FITNESS LANDSCAPES OF GENE REGULATION IN VARIABLE ENVIRONMENTS

Fitness landscapes of gene regulation in variable environments

ACADEMISCH PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus prof. dr. D.C. van den Boom,
ten overstaan van een door het college voor promoties
ingestelde commissie, in het openbaar te verdedigen
in de Agnietenkapel
op donderdag 29 mei 2008, te 12:00 uur

door

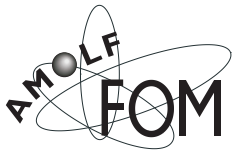
FRANCISCUS JACOBUS POELWIJK

geboren te Amsterdam

Promotiecommissie:

Promotor:	Prof. dr. D. Frenkel
Copromotor:	Dr. ir. S. J. Tans
Overige leden:	Prof. dr. R. Boelens
	Prof. dr. A. M. Dean
	Prof. dr. K. J. Hellingwerf
	Prof. dr. P. H. van Tienderen
	Dr. J. A. G. M. de Visser

Faculteit der Natuurwetenschappen, Wiskunde en Informatica



Nederlandse titel: Fitness-landschappen van genregulatie in een variabele omgeving.

The work described in this thesis was performed at the FOM Institute for Atomic and Molecular Physics (AMOLF) in Amsterdam, The Netherlands. This work is part of the research programme of the 'Stichting voor Fundamenteel Onderzoek der Materie (FOM)', which is financially supported by the 'Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)'.

ISBN-13: 978-90-9023059-7
ISBN-10: 90-9023059-9

A digital version of this thesis can be obtained from `http://www.amolf.nl`. Printed copies can be obtained by request via email to `library@amolf.nl` or by addressing the library at the FOM Institute for Atomic and Molecular Physics (AMOLF), Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands.

Cover design by Frank Poelwijk. The image represents local fitness environments along an evolutionary path (chapter 3). Printed in the Netherlands by Ponsen & Looijen BV graphical company, Wageningen. Typeset with LaTeX.

When a coin is tossed, it does not necessarily fall heads or tails;
it can roll away or stand on its edge.

William Feller, *An Introduction to Probability, vol. I.*

Contents

CHAPTER

1 Introduction

All organisms experience fluctuations in their environment. In response they use regulatory mechanisms, either to attempt to maintain a *status quo*, or to achieve an appropriate change. Regulation has a broad range of manifestations: from the maintenance of constant body temperature in warm-blooded animals, timing and spatial control of embryo morphology, to circadian rhythms which are found across all domains of life. Clearly, regulation is not just superimposed on organisms. It is a product of evolution and itself shaped by fluctuating selective pressures. Therefore a regulatory system is not only a 'device' that realizes its present function, but also a reflection of its evolutionary history. In fact, neither aspect of it can be fully understood in isolation. In this thesis we focus on the evolution of a functionally well-studied bacterial regulatory system that allows *Escherichia coli* to respond to fluctuating levels of a metabolizable sugar. Regulation in this case, as for many bacterial responses, is achieved at the level of gene expression. We addressed issues concerning its (non-)optimality and the specific selective conditions that favor or disfavor regulation. Under artificial fluctuating selective pressures we investigated adaptation towards novel regulatory phenotypes at a quantitative and molecular level. This introduction will give a brief, sometimes historical, overview of the main approaches and concepts used throughout this thesis, such as laboratory evolution experiments and fitness landscapes. The text outlines some of the current questions in the field and the progress that is being made toward answering them.

1.1 Laboratory experiments on evolution

The main advantage of laboratory experiments (or at least controlled experiments) on evolution is that the selective pressures at work are known to a much greater extent than they are in nature. For many evolutionary questions this considerably simplifies

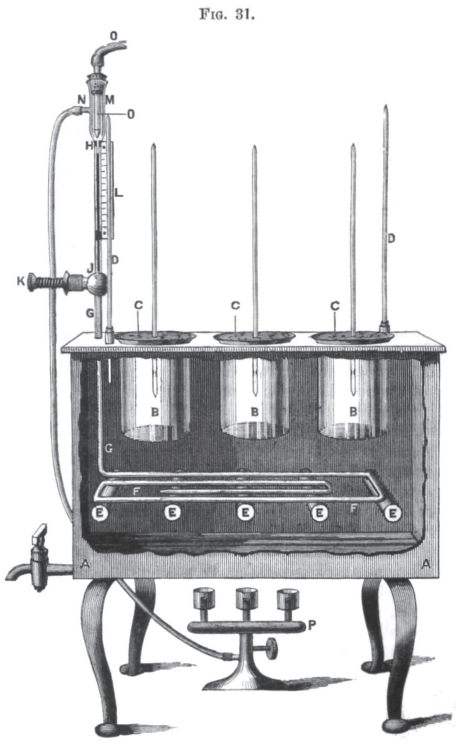


Figure 1.1: Image of the thermostat in which W.H. Dallinger performed his heat-adaptation experiments in the 1880s, taken from the original report [1]. The device consisted of three glass containers (B) kept in a water bath heated by a gas burner (P). A glass tubing containing mercury (F) was located in the water bath and ended in a vertical tube near (H). When the mercury expanded due to heating of the water bath, the mercury meniscus rose and decreased the gas supply (via O) to the burner. The required temperature could be set by screwing the plunger (K) into a mercury reservoir (J). Dallinger reported the temperature to be stable within 0.1^{o}C.

