion rates for each type of spacer (*i.e.* $a^{\star}$ and $b^{\star}$ for the spacers inside operons, a° and b° for the untranscribed spacers, and $a^{\bullet}$ and $b^{\bullet}$ for the UTRs). Here we describe the parameters we chose in the simulation results that we presented.

For the lengths of the promoters, terminators and Shine–Dalgarno sequences we used $\pi + \tau = 50\,\mathrm{bp}$ and $s = 5\,\mathrm{bp}$. The results are not very sensitive to these values; changing s, for instance, merely leads to a shift of all distance distributions. A good fit to the static distance distributions is obtained if 55% of the tandem intergenic regions is assumed to be of type "intra" ($\lambda = 0.55$).

The values of the insertion and deletion rates are less straightforward. We explained on page 106 that the mean length $\bar{n}$ of a given type of spacers is determined by the ratio b/a according to $\bar{n} = (b/a - 1)^{-1}$. A good fit of the static distance distributions is obtained if we take $\bar{n}^{\circ} = 125\,\text{bp}$, $\bar{n}^{\star} = 25\,\text{bp}$, and $\bar{n}^{\bullet} = 10\,\text{bp}$. This procedure fixes three ratios of mutation rates, but not their absolute values; effectively it reduced the number of parameters from six to three.

Mutation rates in the result shown

For the results presented in the plots of Fig. 6.3 we chose the remaining parameters using the following rationale. The rates a and b are mutation–fixation rates—that is, rates at which indels become fixed in the population. As we explained in Chapter 1 (page 18), such rates are a product of the rate at which mutations occur in the population (μ_{ins} and μ_{del} for the insertions and deletions respectively) and the probability that they subsequently become fixed. We now make the plausible assumption that the *bare* insertion and deletion rates per base pair (*i.e.* μ_{ins} and μ_{del}), are equal for all spacers; the differences in the mutation–fixation rates for the differend kinds of spacers are thus assumed to be the result of differences in the selection acting on the spacers.

As in Chapter 5, we assume that the bare insertion and deletion rates are equal ($\mu_{\text{ins}} = \mu_{\text{del}}$) and that the differences between the mutation–fixation rates a and b are due to the costs of dispensable DNA. The fitness cost s per base pair is different for the different kinds of spacers, and is determined by their mean length $\bar{n}$:

$$\bar{n} = \left(\frac{b}{a} - 1\right)^{-1} = \left(\frac{\mu_{\text{del}} P_{\text{K}}(s)}{\mu_{\text{ins}} P_{\text{K}}(-s)} - 1\right)^{-1} = \left(\frac{P_{\text{K}}(s)}{P_{\text{K}}(-s)} - 1\right)^{-1} \tag{A6.31}$$

Here $P_{\text{K}}(s)$ is the Kimura–Ohta fixation probability as defined in Section 1.2.2. From this equation, s can be computed for given $\bar{n}$. Thus, from the values of $\bar{n}^{\circ}$, $\bar{n}^{\star}$, and $\bar{n}^{\bullet}$ the fitness costs s°, $s^{\star}$, and $s^{\bullet}$ can be derived. Hence, all mutation rates are determined up to the multiplicative factor $\mu_{\text{ins}} (= \mu_{\text{del}})$. This factor is eliminated by taking $b^{\bullet} = \mu_{\text{del}} P_{\text{K}}(s^{\bullet})$ as our unit of time.

Alternative choices

The reasoning above determines all mutation rates, but at the cost of some assumptions. Alternatively, one could use the three parameters that remain after fixing the mean spacer lengths as fitting parameters for the functions $\rho(x)$ and $\sigma_y(x)$. As this increases the number of free fit parameters from one (the time t_{a} in units of $b^{\bullet}$) to three (the time t_{a}, the ratio $r^{\bullet}_{\circ} \equiv b^{\bullet}/b^{\circ}$ and the ratio $r^{\bullet}_{\star} \equiv b^{\bullet}/b^{\star}$), this obviously leads to better fits than those in Fig. 6.3 (data not shown). In particular, this procedure leads to lower values for $a^{\star}$ and $b^{\star}$, partly eliminating the peak in Fig. 6.3(b). It is not clear why these rates should be lower; one possibility is that, if transcribed spacers are long, this may be because of internal regulation, which could lead to evolutionary constraints that are not included in the model.

Bibliography

Ackers, G. K., Johnson, A. D., and Shea, M. A., 1982. Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. USA*, **79**(4):1129–1133.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D., 1994. *Molecular Biology of the Cell*. Garland publishing, New Yorker, third edition.

Aranda, C., Colon, M., Ishida, C., Riego, L., Deluna, A., Valenzuela, L., Herrera, J., and Gonzalez, A., 2006. Gcn5p contributes to the bidirectional character of the uga3-glt1 yeast promoter. *Biochem Biophys Res Commun*, **348**(3):989–996.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al., 2000. Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**(1):25–29.

Baumberg, S., editor, 1999. *Prokaryotic Gene Expression*. Frontiers in Molecular Biology. Oxford University Press.

Becskei, A. and Serrano, L., 2000. Engineering stability in gene networks by autoregulation. *Nature*, **405**(6786):590–593.

Bell, P. J., Bissinger, P. H., Evans, R. J., and Dawes, I. W., 1995. A two-reporter gene system for the analysis of bi-directional transcription from the divergent mal6t-mal6s promoter in *Saccharomyces cerevisiae*. *Curr Genet*, **28**(5):441–446.

Benos, P. V., Bulyk, M. L., and Stormo, G. D., 2002. Additivity in protein-DNA interactions: how good an approximation is it? *Nucl. Acids Res.*, **30**(20):4442–4451.

Berg, J., Willmann, S., and Lassig, M., 2004. Adaptive evolution of transcription factor binding sites. *BCM Evol. Biol.*, **4**(1):42.

Berg, O. G., 1988. Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. *J. Biomol. Struc. & Dynam.*, **6(2)**:275–297.

Berg, O. G., 1990. Base-pair specificity of protein-DNA recognition: A statistical-mechanical model. *Biomed. Biochim. Acta*, **49**(8/9):963–975.

Berg, O. G., 1992. The evolutionary selection of DNA base pairs in gene-regulatory binding sites. *Proc. Natl. Acad. Sci. USA*, **86**:7501–7505.

Berg, O. G., 1996. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics*, **142**(4):1379–1382.

Berg, O. G. and Kurland, C. G., 2002. Evolution of microbial genomes: Sequence acquisition and loss. *Mol. Biol. Evol.*, **19**(12):2265–2276.

Berg, O. G. and von Hippel, P. H., 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**:723–750.

Berg, O. G. and von Hippel, P. H., 1988. Selection of DNA binding sites by regulatory proteins ii. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.*, **193**:723–750.

Berg, O. G., Winter, R. B., and von Hippel, P. H., 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, **20**(24):6929–6948.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R., 2005a. Transcription regulation by the numbers 1: Models. *Curr. Opin. Gen. & Dev.*, **15**:116–124.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R., 2005b. Transcription regulation by the numbers 2: Applications. *Curr. Opin. Gen. & Dev.*, **15**:125–135.

Blumenthal, T., 2004. Operons in eukaryotes. *Brief Funct Genomic Proteomic*, **3**(3):199–211.

Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., et al., 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**(6891):851–854.

Boer, V. M., de Winde, J. H., Pronk, J. T., and Piper, M. D. W., 2003. The genome-wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. *J. Biol. Chem.*, **278**(5):3265–3274.

Boorsma, A., de Nobel, H., ter Riet, B., Bargmann, B., Brul, S., Hellingwerf, K. J., and Klis, F. M., 2004. Characterization of the transcriptional response to cell wall stress in *Saccharomyces cerevisiae*. *Yeast*, **21**(5):413–427.

Brikun, I., Suziedelis, K., Stemmann, O., Zhong, R., Alikhanian, L., Linkova, E., Mironov, A., and Berg, D., 1996. Analysis of CRP-CytR interactions at the *Escherichia coli* udp promoter. *J. Bacteriol.*, **178**(6):1614–1622.

Bro, C., Regenberg, B., Lagniel, G., Labarre, J., Montero-Lomeli, M., and Nielsen, J., 2003. Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells. *J. Biol. Chem.*, **278**(34):32141–32149.

Browning, D. F. and Busby, S. J. W., 2004. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.*, **2**(1):57–65.

Buchler, N. E., Gerland, U., and Hwa, T., 2003. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA*, **100**(9):5136–5141.

Busby, S. and Ebright, R. H., 1994. Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell*, **79**(5):743–746.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al., 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, **433**(7025):531–537.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R., et al., 2004. The Gene Ontology Annotation (GOA) Database:

sharing knowledge in Uniprot with Gene Ontology. *Nucl. Acids Res.*, **32**(Database issue):D262–6.

Carlomagno, M. S., Riccio, A., and Bruni, C. B., 1985. Convergently functional, Rho-independent terminator in *Salmonella typhimurium*. *J. Bacteriol.*, **163**(1):362–368.

Cherry, J. L. and Adler, F. R., 2000. How to make a biological switch. *J. Theor. Biol.*, **203**:117–133.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I., 1998. The transcriptional program of sporulation in budding yeast. *Science*, **282**(5389):699–705.

Ciampi, M. S., 2006. Rho-dependent terminators and transcription termination. *Microbiology*, **152**(9):2515–2528.