matters: one does not have to make a 'deconvolution' of the influence of unknown selective conditions. Of course, this implies that important classes of evolutionary issues – specifically those investigating natural selective pressures, but also in general systems that are too large, too slow, or too complex to experiment on – are out of reach. Still, experimental evolution has proven a powerful approach to understand evolutionary processes: short-term experiments can yield estimates for genetic variation and heritability, whereas long-term experiments are a powerful tool to uncover the potential and limits of evolutionary adaptation.

One of the first and at the same time one of the most extraordinary selection experiments was reported by the reverend W. H. Dallinger in 1887 in his address to the Royal Microscopical Society [1]. In the preceding years Dallinger had built an ingenious apparatus (fig. 1.1), which contained three glass vessels kept in a thermostat that

was stable to within almost a tenth of a degree centigrade. The three vessels contained "putrefactive fluids and organisms" that at the time were a rewarding object of study for microscopists. During a six year experiment in which he focused on three species of unicellular flagellate eukaryotes that were most familiar to him, Dallinger raised the temperature of the thermostat by small increments from its initial 16^{o}C to an astonishing 70^{o}C. Although at that point a not further specified catastrophic event destroyed his device, Dallinger's intent "to observe critically how far changes in the organisms led to responsive adaptations and successive survival" had been fulfilled to a surprising level. Even just prior to the destruction, Dallinger had managed to perform the appropriate 'reciprocal transplants': not only did the non-heat-adapted organisms die at elevated temperatures, but also the organisms that could survive 70^{o}C were killed when placed in a nutritive liquid at 16^{o}C.

After Dallinger's experiment, whose clarity of approach and conclusions seem almost like an anachronism, selection experiments have been performed on organisms throughout all domains of life. For instance, extensive selection for morphological changes in water fleas was performed by Woltereck [2], and around 1910 *Drosophila* became established as an important organism to study evolution (e.g. [3]). In the early days (see [4]) one of the main questions was whether selection is able to produce permanent changes outside the ranges of naturally occurring variation. Somewhat later, studies focused on *what* changed under selection – the gene under consideration itself or genes at other loci that determine whether it is expressed. A third important area of exploration were the limits of the response to selection [5]. One study should be mentioned here for its sheer duration and its remarkable results: the current record holder for duration, the Illinois Long-Term Selection Experiment for grain protein and oil concentration in maize (*Zea mays*) started in 1896 and is still running [6]. Even after about 110 generations, lines selected for high protein and high oil contents have not yet plateaued, while selection for low oil content has hit the 0% limit. As this study has made it into the genetics era, it is a rich source of information on plant evolution. This experiment has an obvious relevance not only for evolutionary adaptation itself, but also for studying a key event in early human civilization.

Surprisingly, after the initial selection experiments on microbes, the combination evolution and micro-organisms (particularly bacteria) has been unpopular for most of the twentieth century. One of the main causes was their small size and relative lack of features, which hampered the creation of a microbial phylogeny [7]. Microbiology became a field that was almost void of evolutionary thinking. It was not until an appreciable level in nucleic acid sequencing technology [8, 9] was reached by the 1980s that renewed consideration was given to microbes in evolution, and that the advantages of working with organisms whose "cycle of life is so relatively short, and [whose] generations succeed each other so rapidly", as Dallinger had put it [1], were re-realized. Apart from their short generation times, bacteria are small (up to 10^9 in a ml), clonal, easy to grow (which also creates possibilities for parallel lines), easy to store (many years in

a -80^{o}C freezer) and amenable to artificial mutation. This explains their rise as model systems in evolutionary studies, in which their adaptation to artificial environments and population dynamics [10, 11] were explored, as well as the influence of mutation rates [12], and the interactions with bacteriophages [13], to name a few. The rigorous control the researcher can exert over the environment is a reason for the more recent popularity of microbial systems in the study of ecological questions [14]: for example diversification [15, 16], predator-prey dynamics [17], and the maintenance of genetic variation [18] are studied, often in a mathematical framework.