Daran-Lapujade, P., Jansen, M. L. A., Daran, J.-M., van Gulik, W., de Winde, J. H., and Pronk, J. T., 2004. Role of transcriptional regulation in controlling fluxes in central carbon metabolism of *Saccharomyces cerevisiae*. A chemostat culture study. *J. Biol. Chem.*, **279**(10):9125–9138.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., and Steinmetz, L. M., 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA*, **103**(14):5320–5325.

de Daruvar, A., Collado-Vides, J., and Valencia, A., 2002. Analysis of the cellular functions of *Escherichia coli* operons and their conservation in *Bacillus subtilis*. *J. Mol. Evol.*, **55**(2):211–221.

de Hoon, M., Makita, Y., Nakai, K., and Miyano, S., 2005. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Computational Biology*, **1**(3).

Devaux, F., Marc, P., Bouchoux, C., Delaveau, T., Hikkel, I., Potier, M. C., and Jacq, C., 2001. An artificial transcription activator mimics the genome-wide properties of the yeast Pdr1 transcription factor. *EMBO Rep*, **2**(6):493–498.

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I., 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**:2381–2390.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F., 1998. Rates of spontaneous mutation. *Genetics*, **148**(4):1667–1686.

Elowitz, M. B. and Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**(6767):335–338.

Escolar, L., Perez-Martin, J., and de Lorenzo, V., 2000. Evidence of an unusually long operator for the Fur repressor in the aerobactin promoter of *Escherichia coli*. *J. Biol. Chem.*, **275**(32):24709–24714.

Farnham, P. J. and Platt, T., 1981. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. *Nucl. Acids Res.*, **9**(3):563–577.

Fields, D. S., He, Y., Al-Uzri, A. Y., and Stormo, G. D., 1997. Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, **271**:178–194.

Fleming, J. A., Lightcap, E. S., Sadis, S., Thoroddsen, V., Bulawa, C. E., and Blackman, R. K., 2002. Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by ps-341. *Proc. Natl. Acad. Sci. USA*, **99**(3):1461–1466.

Francois, P. and Hakim, Y., 2004. Design of genetic networks with specified functions by evolution *in silico*. *Proc. Natl. Acad. Sci. USA*, **101(2)**:580–585.

Fukuda, Y., Nakayama, Y., and Tomita, M., 2003. On dynamics of overlapping genes in bacterial genomes. *Gene*, **323**:181–187.

Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J., and Brown, P. O., 2001. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell*, **12**(10):2987–3003.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**(12):4241–4257.

Gerlach, P., Sogaard-Andersen, L., Pedersen, H., Martinussen, J., Valentin-Hansen, P., and Bremer, E., 1991. The cyclic AMP (cAMP)-cAMP receptor protein complex functions both as an activator and as a corepressor at the *tsx-p2* promoter of *Escherichia coli* k-12. *J. Bacteriol.*, **173(17)**:5419–5430.

Gerland, U., David Moroz, J., and T., H., 2002. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Natl. Acad. Sci. USA*, **99(19)**:12015–12020.

Gerland, U. and Hwa, T., 2002. On the selection and evolution of regulatory DNA motifs. *J. Mol. Evol.*, **55**:386–400.

Gilbert, S., 2003. *Developmental biology*. Sunderland (Massachusetts): Sinauer.

Gillespie, J. H., 1991. *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press.

Gowers, D. M. and Halford, S. E., 2003. Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO J.*, **22**(6):1410–1418.

Graber, J. H., McAllister, G. D., and Smith, T. F., 2002. Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites. *Nucl. Acids Res.*, **30**(8):1851–1858.

Graber, J. H., Salisbury, J., Hutchins, L. N., and Blumenthal, T., 2007. *C. elegans* sequences that control trans-splicing and operon pre-mRNA processing. *RNA*, **13**(9):1409–1426.

Graham, I. and Duke, T., 2005. The logical repertoire of ligand-binding proteins. *Phys. Biol.*, **2**(19):159–165.

Hajarnavis, A., Korf, I., and Durbin, R., 2004. A probabilistic model of 3' end formation in *Caenorhabditis elegans*. *Nucl. Acids Res.*, **32**(11):3392–3399.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., et al., 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**(7004):99–104.

Harris, K., Lamson, R. E., Nelson, B., Hughes, T. R., Marton, M. J., Roberts, C. J., Boone, C., and Pryciak, P. M., 2001. Role of scaffolds in MAP kinase pathway specificity revealed by custom design of pathway-dedicated signaling proteins. *Curr Biol*, **11**(23):1815–1824.

Harrison, S. C., 1991. A structural taxonomy of DNA-binding domains. *Nature*, **353**(6346):715–719.

Hermsen, R., ten Wolde, P. R., and Teichmann, S., 2008. Chance and necessity in chromosomal gene distributions. *Trends Genet.*, **24**(5):216–219.

Hornung, G. and Barkai, N., 2008. Noise propagation and signaling sensitivity in biological networks: A role for positive feedback. *PLoS Computational Biology*, **4**(1):e8–.

Huerta, A. M. and Collado-Vides, J., 2003. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J Mol Biol*, **333**(2):261–278.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al., 2000. Functional discovery via a compendium of expression profiles. *Cell*, **102**(1):109–126.

Ishida, C., Aranda, C., Valenzuela, L., Riego, L., Deluna, A., Recillas-Targa, F., Filetici, P., Lopez-Revilla, R., and Gonzalez, A., 2006. The UGA3-GLT1 intergenic region constitutes a promoter whose bidirectional nature is determined by chromatin organization in *Saccharomyces cerevisiae*. *Mol. Microbiol.*, **59**(6):1790–1806.

Istrail, S. and Davidson, E. H., 2005. Gene regulatory networks special feature: Logic functions of the genomic *cis*-regulatory code. *Proc. Natl. Acad. Sci. USA*, **102**(14):4954–4959.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T., 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.*, **16**(3):332–346.

Jacob, F. and Monod, J., 1961a. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**:318–356.

Jacob, F. and Monod, J., 1961b. On the regulation of gene activity. In *Cold Spring Harbor Symp Quant Biol*, volume 26, pages 193–211.

Jaynes, E. T. and Bretthorst, G. L., 2003. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK.

Jeeves, M., Evans, P. D., Parslow, R. A., Jaseja, M., and Hyde, E. I., 1999. Studies of the *Escherichia coli* Trp repressor binding to its five operators and to variant operator sequences. *Eur. J. Biochem.*, **265**(3):919–928.

Kabata, H., Kurosawa, O., Arai, I., Washizu, M., Margarson, S. A., Glass, R. E., and Shimamoto, N., 1993. Visualization of single molecules of RNA polymerase sliding along DNA. *Science*, **262**(5139):1561–1563.

Kalodimos, C. G., Biris, N., Bonvin, A. M. J. J., Levandoski, M. M., Guennuegues, M., Boelens, R., and Kaptein, R., 2004. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**(5682):386–389.

Kao-Huang, Y., Revzin, A., Butler, A. P., O'Conner, P., Noble, D. W., and von Hippel, P. H., 1977. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: measurement of DNA-bound *Escherichia coli* lac repressor in vivo. *Proc. Natl. Acad. Sci. USA*, **74**(10):4228–4232.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., and Karp, P. D., 2005. Ecocyc: a comprehensive database resource for *Escherichia coli*. *Nucl. Acids Res.*, **33**(suppl 1):D334–337.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. *Genetics*, **47**(6):713–719.

Kimura, M., 1979. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. USA*, **76**(7):3440–3444.

Kimura, M. and Ohta, T., 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**(3):763–771.

Kingsford, C. L., Ayanbule, K., and Salzberg, S. L., 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*, **8**(2):R22.

Koh, Y. S. and Roe, J. H., 1995. Isolation of a novel paraquat-inducible (*pqi*) gene regulated by the soxRS locus in *Escherichia coli*. *J. Bacteriol.*, **177**(10):2673–2678.

Kuhlman, T., Zhang, Z., Saier, Milton H., J., and Hwa, T., 2007. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **104**(14):6043–6048.

Kussell, E. and Leibler, S., 2005. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, **309**(5743):2075–2078.

Lagorce, A., Hauser, N. C., Labourdette, D., Rodriguez, C., Martin-Yken, H., Arroyo, J., Hoheisel, J. D., and Francois, J., 2003. Genome-wide analysis of the response to cell wall mutations in the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **278**(22):20345–20357.

Lamblin, A. and Fuchs, J., 1994. Functional analysis of the *Escherichia coli* k-12 *cyn* operon transcriptional regulation. *J. Bacteriol.*, **176**(21):6613–6622.

Lawrence, J., 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current Opinion in Genetics & Development*, **9**(6):642–648.

Lawrence, J. G. and Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA*, **95**(16):9413–9417.

Lawrence, J. G. and Roth, J. R., 1996. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**(4):1843–1860.

Lee, Y. S. and Hwang, D. S., 1997. Occlusion of RNA polymerase by oligomerization of DnaA protein over the dnaA promoter of *Escherichia coli*. *J. Biol. Chem.*, **272**(1):83–88.

Lesnik, E. A., Sampath, R., Levene, H. B., Henderson, T. J., McNeil, J. A., and Ecker, D. J., 2001. Prediction of Rho-independent transcriptional terminators in *Escherichia coli*. *Nucl. Acids Res.*, **29**(17):3583–3594.

Lipshtat, A., Loinger, A., Balaban, N. Q., and Biham, O., 2006. Genetic toggle switch without cooperative binding. *Phys Rev Lett*, **96**(18):188101.

Lisser, S. and Margalit, H., 1993. Compilation of *E.coli* mRNA promoter sequences. *Nucl. Acids Res.*, **21(7)**:1507–1516.