Clearly, micro-organisms, and especially bacteria, being unicellular haploid and mainly reproducing asexually, are in many aspects different from sexual multi-cellular 'higher' organisms. The differences play out even in a very conceptual way: think e.g. about the distinction between lifetime versus generation time that is problematic in a bacterium, or the concept of the individual. But in many other aspects the evolutionary processes are similar and the understanding of mechanisms at work in bacteria must turn out fruitful for the study of higher organisms too. In the words of C. R. Woese: "[Bacterial evolution] holds the key to the evolution of the eucaryotic cell." [7]. And this is of course apart from the value of studying them for their own sake. An additional important benefit of studying bacterial evolution is their relative (!) biochemical simplicity, which facilitates our access to the molecular level.

1.2 The functional synthesis

Apart from microbiology, a second element that was conspicuously lacking in the study of evolution up until the mid-1980s was the explicit link between the genetic level and the level of phenotype and fitness. Evolutionary biology on one hand focused on the genetic level where polymorphisms were studied, phylogenetic relations constructed, and where statistical inferences were made about the contribution of positive selection and neutral drift to the observed genetic variation. On the other hand, phenotypic evolution and influences on fitness were mostly approached in a comparative and descriptive way. It was a major problem that often even for related species many genetic differences had accumulated over evolutionary time, of which it was hard to assess the importance for phenotype and adaptation. With the advent of novel molecular techniques and catalyzed by the return of microbiology to the evolutionary stage, attention was directed to the biochemical and structural basis of adaptation [19]. This integration between evolutionary biology and experimental molecular biology has recently been labeled 'the functional synthesis' [20], reminiscent of the earlier 'modern synthesis' of evolution and genetics.

Two elegant early examples are the so-called gene resurrection [21] studies [22, 23], in which phylogenetic analysis provided the sequence of extinct biological molecules, an ancient lysozyme and a ribonuclease, which were reconstructed and assayed in the laboratory. Remarkably, this approach had already been foreseen by Pauling and Zuck-

erkandl [24], but in a time that the appropriate molecular methods were not yet available. The functional synthesis opens new perspectives on old questions and allows new questions to be posed. For example, in what order have the molecular alterations taken place during adaptation? How many mutations are needed to produce new phenotypes, and does adaptation proceed by many mutations of small effect or few mutations of large effect [25–27]. And also: what is the biochemical or physical 'paleoenvironment' that ancient molecules experienced [28]?

Important to note is that the benefits of the functional synthesis go both ways: the historical perspective of evolutionary biology is also essential for the mechanistic understanding of bio-molecules in a reductionist field as molecular biology. The review by Golding and Dean [19] contains clear examples where phylogeny has pointed at specific residues that were key for molecular function, but had been overlooked by other approaches. Even more so, as often multiple solutions exist for performing the same molecular task, evolutionary analysis could reveal the reason why a certain solution is picked. Similarly, constraints on the evolutionary process and neutral drift (to both of which we will devote some attention later) could have co-shaped the present-day molecules.

Not to be omitted in the discussion of the functional synthesis is the emergent field of synthetic biology [29, 30]. A reductionist science *par excellence*, the prospects for the role of this discipline in evolutionary study should be favorable. The power of the synthetic approach is the necessity of a precise functional understanding of the systems under investigation: one is not allowed to overlook details, otherwise the system simply does not work. Synthetic biology provides a bottom-up perspective that makes it possible to single out specific biological components for evolutionary or mutational analysis [30]. Another niche for synthetic biology is the exploration of early life [31, 32]. What are the minimal biochemical requirements for replication, what biochemical characteristics of DNA and RNA make it suitable for carrying genetic information and replication [33]? Here again, the answers to evolutionary questions have consequences not only for evolutionary biology; they should also foster biotechnology and bio-engineering.

Most of the discussion sofar has addressed *adaptive* changes in organisms. A brief caveat is in order here. In the second half of the twentieth century the adaptionist program took a double blow. First, when confronted with the high amount of sequence difference between related species, it was proposed that many of the observed genetic changes should be neutral with respect to fitness (or also with respect to phenotype) [34,35]. Provocatively, the phenomenon was labeled "non-darwinian evolution" in one of the two initial papers [35]. The issue of how much change at the genetic level is neutral and how much is selected for is still actual today [27]. Addressing it satisfactorily will not be possible without explicitly taking into account protein structure and function. Secondly, a seminal paper by Gould and Lewontin [36] argued that not all changes need to be adaptive, even when they do have effect on the phenotype or the fitness.