Liu, Y. and Xiao, W., 1997. Bidirectional regulation of two DNA-damage-inducible genes, MAG1 and DDI1, from *Saccharomyces cerevisiae. Mol. Microbiol.*, **23**(4):777–789.

Lynch, A. S. and Lin, E. C., 1996. Transcriptional control mediated by the ArcA two-component response regulator protein of escherichia coli: characterization of DNA binding at target promoters. *J. Bacteriol.*, **178**(21):6238–6249.

Madan Babu, M. and Teichmann, S. A., 2003. Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites. *Trends Genet.*, **19(2)**:75–79.

Mandel-Gutfreund, Y. and Margalit, H., 1998. Quantitative parameters for amino acid-base interactions: implications for prediction of protein-DNA binding sites. *Nucl. Acids Res.*, **26(10)**:2306–2312.

McCammon, M. T., Epstein, C. B., Przybyla-Zawislak, B., McAlister-Henn, L., and Butow, R. A., 2003. Global transcription analysis of Krebs tricarboxylic acid cycle mutants reveals an alternating pattern of gene expression and effects on hypoxic and oxidative genes. *Mol Biol Cell*, **14**(3):958–972.

Meibom, K., Sogaard-Andersen, L., Mironov, A., and Valentin-Hansen, P., 1999. Dissection of a surface-exposed portion of the cAMP-CRP complex that mediates transcription activation and repression. *Mol. Microbiol.*, **32**(3):497–504.

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.

Mnaimneh, S., Davierwala, A. P., Haynes, J., Moffat, J., Peng, W.-T., Zhang, W., Yang, X., Pootoolal, J., Chua, G., Lopez, A., et al., 2004. Exploration of essential gene functions via titratable promoter alleles. *Cell*, **118**(1):31–44.

Mossing, M. C. and Record, M. T. J., 1985. Thermodynamic origins of specificity in the lac repressor-operator interaction. adaptability in the recognition of mutant operator sites. *J Mol Biol*, **186**(2):295–305.

Müller-Hill, B., 1996. *The Lac Operon: A Short History of a Genetic Paradigm.* Walter de Gruyter, Berlin.

Müller-Hill, B., 1998. Some repressors of bacterial transcription. *Curr. Opin. Micobiol.*, **1**:145–151.

Murata, Y., Watanabe, T., Sato, M., Momose, Y., Nakahara, T., Oka, S.-i., and Iwahashi, H., 2003. Dimethyl sulfoxide exposure facilitates phospholipid biosynthesis and cellular membrane proliferation in yeast cells. *J. Biol. Chem.*, **278**(35):33185–33193.

Mustonen, V. and Lassig, M., 2005. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proc. Natl. Acad. Sci. USA*, **102**(44):15936–15941.

Omelchenko, M. V., Makarova, K. S., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V., 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ. Genome Biol*, **4**(9):R55.

Pal, C. and Hurst, L. D., 2004. Evidence against the selfish operon theory. *Trends Genet.*, **20**(6):232–234.

Pedersen, H., Dall, J., Dandanell, G., and Valentin-Hansen, P., 1995. Gene-regulatory modules in *Escherichia coli*: nucleoprotein complexes formed by cAMP-CRP and CytR at the nupG promoter. *Mol. Microbiol.*, **17**(5):843–853.

Pérez-Rueda, E. and Collado-Vides, J., 2000. The repertoire of DNA-binding transcriptioal regulators in *Escherichia coli* k-12. *Nucl. Acids Res.*, **28**(8):1838–1847.

Perocchi, F., Xu, Z., Clauder-Munster, S., and Steinmetz, L. M., 2007. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucl. Acids Res.*, :gkm683–.

Phalip, V., Kuhn, I., Lemoine, Y., and Jeltsch, J. M., 1999. Characterization of the biotin biosynthesis pathway in *Saccharomyces cerevisiae* and evidence for a cluster containing BIO5, a novel gene involved in vitamer uptake. *Gene*, **232**(1):43–51.

Postle, K. and Good, R. F., 1985. A bidirectional Rho-independent transcription terminator between the *E. coli* tonB gene and an opposing gene. *Cell*, **41**(2):577–585.

Price, M. N., Arkin, A. P., and Alm, E. J., 2006. The life-cycle of operons. *PLoS Genet*, **2**(6):e96.

Price, M. N., Huang, K. H., Alm, E. J., and Arkin, A. P., 2005a. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucl. Acids Res.*, **33**(3):880–892.

Price, M. N., Huang, K. H., Arkin, A. P., and Alm, E. J., 2005b. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, **15**(6):809–819.

Ptashne, M., 2004. *A genetic switch: phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd ed edition.

Ptashne, M. and Gann, A., 2002. *Genes and Signals.* Cold Spring Harbour Laboratory Press, NY.

Pul, U., Wurm, R., Lux, B., Meltzer, M., Menzel, A., and Wagner, R., 2005. LRP and H-NS–cooperative partners for transcription regulation at *Escherichia coli* rRNA promoters. *Mol. Microbiol.*, **58**(3):864–76.

Pupo, G., Karaolis, D., Lan, R., and Reeves, P., 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infection and Immunity*, **65**(7):2685–2692.

Resnik, P., 1998. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**:95–130.

Richet, E., 2000. Synergistic transcription activation: a dual role for CRP in the activation of an *Escherichia coli* promoter depending on MalT and CRP. *EMBO J.*, **19**:5222–5232.

Richet, E. and Sogaard-Andersen, L., 1994. CRP induces the repositioning of MalT at the *Escherichia coli malKp* promoter primarily through DNA bending. *EMBO J.*, **13**(19):4558–4567.

Richter, P. H. and Eigen, M., 1974. Diffusion controlled reaction rates in spheroidal geometry. application to repressor–operator association and membrane bound enzymes. *Biophys Chem*, **2**(3):255–263.

Riggs, A. D., Bourgeois, S., and Cohn, M., 1970. The lac repressor-operator interaction. 3. kinetic studies. *J Mol Biol*, **53**(3):401–417.

Rojo, F., 2001. Mechanisms of transcriptional repression. *Curr. Opin. Micobiol.*, **4**:145–151.

Rosenfeld, N., Elowitz, M. B., and Alon, U., 2002. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol*, **323**(5):785–793.

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J. G., Mermod, N., and Bucher, P., 2002. High-throughput selex-sage method for quantitative modeling of transcription-factor binding sites. *Nat Biotech*, **20**(8):831–835.

Sahara, T., Goda, T., and Ohgiya, S., 2002. Comprehensive expression analysis of time-dependent genetic responses in yeast cells to low temperature. *J. Biol. Chem.*, **277**(51):50015–50021.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J., 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA*, **97**(12):6652–6657.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J., 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* k-12. *Nucl. Acids Res.*, **29**(1):72–74.

Savageau, M. A., 1974. Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, **252**(5484):546–549.

Savageau, M. A., 1977. Design of molecular control mechanisms and the demand for gene expression. *Proc. Natl. Acad. Sci. USA*, **74**(12):5647–5651.

Schell, M. A., 1993. Molecular biology of the LysR family of transcriptional regulators. *Ann. Rev. Microbiol.*, **47**(1):597–626.

Sengupta, A. M., Djordjevic, M., and Shraiman, B., 2002. Specificity and robustness in transcription control networks. *Proc. Natl. Acad. Sci. USA*, **99**:2072–2077.

Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U., 2003. Detailed map of a *cis*-regulatory input function. *Proc. Natl. Acad. Sci. USA*, **100**(13):7702–7707.

Shea, M. A. and Ackers, G. K., 1985. The o_r control system of bacteriophage lambda. a physical-chemical model for gene regulation. *J. Mol. Biol.*, **181**:211–230.

Shearwin, K. E., Callen, B. P., and Egan, J. B., 2005. Transcriptional interference - a crash course. *Trends Genet.*, **21**(6):339–345.

Shen, J. and Gunsalus, R., 1997. Role of multiple ArcA recognition sites in anaerobic regulation of succinate dehydrogenase (sdhCDAB) gene expression in *Escherichia coli*. *Mol. Microbiol.*, **26**(2):223–236.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**:64–68.

Shin, M., Kang, S., Hyun, S., Fujita, N., Ishihama, A., Valentin-Hansen, P., and Choy, H., 2001. Repression of *deoP2* in *Escherichia coli* by CytR: conversion of a transcription activator into a repressor. *EMBO J.*, **20**(19):5392–5399.

Shine, J. and Dalgarno, L., 1975. Determinant of cistron specificity in bacterial ribosomes. *Nature*, **254**(5495):34–38.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, **9**(12):3273–3297.

Stormo, G. D. and Fields, D. S., 1998. Specificity, free energy and information content in protein-DNA interactions. *TiBS*, **23**:109–113.

Summers, A., 1992. Untwist and shout: a heavy metal-responsive transcriptional regulator. *J. Bacteriol.*, **174(10)**:3097–3101.

Tai, S. L., Boer, V. M., Daran-Lapujade, P., Walsh, M. C., de Winde, J. H., Daran, J.-M., and Pronk, J. T., 2005. Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **280**(1):437–447.

Tirosh, I., Berman, J., and Barkai, N., 2007. The pattern and evolution of yeast promoter bendability. *Trends Genet.*, **23**(7):318–321.

Tonks, L., 1936. The complete equation of state of one, two and three-dimensional gases of hard elastic spheres. *Physical Review*, **50**(10).

Tretyachenko-Ladokhina, V., Ross, J., and Senear, D., 2001. Thermodynamics of *E. coli* cytidine repressor interactions with DNA: Distinct modes of binding to different operators suggests a role in differential gene regulation. *J. Mol. Biol.*, **316**:531–546.

van Helden, J., del Olmo, M., and Perez-Ortin, J. E., 2000. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucl. Acids Res.*, **28**(4):1000–1010.

van Kampen, N. G., 1992. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam.

van Nimwegen, E., 2007. Finding regulatory elements and regulatory motifs: a probabilistic framework. *BMC Bioinformatics*, **8(Suppl 6)**(S4).

von Hippel, P. H. and Berg, O. G., 1986. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. USA*, **83**:1608–1612.

Wade, J., Belyaeva, T., Hyde, E., and Busby, S., 2001. A simple mechanism for co-dependence on two activators at an *Escherichia coli* promoter. *EMBO J.*, **20**(24):7160–7167.

Wagner, R., 2000. *Transcription Regulation in Prokaryotes*. Oxford University Press, NY.

Warren, P. B. and ten Wolde, P. R., 2004a. Enhancement of the stability of genetic switches by overlapping upstream regulatory domains. *Phys. Rev. Lett.*, **92**:128101 1–4.

Warren, P. B. and ten Wolde, P. R., 2004b. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J. Mol. Biol.*, **342**:1379–1390.

Warren, P. B. and ten Wolde, P. R., 2005. Chemical models of genetic toggle switches. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, **109**(14):6812–23.

Wilson, R., Urbanowski, M., and Stauffer, G., 1995. DNA binding sites of the LysR-type regulator GcvA in the gcv and gcvA control regions of *Escherichia coli*. *J. Bacteriol.*, **177**(17):4940–4946.

Winter, R. B., Berg, O. G., and von Hippel, P. H., 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli* lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry*, **20**(24):6961–6977.

Winter, R. B. and Von Hippel, P. H., 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The *Escherichia coli* lac repressor-operator interaction: equilibrium measurements. *Biochemistry*, **20**(24):6948–6960.

Xu, J. and Johnson, R., 1995. aldB, an RpoS-dependent gene in *Escherichia coli* encoding an aldehyde dehydrogenase that is repressed by Fis and activated by Crp. *J. Bacteriol.*, **177**(11):3166–3175.

Yachie, N., Arakawa, K., and Tomita, M., 2006. On the interplay of gene positioning and the role of Rho-independent terminators in *Escherichia coli*. *FEBS Lett*, **580**(30):6909–6914.

Yokobayashi, Y., Weiss, R., and Arnold, F. H., 2002. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci. USA*, **99**(26):16587–16591.

Yoshimoto, H., Saltsman, K., Gasch, A. P., Li, H. X., Ogawa, N., Botstein, D., Brown, P. O., and Cyert, M. S., 2002. Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **277**(34):31079–31088.

Yuh, C.-H., Bolouri, H., and Davidson, E. H., 1998. Genomic *cis*-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science*, **279**(5358):1896–1902.

Zivanovic, Y., Lopez, P., Philippe, H., and Forterre, P., 2002. *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. *Nucl. Acids Res.*, **30**(9):1902–1910.

Summary

A defining property of living systems is their ability to respond to signals. These signals are of a physical or chemical nature: for instance, many organism detect light intensities (*seeing*), mechanical forces (*feeling*), and the presence of certain molecules in the environment (*smelling* and *tasting*). Their responses to such clues are not arbitrary. They have evolved to allow organisms to adjust their behavior to varying environments and circumstances, and ultimately to increase their chances to survive and create offspring.

It is no surprise that *humans* can see, smell, taste and feel, nor that they adjust their behavior on the basis of such sensory information. But it is less obvious how micrometer-sized single-cellular creatures such as bacteria can obtain, weight and exploit knowledge of a multitude of physico-chemical quantities. Yet, they obviously do. Bacteria swim towards food sources and to warmer places, they synchronize their biological clocks to circadian rhythms, monitor the density of their colony, measure the osmolarity of their environment and assess which types of sugars are available.

Bacteria often have to make *logical* decisions. A famous example is the sugar utilization system in the bacterium *Escherichia coli.* In order to uptake and digest different sugars, such as glucose, lactose, galactose and arabinose, *E. coli* needs to produce particular sets of proteins that catalyze the required metabolic reactions. However, these sugars are not always present in the environment. As the production of the proteins requires an investment in terms of energy and other resources, it would be quite inefficient to produce them constitutively. Hence, *E. coli* decides when to make these enzymes and when not to, depending on the availability of the sugars. As it turns out, glucose is *E. coli*'s preferred source of energy, because it allows for the highest growth rate. Therefore, *E. coli* produces the proteins necessary for the digestion of other sugars only if these sugars are present and no glucose is found in the environment. This illustrates that the bacteria integrates several input signals (sugar availabilities) to make the decision. In this example, the decision procedure can be described by the Boolean logic function ANDN (A **AND N**ot B). In general, many of the decisions taken by cells can be categorized using the language of Boolean logic.

Cells implement many decisions at the level of *transcription.* Transcription is the molecular process by which genes (stretches of DNA containing the information required for the synthesis of one protein[6]) are copied (transcribed) to a different molecular medium: RNA. These RNA copies are produced by a multi-subunit

---

[6] Some genes actually code for a stable RNA molecule instead of a protein.

molecular machine called RNA polymerase (RNAp). Each of these RNA molecules, so-called *messenger* RNAs or mRNAs, is subsequently used as a template for the assembly of one particular protein. As a consequence, the rate at which a gene is transcribed determines (to a large extent) how many copies of the corresponding protein are produced.

The term *transcription regulation* refers to the processes used by cells to regulate the rate at which specific genes are transcribed. The resulting changes in protein concentrations are of crucial importance, because proteins determine many of the cell's properties. Many proteins serve as enzymes, which modulate the rate at which chemical processes occur in the cell; as structural elements, they constitute many of the cellular structures; and other proteins operate as tiny motors converting chemical energy to mechanical motion. In conclusion, by modulating the transcription rates of sets of genes, cells can drastically alter their protein content, and consequently their behavior and appearance.

The regulation of transcription is mediated by a special family of proteins. These proteins are called *transcription factors* and function by virtue of their ability to bind rather specifically to particular DNA sequences. Such sequences are typically located close to the starting points of genes, where RNAp initiates the transcription process. When transcription factors bind to their binding sites, they can influence the efficiency of the first steps of the transcription process and hence change the transcription rate.

The work described in this dissertation concerns with, on the one hand, the mechanisms of transcription regulation, and on the other hand, the consequences of these processes for the organization of genomes. The analyses are based on theoretical models, but published experimental data are being used to test these models and their predictions. In our work, we mainly focus on prokaryotes, featuring the bacteria *Escherichia coli* in the leading role — even though various other organisms play a supporting part.

Mechanisms of transcription regulation

Experiments have identified several mechanisms by which bacteria regulate transcription rates. Most of these mechanisms rely on fact that transcription factors are able to either recruit other molecules to the DNA or, conversely, to prevent them from binding. For instance, if a transcription factor recruits RNAp to its binding site on the DNA (called the promoter), it activates transcription. If, on the other hand, it obstructs the binding of RNAp, then transcription is inhibited. This way, the transcription rate of genes can be made to depend on the concentrations of certain (active forms of) transcription factors.

Interestingly, transcription factors need relatively simple physical properties in order to function. In bacteria, transcription factors freely diffuse through the cell; because of the small diameter of a bacterial cell (in the order of a micrometer), a protein needs about 0.1 s to diffuse from one end of the cell to the other, which makes active transport unnecessary. Transcription factors interact with the DNA and with other molecules (including other transcription factors) due to electrostatic interactions and hydrogen bonds. They can preclude another molecule (*e.g.* a transcription factor or RNAp) from binding to the DNA by binding strongly to a site that overlaps with the binding site of the another molecule.

In the Chapters 2 and 3 of this thesis, we ask what kind of functionality can in principle be obtained with these mechanisms, recruitment and hindrance, only. For this purpose, we formulate a model of transcription regulation. In this model, transcription factors can bind specifically to sites on the DNA. They can also recruit other molecules to the DNA if they bind sufficiently close together. Binding sites can overlap; molecules thus compete for binding to particular sites on the DNA. Subsequently, we use the formalism of statistical mechanics to calculate which fraction of the time molecules are bound. We use this model and an evolutionary algorithm to design transcription-regulatory systems that perform a pre-defined function and we thus explore the space of possible mechanism.

The simple mechanisms of recruitment and hindrance turn out to be immensely versatile. We show that, using rather complex distributions of binding sites, transcription regulation can perform all possible Boolean logic operations with two inputs. The functional designs often consist of modules of tandem binding sites to which transcription factors can recruit each other. This cooperative behavior leads to sharp responses of the transcription rate as a function of TF concentrations. But more intricate effects can be obtained if the modules (partly) overlap with each other, introducing competition for binding at the level of these complexes. Which module dominates in such a competition can depend strongly on the concentrations of the different transcription factors, which can be exploited to integrate different signals.

The complex designs that we describe are not unrealistic. We demonstrate that many real promoters in the bacterium *Escherichia coli* contain large numbers of transcription factor binding sites, and that overlap between these sites is extremely common. Also, transcription factors often bind to more than one binding site in a given promoter region — in exceptional cases, up to eleven sites for a single transcription factor have been documented.

Another world of possibilities is entered if we allow the systems to use feedback. In the simplest case, this means that the gene that is being regulated codes for a transcription factor that influences its own transcription rate (called auto-regulation). It has been shown that this allows for fine-tuning of the dynamical properties of these systems — *e.g.* their robustness to noisy signals or their response speed. We demonstrate that auto-regulation can also be used to achieve a more efficient repression mechanism and that it allows for alternative ways to integrate signals.

Again, the mechanisms we find are realistic. Auto-regulation is very common in *E. coli*: 59% of the transcription factors are known to bind to their own promoter region, and this is likely a lower bound since our current knowledge of regulatory interactions is far from complete. The mechanisms we discovered shed new light on the possible functions of these feedback systems, most of which are not yet elucidated.

Chromosome organization

The processes of transcription and transcription regulation also has a considerable impact on the way genes are distributed on chromosomes. To start with, all regulatory sequences, such as binding sites for RNA polymerase and transcription factors, take up space on the DNA and thereby influence the spacing between genes. Indeed, regulatory sequences directly before and after genes leave footprints on the frequency distribution of distances between genes. (Here distances are measured in base pairs.) Therefore the statistical properties of the distances between genes reveal many properties of the gene regulatory mechanisms used in the organisms.

Distances between genes

In order to study the frequency distributions of distances between genes in detail, we compare them with random models. We exploit that, mathematically, these random models are equivalent to models of one-dimensional gases. In this analogy, genes correspond to gas particles and the DNA acts as a one-dimensional, finite space.

In the most naive model, the genes are distributed completely at random. This is formally equivalent to an ideal gas. This model does not describe the data well, because genes usually do not overlap. Therefore, a second model is proposed; here we assume that genes are distributed at random, except that they do not overlap. This model is analogous to the so-called Tonks gas: a gas of hard particles in one dimension. The Tonks gas model offers a better description of the gene distributions, but fails to explain why genes have a tendency to keep a certain minimal distance from each other. This inspires the final random model, called the Constant-Force model. In the Constant-Force model, we assume that the genes are accompanied by regulatory sequence that occupy space and therefore “push” the genes apart. This leads to a picture in which the genes are distributed at random, except that they do not overlap and repel each other at short distances. The Constant-Force model provides a good fit to the gene distributions in species such as *E. coli* and *Saccharomyces cerevisiae* (Baker’s yeast).

The typical lengths of upstream and downstream regulatory regions are fit parameters of our model. This means that, by fitting our model to the distribution of genes in a particular organism, we can estimate the lengths of these upstream and downstream regulatory sequences using only the positions of genes as input. We use this to estimate lengths for various organisms.

The genomic data deviate from the Constant-Force model on several points. These deviations lead to interesting biological predictions. For instance, in most *fungi*, the distribution of distances between divergent gene pairs — neighboring genes that are transcribed from opposite DNA strands and in diverging directions—is bi-modal, strongly suggesting that their genomes contain many bi-directional promoters. Similarly, in *E. coli* we find a significant access of convergent gene pairs — neighboring gene pairs that are transcribed from opposite strands and in a converging orientation — that are unusually closely spaced; we predict that these gene pairs share a bi-directional terminator. We test all these predictions using expression data, Gene Ontology annotations and terminator predictions; the results indeed corroborate our hypotheses.

Operons

A special feature of most (if not all) prokaryotes and a few eukaryotes, is that their genes are organized in so-called *operons*. An operon is a cluster of several genes that are in one transcription unit. This means that the transcription machinery produces one long messenger RNA that contains all these genes. Genes in one operon are usually very closely spaced, and are always coded on the same strand—in so-called *tandem* orientation. As a result, the set of tandem neighboring gene pairs in such genomes consist of two populations: those pairs of genes that are in the same operon, and those that are in different operons. Correspondingly, the sequences between those genes (intergenic regions) are either inside an operon or between two operons. This division is visible in the distribution of the distances between tandem gene pairs: it is largely consistent with our random model, except that a considerable excess of gene pairs is found at short distances. Thus, the distribution of distances reveals the presence of operons.

A subject of ongoing debate is *why* genes are organized in operons. One school of thought argues that operons are used to co-regulate genes. Indeed, if several genes need to be expressed in a correlated fashion — perhaps because they have a related function — this could be achieved by placing them in one operon. Others argue that operon formation relies on horizontal gene transfer: the exchange of genes between organisms of different species. The horizontal transfer of a set of functionally interdependent genes may be more successful if they are organized in one cluster (an operon) than if they are dispersed on the genome. Therefore, operons may be “selfish” structures: their abundance could be due to their reproductive success rather than due to their added value for the organism.

These two arguments have one thing in common: they both silently assume that operons would not exist in the absence of any selective pressure to create them. In Chapter 5 we suggest quite the opposite: even if operons do not have any selective advantage (neither at the level of organisms, nor at the level of clusters of genes), they are expected to emerge. The reason is that two neighboring tandem genes are naturally in the same operon, *unless* there is a transcriptional terminator sequence in the intergenic region between them. This means that, in a sense, operons are the default: only if there is sufficient and persistent evolutionary pressure to regulate genes independently, transcriptional terminators and private promoters are expected to emerge in the course of evolution. Moreover, existing terminators are continually challenged by myriad mutations. On evolutionary time scales they will survive only if they are under constant and sufficient purifying selection. Whenever this is not the case, the terminator will be lost, and operons form immediately.

The above picture holds only for prokaryotes. In prokaryotes, all that is required to produce a new operon is the removal of a transcriptional terminator between tandem genes. In eukaryotes, this is generally not enough, because the eukaryotic ribosomes, which use the mRNA templates to assemble proteins, cannot deal with mRNAs containing more than one gene, unless the mRNA contains an internal ribosome entry site (IRES). Without such a special sequence, the ribosome translates the first gene on the mRNA only, and ignores the other genes. This explains why operons are much more rare in eukaryotes than in prokaryotes.

To prove the concept, we present a simplified model of genome evolution and developed a novel simulation scheme based on population genetics. In simulations of this model, operons and shared terminators indeed emerge spontaneously. Moreover, the model reproduces the spacing of genes in the model prokaryotes *Escherichia coli* and *Bacillus subtilis*, including the characteristic close spacing of genes in operons and the differences in spacing between convergent, divergent and tandem gene pairs. As a side effect, it also explains why promoters and terminators usually tightly flank the genes they regulate.

Evolution of intergenic distances

Intergenic regions grow and shrink due to insertions and deletions. In intergenic regions, these mutations are typically only a few base pairs long. As the occurrence of mutations is a stochastic process, one would expect that the lengths of intergenic regions perform a “random walk” on evolutionary time scales. In Chapter 6 we propose a stochastic model for this evolutionary “diffusion” of intergenic regions.

This idea can be tested using data from closely related species. If a speciation event occurs, giving rise to two different species, the intergenic distances in the two resulting species are initially expected to be equally long. However, on evolutionary time scales, random insertions and deletions will lead to a decorrelation of these distances. This process is the combined effect of the random walks performed by the two species independently; therefore we can test our model by comparing the intergenic distances of two related species.

We compare our model to the divergence of the intergenic distances in *Escherichia coli* and *Salmonella enterica subsp. enterica serovar Typhi*. We focus on intergenic regions between tandem gene pairs. As we described, such intergenic regions come in two kinds: those “between” operons and those “inside” operons. The intergenic regions in the these two groups have rather different compositions: for instance, those of type “between” contain a transcriptional terminator, whereas those “inside” do not. Also, the typical lengths of the two kinds of intergenic regions are very different. This implicates that the evolutionary diffusion of these different types is also not the same.

The main ingredients of the model are as follows. We assume that intergenic regions consist of, on the one hand, elements that have a fixed length (*e.g.* transcriptional terminators and promoters), and, on the other hand, “spacers”, whose length can change substantially. The rate at which mutations occur in spacers is assumed to depend linearly on their length, since a longer spacer contains more places where an insertion or deletion can take place. Therefore, longer intergenic regions are expected to change faster. These considerations can be formalized using Master equations and Fokker–Planck equations, which allow for quantitative predictions.

The diffusion model fits the data of *E. coli* and *S. Typhi* well. The data clearly show bigger changes in longer intergenic regions, and the diffusion of the two types of spacers is indeed different. The fit parameters also show that the divergence between *E. coli* and *S. Typhi* cannot be explained by insertions and deletions of single base pairs only. Indeed, calculations show that larger mutations can have a large influence on the speed of the evolution of the lengths of intergenic regions, even if they occur at a low rate.

The model can also be used to compute what happens if an operon splits in two, or if two operons merge. This is particularly relevant in the light of our suggestion in Chapte 5 that operons may form by merging processes that result from the loss of terminator sequences. If an operon splits in two, one intergenic region has to switch from type “inside” to type “between”. Conversely, if two operons merge, an intergenic region has to change from type “between” to type “inside”. Consequently, also the mode of diffusion of this intergenic region changes. Calculation of these processes allow for the identification of intergenic regions in which such a merging or splitting event may have taken place. We discuss these candidate regions in detail.

Samenvatting

Een van de bijzondere eigenschappen van levende wezens is dat ze kunnen reageren op prikkels. Veel organismen nemen bijvoorbeeld veranderingen waar in lichtintensiteit (*zien*), mechanische krachten (*voelen*), en de concentratie van bepaalde moleculen in de omgeving (*ruiken* en *proeven*). Ze reageren natuurlijk niet voor niets op zulke waarnemingen: deze reacties zijn gedurende miljoenen jaren van evolutie ontwikkeld zodat organismen zich kunnen aanpassen aan veranderingen in hun omgeving, en uiteindelijk een grotere kans hebben te overleven en zich voort te planten.

Het zal geen verrassing zijn dat *mensen* kunnen zien, voelen, ruiken en proeven, en ook niet dat zij hun gedrag aanpassen naar aanleiding van hun zintuiglijke waarnemingen. Maar het is een stuk minder vanzelfsprekend hoe ééncellige organismen die niet groter zijn dan een duizendste van een millimeter, zoals bacteriën, kennis over allerlei fysisch-chemische grootheden kunnen verkrijgen, afwegen en uitbuiten. Toch doen ze dat. Bacteriën zwemmen op voedsel af, zoeken een warm plekje op, zetten hun biologische klok gelijk, beoordelen de dichtheid van hun kolonie, meten de osmotische druk van hun omgeving en houden in de gaten welke soorten suikers er voorhanden zijn — om maar een paar dingen te noemen.

Transcriptieregulatie is één van de belangrijkste mechanismen die cellen gebruiken om te reageren op hun omgeving, en bovendien het onderwerp van dit proefschrift. De Hoofdstukken 2 en 3 gaan over transcriptieregulatie in bacteriën. We onderzoeken de mechanismen van transcriptieregulatie en wat hun mogelijkheden en beperkingen zijn. In de Hoofdstukken 4, 5 en 6 beschouwen we hoe transcriptieregulatie de indeling van genomen beïnvloedt. Alle analyses zijn gebaseerd op theoretische modellen, maar we gebruiken experimentele gegevens om deze modellen en hun voorspellingen te testen. We concentreren ons voornamelijk op prokaryoten (organismen die geen celkern hebben), met de darmbacterie *Escherichia coli* in de hoofdrol — maar voor verschillende andere organismen is een bijrol weggelegd.

Transcriptie en transcriptieregulatie

Cellen nemen veel van hun beslissingen op het niveau van *transcriptie*. Transcriptie is het moleculaire proces waarbij genen (stukken DNA die de instructies bevatten voor het maken van één enkel type eiwit[7]) worden gekopieerd (getranscribeerd). Dit kopieerproces wordt uitgevoerd door een moleculaire machine die RNA-

---

[7] Sommige genen coderen voor een stabiel RNA-molecuul in plaats van een eiwit.

polymerase heet (RNAp). Het kopie heeft een net andere structuur dan DNA en wordt mRNA genoemd. Elk mRNA-molecuul wordt vervolgens gebruikt als blauwdruk voor het maken van een specifiek eiwit. Daardoor bepaalt de frequentie waarmee de transcriptie van een bepaald gen plaatsvindt grotendeels hoeveel kopieën van het corresponderende eiwit in de cel aanwezig zijn.

Transcriptieregulatie is het proces waarmee cellen reguleren hoe vaak bepaalde genen worden getranscribeerd. Indirect worden hiermee dus de eiwitconcentraties in de cel gereguleerd. Dit is heel belangrijk, omdat eiwitten allerlei eigenschappen van de cel bepalen. Veel eiwitten functioneren bijvoorbeeld als enzymen, die de snelheid van chemische reacties in de cel beïnvloeden. Andere eiwitten zijn de bouwstoffen waaruit de meeste structuren in de cel zijn opgebouwd. En weer andere eiwitten werken als minuscule motortjes die chemische energie omzetten in mechanische beweging. Kortom, door de transcriptiesnelheid van groepen van genen te reguleren, kunnen cellen hun samenstelling drastisch aanpassen en daardoor ook hun gedrag en vorm.

Bij de regulatie van transcriptie speelt een speciale familie van eiwitten een grote rol. Deze eiwitten heten *transcriptiefactoren*. Ze functioneren doordat ze kunnen binden aan specifieke stukken DNA. Deze “parkeerplaatsen” zijn meestal vlak bij de beginpunten van genen te vinden, daar waar de RNAp begint met het transcriptieproces. Wanneer transcriptiefactoren binden aan hun bindingsplaatsen kunnen ze de efficiëntie van de eerste stappen van het transcriptieproces beïnvloeden en daardoor ook de frequentie veranderen waarmee transcriptie plaatsvindt.

Logische beslissingen

Bacteriën moeten vaak *logische* beslissingen nemen. Een beroemd voorbeeld is de manier waarop *E. coli* zijn maaltijd kiest. *E. coli* haalt zijn energie uit suikers; maar om de verschillende soorten suikers, zoals glucose, lactose, galactose en arabinose te kunnen opnemen en verteren, moet hij een aantal eiwitten (enzymen) produceren die de benodigde stofwisselingsreacties stimuleren. De verschillende suikers zijn meestal niet allemaal aanwezig in de omgeving van de bacterie. Omdat de productie van de eiwitten onder andere energie kost, zou het niet erg efficiënt zijn om ze continu aan te maken. Daarom beslist *E. coli* afhankelijk van de beschikbaarheid van de suikers welke enzymen hij wel of niet wil produceren.

In feite gaat *E. coli* nog een stapje verder. Omdat hij het snelst kan groeien als hij glucose eet, is dat zijn favoriete maaltijd. Daarom produceert *E. coli* de eiwitten die nodig zijn voor de vertering van, zeg, lactose, enkel als er lactose beschikbaar is en geen glucose. Dit laat zien dat de bacterie meerdere gegevens — de beschikbaarheid van de verschillende suikers — betrekt bij zijn beslissing. De besluitprocedure is, kort gezegd: maak eiwitten LacY, LacZ en LacA alleen als er lactose is EN NIET glucose. De relatie EN NIET is een voorbeeld van een Boolese functie (zie Kader 1 voor meer uitleg).

Deze besluitprocedure wordt met behulp van transcriptieregulatie uitgevoerd. Dat werkt als volgt. *E. coli* bevat een speciale transcriptiefactor, genaamd LacI.

Kader 1: Boolese logica en logische poorten

In veel wetenschapsgebieden wordt een speciale indeling van logische redeneringen gebruikt. Deze wordt de Boolese logica genoemd, naar de 19e-eeuwse Britse wiskundige en filosoof George Boole, die er de grondlegger van is.

We kunnen de Boolese logica introduceren aan de hand van een paar voorbeelden. Sommige beslissingen zijn van het type A EN B. Hier staan A en B voor willekeurige uitspraken. Een goed voorbeeld is als iemand zegt: “Ik ga alleen zeilen als het zonnig weer is en het bovendien waait”. Als we de uitspraak “Het is zonnig weer” nu A noemen en de uitspraak “Het waait” aanduiden met B, dan heeft de beslissing om te gaan zeilen inderdaad de vorm “Ik ga zeilen als A EN B”. Een andere uitspraak kan zijn “Ik ga alleen fietsen als het zonnig weer is en het bovendien *niet* waait”. Deze zin is van het type A EN NIET B. Op deze manier bestaat er een hele verzameling aan Boolese functies, waaronder EN, OF, EN NIET, OF NIET, etcetera.

Een apparaatje dat een van de Boolese beslissing kan uitvoeren, heet een *logische poort*. Zulke logische poorten vormen de basis van alle digitale elektronica. Het voorbeeld in de tekst over de maaltijdkeuze van *E. coli* laat zien dat bacteriën blijkbaar een mechanisme bevatten dat kan functioneren als een logische poort. Eigenlijk is de bacterie dus een soort analoog computertje.

Deze transcriptiefactor bindt in de afwezigheid van lactose op een speciale plek aan het DNA. Daardoor verhindert hij de transcriptie van de genen die coderen voor de eiwitten LacY, LacZ en LacA; deze eiwitten worden in afwezigheid van lactose dus niet geproduceerd. Maar, als er wel lactose in de omgeving is, dan bindt lactose aan de transcriptiefactor LacI[8]. LacI verandert hierdoor van vorm; in deze vorm bindt het veel slechter aan het DNA en blokkeert het de transcriptie niet langer. De transcriptiefactor CRP doet iets soortgelijks, maar dan voor glucose: CRP activeert transcriptie als er geen glucose aanwezig is. Op deze manier worden de lactose-genen alleen gebruikt als er lactose is maar geen glucose.

Mechanismen van transcriptieregulatie

Door middel van ingewikkelde experimenten zijn een aantal mechanismen ontdekt die bacteriën gebruiken om de transcriptiesnelheid te reguleren. De meeste van deze mechanismen werken doordat transcriptiefactoren andere moleculen helpen bij hun binding aan het DNA, of precies andersom, ze daarbij hinderen. Bijvoorbeeld, als een transcriptiefactor de RNAp helpt bij het binden aan zijn bindingsplaats (deze wordt de *promoter* genoemd), dan wordt de transcriptie geactiveerd. Als, aan de andere kant, de transcriptiefactor de binding van RNAp blokkeert, dan wordt transcriptie onderdrukt.

In de Hoofdstukken 2 en 3 van dit proefschrift bestuderen we welke types

---

[8] Dit is niet helemaal waar: eigenlijk bindt niet lactose aan LacI, maar allolactose. *E. coli* zet lactose om in allolactose.

Kader 2: Evolutionaire algoritmes

Biologische evolutie is het samenspel van mutaties en (natuurlijke en seksuele) selectie. Mutaties zorgen er voor dat verschillende individuen in een populatie niet precies hetzelfde genoom hebben. Individuen die door hun genoom beter in staat zijn zich voort te planten dan anderen, krijgen gemiddeld meer nakomelingen; genen die de voortplantingskansen verhogen hebben dus grote kans om in de loop van de tijd in steeds grotere aantallen voor te komen. Door miljoenen rondes van mutaties en selectie kunnen zo organismen ontstaan die heel goed zijn aangepast op hun omgeving. In de natuur heeft het proces van evolutie in de loop der miljoenen jaren heel geavanceerde organismen opgeleverd.

Een evolutionair algoritme is een type computerprogramma dat een evolutionair proces nabootst om oplossingen te vinden voor een bepaald complex (ontwerp)probleem. De computer slaat eerst een groot aantal willekeurige ontwerpen in zijn geheugen op. Die eerste, willekeurige ontwerpen zijn in het algemeen heel slechte oplossingen voor het betreffende probleem. Vervolgens kiest het programma de beste ontwerpen uit, kopieert ze een aantal keer, en brengt er hier en daar willekeurige wijzigingen in aan. Sommige ontwerpen zijn er waarschijnlijk slechter op geworden, maar een paar zijn wellicht iets verbeterd. Weer selecteert het programma de beste ontwerpen. Na heel veel rondes van muteren en selecteren kan zo een heel geavanceerde ontwerp worden geconstrueerd. Een leuk aspect aan deze methode is dat je op deze manier iets ingewikkelds kunt ontwerpen, zonder dat je de oplossing zelf hoeft te bedenken.

Wij gebruikten een evolutionair algoritme om DNA-sequenties te vinden die als logische poorten kunnen functioneren. Het algoritme vond vaak oplossingen die we zelf nog niet hadden bedacht.

beslissingen in principe kunnen worden geïmplementeerd met de mechanismen die bekend zijn. Om dat uit te zoeken, formuleren we een kwantitatief model van transcriptieregulatie. We combineren dit model met een evolutionair algoritme basale om transcriptieregulatiesystemen te ontwerpen die een door ons gekozen functie vervullen. Zo kunnen we het scala aan mogelijke ontwerpen verkennen. In Kader 2 is meer te lezen over evolutionaire algoritmes.

Het blijkt dat de eenvoudige mechanismen van transcriptieregulatie enorm veelzijdig zijn. Met behulp van vrij complexe patronen van bindingsplaatsen kunnen alle mogelijke logische beslissingen met twee input-signalen worden geïmplementeerd. De beste ontwerpen bestaan uit modules van bindingsplaatsen die allemaal direct naast elkaar liggen. De transcriptiefactoren die aan deze plaatsen binden, helpen elkaar bij het binden. Dit coöperatieve gedrag leidt tot een scherpe reactie van de transcriptiefrequentie als functie van de concentraties van de transcriptiefactoren. Meer geavanceerde effecten kunnen worden bereikt als de modules (gedeeltelijk) met elkaar overlappen. Dat introduceert competitie op het niveau van modules, die immers niet tegelijk gebonden kunnen zijn. Welke module domineert, kan in zulke situaties sterk afhangen van de concentraties van de verschillende transcriptiefactoren. Dit kan worden uitgebuit om verschillende signalen tegen elkaar af te wegen.

De mogelijkheden worden nog meer vergroot als we terugkoppeling (feedback) toelaten. In het meest eenvoudige geval codeert het gereguleerde gen voor een transcriptiefactor die op zijn beurt zijn eigen transcriptiefrequentie reguleert. Dit heet auto-regulatie. Het is aangetoond dat dit het mogelijk maakt om de dynamische eigenschappen van de systemen af te stellen — bijvoorbeeld, de gevoeligheid voor ruis of de reactie-snelheid. Onze resultaten tonen aan dat auto-regulatie ook een efficiënter repressiemechanisme mogelijk maakt en alternatieve mechanismen biedt voor het integreren van signalen. De mechanismen die we vinden werpen een nieuw licht op de mogelijke functies van feedback-systemen in transcriptieregulatie.

Chromosoom-organisatie

De processen van transcriptie en transcriptieregulatie hebben een grote invloed op de manier waarop genen verdeeld zijn over chromosomen. Alle stukken DNA die een rol spelen in transcriptieregulatie, zoals de bindingsplekken voor RNAp en transcriptiefactoren, nemen plaats in beslag op het DNA. Daardoor beïnvloeden ze de afstanden tussen genen. Inderdaad zien we dat deze sequenties direct vóór en na genen hun sporen achterlaten op de kansverdelingen van afstanden tussen genen. Omgekeerd onthullen de statistische eigenschappen van de afstanden tussen genen allerlei informatie over de regulatiemechanismen die door het organisme worden gebruikt.

Afstanden tussen genen

Om de kansverdelingen van de afstanden tussen genen goed te bestuderen, vergelijken we ze met modellen. We maken er gebruik van dat, wiskundig gezien, deze modellen precies overeenkomen met modellen van één-dimensionale gassen. In deze analogie komen genen overeen met gasdeeltjes en het DNA speelt de rol van een één-dimensionale, eindige ruimte.

Het beste model is het Constantekrachtmodel. In dat model nemen we aan dat de genen worden vergezeld door sequenties ten behoeve van de regulatie die plaats innemen en daarom de genen als het ware uit elkaar houden. De genen zijn willekeurig verdeeld, behalve dat ze zelden overlappen en elkaar op korte afstanden “afstoten”. Dit model komt erg goed overeen met de verdelingen in organismen zoals *E. coli* en *Saccharomyces cerevisiae* (bakkersgist).

De genoomdata wijken op verschillende punten af van het Constantekrachtmodel. Deze afwijkingen leiden tot interessante biologische voorspellingen. Bijvoorbeeld, in de meeste *schimmels* heeft de kansverdeling van afstanden tussen divergente genen — naburige genen die in tegengestelde richting en in divergente oriëntatie worden afgelezen — twee pieken, wat sterk suggereert dat deze genomen veel bi-directionele promoters bevatten. Net zoiets is het geval in *E. coli*: we vinden een flink overschot aan convergente gen-paren — naburige genen die in te-

Kader 3: Promoters en terminators, hoofdletters en punten

Je kunt je DNA voorstellen als een lange reeks letters, net als een tekst. In die vergelijking is een gen zoiets als een *zin*.qpenvaosvkHet verwarrende is dat tussen de genen op het DNA ook veel letters staan die geen betekenis hebben. Bovendien kunnen genen ook ondersteboven worden geschreven. Dit is allemaal niet zo'n probleem, omdat er codes op het DNA staan die vertellen waar een gen begint en waar hij eindigt. Het beginsignaal heet een *promoter* en kan het best vergeleken worden met een hoofdletter, die immers het begin van een zin aangeeft. Het eindsignaal is een *terminator* en heeft dezelfde functie als een punt.

Bij het bekijken van de afstanden tussen genen van *schimmels* vonden we aanwijzingen dat veel promoters bi-directioneel kunnen zijn. Dat wil zeggen dat ze twee kanten op functioneren. Zo zou je dat kunnen illustreren. Zoals u ziet, gebruikt deze zin dezelfde hoofdletter als de vorige zin. Dat kan natuurlijk alleen maar omdat beide zinnen met een heel speciale letter beginnen die ook op z'n kop bruikbaar is.

In *Escherichia coli* vinden we juist sterke aanwijzingen voor veel bi-directionele terminators. Deze zin gebruikt dezelfde punt als de vorige zin. Dat kan omdat een punt er omgekeerd precies hetzelfde uitziet als rechtop. Dat geldt ook voor bi-directionele terminators; deze sequenties zijn bij benadering palindromen (sequenties die symmetrisch zijn, waardoor je ze in twee richtingen kunt lezen, zoals het woord "meetsysteem").

gengestelde richting en in convergente oriëntatie worden afgelezen — die bijzonder dicht bij elkaar verblijven; we voorspellen dat deze gen-paren een bi-directionele terminator delen (zie Kader 3).

Operons

Een speciale eigenschap van de meeste (zo niet alle) prokaryoten en een paar eukaryoten is dat hun genen zijn georganiseerd in zogenaamde *operons*. Een operon is een kluster van genen die samen worden getranscribeerd tot één mRNA. Genen in een operon bevinden zich meestal heel dicht bij elkaar en worden ook in dezelfde richting afgelezen; ze hebben een zogenaamde *tandem* oriëntatie. Daardoor bestaat de verzameling van tandem gen-paren in zulke genomen uit twee groepen: de gen-paren die in hetzelfde operon zitten, en de paren die in verschillende operons zitten. De sequenties die tussen deze genen in liggen bevinden zich daardoor ofwel *in* een operon of *tussen* twee operons. Deze tweedeling is ook zichtbaar in de verdeling van afstanden tussen tandem genen: deze is grotendeels consistent met ons model, behalve dat een duidelijk overschot aanwezig is op korte afstanden. Door dat overschot verraadt de verdeling van afstanden de aanwezigheid van operons.

De vraag *waarom* genen zijn georganiseerd in operons, is een onderwerp van continu debat. De meningen zijn grofweg verdeeld in twee kampen. Het eerste kamp betoogt dat operons gebruikt worden om genen te co-reguleren. Als een aantal genen in een gecorreleerde manier tot expressie moeten worden gebracht — bijvoorbeeld omdat ze een gerelateerde functie hebben — dan kan dit inderdaad worden bewerkstelligd door ze in één operon te plaatsen. Het andere kamp is van mening dat de formatie van operons het gevolg is van de “horizontale overdracht” van genen: het feit dat genen soms worden overgebracht van het ene naar het andere organisme. Operons die verschillende of zelfs alle benodigde genen bevatten voor een bepaalde functie, zouden een grotere kans kunnen hebben om succesvol te worden overgebracht naar andere organismen dan losse genen. Daarom zouden operons “egoïstische” structuren kunnen zijn: hun bestaan zou dan het gevolg zijn van hun succesvolle verspreiding en niet zozeer van hun toegevoegde waarde voor het organisme.

De twee genoemde argumenten hebben een ding gemeenschappelijk: ze nemen beiden stilzwijgend aan dat operons niet zouden bestaan in afwezigheid van enige selectiedruk om ze te creëren. In Hoofdstuk 5 beargumenteren wij precies het omgekeerde: zelfs als operons geen enkel selectief voordeel met zich meebrengen—noch op het niveau van het organisme, noch op het niveau van het operon zelf—dan nog zijn operons te verwachten. De reden is dat twee tandem buurgenen van nature in hetzelfde operon zijn, *tenzij* er zich tussen hen in een terminatorsequentie bevindt. Dit betekent dat, in zekere zin, operons de “default”-indeling zijn: alleen als er voldoende en aanhoudende evolutionaire druk bestaat om de genen onafhankelijk te reguleren, kan men verwachten dat in de loop van de evolutie terminators en onafhankelijke promoters ontstaan. Tegelijkertijd worden bestaande terminators continu op de proef gesteld door allerlei soorten mutaties. Op evolutionaire tijdsschalen zullen zij enkel overleven als ze constant onder voldoende selectiedruk staan. Wanneer dat niet het geval is, zal de terminator verdwijnen en een operon ontstaan.

Om aan te tonen dat dit concept hout snijdt, presenteren we een eenvoudig model voor de evolutie van genomen en ontwikkelden we een nieuw simulatieschema gebaseerd op het wetenschapsgebied van de populatie-genetica. In simulaties van dit model ontstaan inderdaad spontaan operons en gedeelde terminators. Bovendien reproduceert het model de verdeling van genen in de organismen *E. coli* en *Bacillus subtilis*, inclusief de karakteristieke clustering van genen in operons en de verschillen in de afstanden tussen convergente, divergente en tandem genparen. *En passant* verklaart het ook waarom promoters en terminators zich in het algemeen heel dicht bij het bijbehorende gen bevinden.

De evolutie van afstanden tussen genen

Op evolutionaire tijdsschalen groeien en krimpen de afstanden tussen genen als het gevolg van invoegingen en verwijderingen van stukjes DNA. In de regio’s tussen genen zijn deze stukjes typisch heel kort. Aangezien het vóórkomen van deze mutaties een kansproces is, zou je verwachten dat de lengtes van deze regio’s op de lange duur een zogenaamde “random walk” beschrijven. In Hoofdstuk 6 beschrijven we een stochastisch model voor deze evolutionaire “diffusie” van sequenties tussen genen.

Dit idee kan getest worden met behulp van gegevens over gerelateerde organismen. Direct nadat twee soorten ontstaan uit een gezamenlijke voorouder, zouden de regio's tussen genen in beiden organismen even lang moeten zijn. Maar in de loop der tijd zullen invoegingen en verwijderingen ervoor zorgen dat de afstanden gaan verschillen. We kunnen ons model dus testen door de afstanden tussen genen in twee gerelateerde organismen te vergelijken met berekeningen aan het model. We vergelijken ons model met de gegevens van *Escherichia coli* en *Salmonella enterica subsp. enterica serovar Typhi*.

Het model kan ook worden gebruikt om te berekenen wat er gebeurt als een operon opsplitst of wanneer twee operons samensmelten. Door berekeningen aan dit proces kunnen we regio's opsporen waarin wellicht recentelijk een samenvoeging of splitsing heeft plaatsgevonden.

Acknowledgments

The final chapter of my thesis is also the final chapter of my work at AMOLF and my life in Amsterdam. Sadly, it means I'll have to say goodbye to many great colleagues and friends.

The first one to thank is, of course, Pieter Rein. For me, PR has been an ideal promotor. Few people are so genuinely eager to learn and understand as PR. He likes to share his knowledge, wisdom and experience, but seems to enjoy learning from his students even more. PR gave me total freedom to define my own projects, but at the same time was always available for a brainstorm, crucial insights or valuable suggestions. Apart from an excellent supervisor, PR is also a great person. I truly enjoyed our chats about topics ranging from politics to boats, from books to nostalgia, and from his kids to music. And if, just before a deadline, the hard disk of my laptop would crash or some other technical problem occurred, PR would not leave the building before we solved it together.

I am also greatly indebted to Sarah. The several visits in Cambridge were a great experience for me—both scientifically and personally—and have resulted in a nice publication (Chapter 4). Your knowledge of and insight into biological data have given me the opportunity to make a lot of progress in a very short time.

Harald! For several years, we were colleagues both at AMOLF and as board members of the “Nationale DenkTank”. We worked hard to make “De Nationale DenkTank” a reality. I checked: since June 2004, each of us sent the other one email per day on average—even though we walked into each others office all the time. Yesterday, the third DenkTank project started and I'm sure it will be just as successful as the previous episodes. I'm grateful of the things I learned from you and of the things we learned together. Thanks!

Interacting with the other group members was a pleasure as well. Part of the work presented in Chapter 3 was performed by Bas Ursem as part of his internship in our group. Bas, thanks for your enthusiasm; I'm sure our project will turn out to be a great success! Siebe, I really hope and trust that you and Siri will have a great time in Ålesund. Christian, thanks for the good times we shared; can I frame one of your great pictures of Amsterdam when I move abroad? Also many thanks to Marco, Jeroen, Rosalind, Pim, Sanne, Thorsten, Wiet and Filipe.

The friendly atmosphere at the “overloop” part of AMOLF should, I think, largely be attributed to the other group leaders, who set the tone: Bela, Daan, Sander, Marileen, Harm-Geert and Gijsje. I have always liked the cross-fertilization between these groups and have profited directly from the expertise of most of you.

Sander collaborated with us on Chapter 2. Bela was the first I would go to with questions about foundations of maths, statistical physics, notation, LaTeX — or any other issue that required a playful but "keen philosophical" mind (as he would call it, tongue-in-cheek). When it comes to simulation techniques (or any other topic, for that matter), there's no better oracle than Daan. And coffee-corner conversations with Harm-Geert always turned into an amazing scientific roller coaster.

In several stages, Dennis Bray, Nick Buchler, Rosalind and Frank were so kind as to critically read the various manuscripts; thank you for your useful suggestions!

Thanks also to Simon. Science, the "DenkTank", photography — we always had enough common interests to chat about. A number of discussions with you have definitely influenced Chapters 2 and 5.

To my office mates (Matt, Fabiana, Pep, Sanne, Ana, Jan-Willem): thanks for the great working atmosphere! You all helped me in many ways. I hope you did not suffer too much from the many phone calls during the DenkTank periods; I again apologize! And I admire Sanne and Ana for their patience with my typesetting obsession.

I also thank the people that made the "Nationale DenkTank" possible. To begin with, Liedewij, Simon, Yves, Ruud, Harald and Stefan, who conceived and co-authored the initial plan. A bit later, all board members: Harald and Stefan again, and Bart, Paul, Wouter, Edda, Claire, Jurjen, Pieter, Lindy, and Nouchka. Thanks to Rolf and Pris too, for their hard work and good company. I am also grateful for the support at AMOLF, mainly thanks to Albert Polman and Piet Kistemaker.

Some other people that contributed strongly to the AMOLF experience: Niels (hiking trips!), Rhoda (tuinkabouter!), Laura, Paige (♪), Daan, Koos, Julien (eBooks!), Aileen, Behnaz (Persian food & ditto graphic design!), Benjamin, Bianca, Chantal, Eva, Gertjan, Kostya, Izabela, Live, Jerien, Manju & Sunitha (what a memorable diner!), Marina, Marjon, Nefeli, Nienke, Patrick, Svenja, Ymkje — it's impossible to list all of you!

I owe much, if not all, to my parents, who have supported me throughout my school and studies, in all possible ways. Pa, Louise, Mam, Djien-Gie, Eline, Floris: I have the paradoxical tendency to neglect the people that are most dear to me; I hope some of my love shines through nonetheless.

And last but not least, I want to express my love and gratitude to Alette, who had to compete with a computer screen for far too long. You're the reason why, every day, I fall asleep with a smile on my face.

Curriculum vitae

Rutger Hermsen was born in Leiderdorp, The Netherlands, on July 14, 1978. He grew up in Koudekerk aan den Rijn and initially attended the Stedelijk Gymnasium Leiden. He relocated to Soest in 1994 and graduated from the Johan van Oldenbarneveltgymnasium in Amersfoort in 1996.

Subsequently, he studied physics, mathematics and philosophy at the University Utrecht. He passed his propaedeutic exam in mathematics *cum laude* in 1997 and obtained his M. Sc. degrees in the Philosophy of Science and Theoretical Physics in 2003. His first Masters thesis was supervised by Prof. Dr. Dennis Dieks, from the Institute for History and Foundations of Science, and Dr. Lev Beklemishev, at the Department of Philosophy. This thesis dealt with the connections between physics and the mathematical notion of computability. A second Masters thesis, in Theoretical Physics, was supervised by Dr. René van Roij at the Institute for Theoretical Physics and presented calculations on interfacial phenomena in hard-rod fluids.

Starting from 2004, Rutger worked as a Ph.D. student in the Biochemical Networks Group of Prof. Dr. Pieter Rein ten Wolde, at the FOM institute AMOLF in Amsterdam. The results of this work are presented in this dissertation.

In the beginning of 2005, Rutger was a co-founder of the foundation “De Nationale DenkTank” (The National Think Tank). Every year, this foundation selects 25 talented students and young professionals to participate in a newly formed National Think Tank. This multidisciplinary team spends three months studying a complex societal issue in search for creative solutions. Together, the Think Tanks of the consecutive years form an active network and a hotbed for new ideas and initiatives.