Apart from pointing to random drift, they argued that traits of organisms could for example be imposed by their developmental program, or when a part of an organism acquires a new function, by the former function of that part (cooption). These limits to adaptation, or constraints (as they are excellently discussed in [37]), became a popular concept in evolutionary reasoning. However, a lack of solid quantitative basis also surrounded it with an aura of vagueness [38, 39]. With the rise of the functional synthesis more quantitative inferences can be made about the origin and extent of adaptive constraints. It is, in fact, essential to our understanding of natural selection to know whether it is limited by genetic or functional constraints [40], and what is the role of the specific selective pressures [41, 42].

1.3 The fitness landscape

A concept that is married very naturally to the functional synthesis is that of the fitness landscape, adaptive landscape, or adaptive topography. Introduced by Wright in 1932, it served as a metaphor to think about the effects of mutation, selection and drift. Wright's landscape (fig. 1.2) depicts organismal fitness as a function of its 'gene combinations' (whose high-dimensionality is reduced to two dimensions as a simplification). Well-adapted combinations are located on the peaks of the landscape, while valleys indicate poorly adapted combinations. When there are multiple peaks present, this implies that there are multiple solutions to the same evolutionary challenge. To give a well-known example, the gaits of ostrich, antelope, and kangaroo are different solutions to the same problem of animal locomotion [43].

Several kinds of fitness landscapes have been considered (but also confused) since Wright's introduction of the concept. When promoting fitness landscapes from a metaphor to a mathematical construction aimed at a quantitative understanding of adaptation, a dichotomy arose in what specified the horizontal axes of the landscape [45]. The axes in the first type of landscape define a (DNA or protein) sequence space, being the set of all genotypes. The other type of landscape specifies fitness as a function of genotype frequencies in a population. For example, in a population with two alleles A and a, it specifies the population averaged fitness as a function of the frequency of allele A. As the latter type of fitness landscape is derived from the former and moreover makes implicit assumptions about the evolutionary dynamics (as is clearly explained in [45]), we will only use the former type. Simply stated: it is a mapping from genotype to fitness. To complicate matters a bit, the term fitness landscape is also used for mappings from phenotype to fitness, where the axes specify the value of certain phenotypic parameters such as the catalytic rate of an enzyme. Since it is usually clear from the context which one is meant, the term fitness landscape will be used here interchangeably for genotype-fitness or phenotype-fitness landscapes. When the distinction is not clear, or requires attention, the specific mapping will be indicated.

The concept of a fitness landscape is useful, since its shape determines what kind

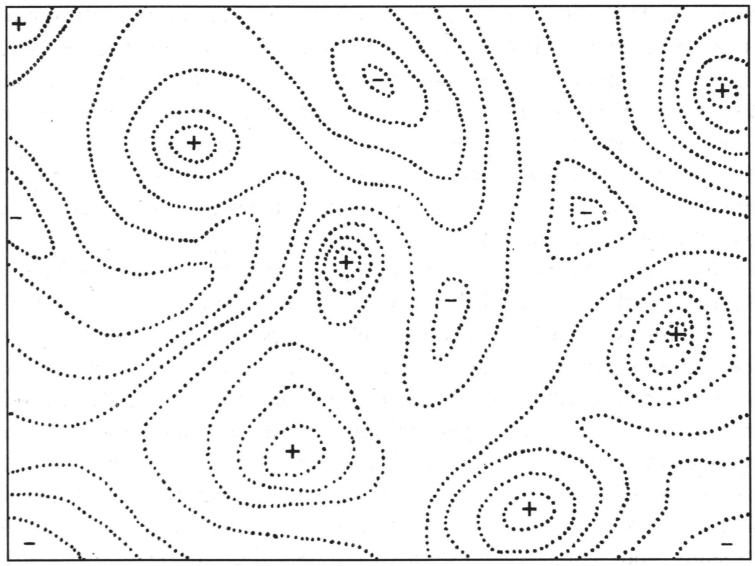


Figure 1.2: Sewall Wright's representation of an adaptive landscape, or fitness landscape from ref [44]. His original caption reads: Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.

of evolutionary dynamics can be expected. If a population of individuals is located on or around a peak, natural selection will tend to keep it there. This can be understood as follows: when an individual in a population acquires a mutation that lowers its fitness (makes a step away from the peak), it will most probably be removed from the population by selection, so that the population remains at the peak. When a population is located on a slope of a peak, selection will push it upwards: an individual acquiring a beneficial mutation (makes a step towards the peak) is fitter than the rest of the population, outcompetes the less fit individuals, and may increase its frequency in the population until it reaches fixation. In that case the new population approached the adaptive peak. If the landscape is smooth, adaptation can proceed by consecutive steps until a high fitness is reached. If the landscape is rugged, adaptation will probably not reach a high peak: the population gets stuck in a local optimum, as downwards movements are prohibited by selection. Implicit in the above description is the idea of an evolutionary trajectory (through DNA sequence space, or protein space) consisting of consecutive single mutational changes. It is based on the assumption that mutations occur only one at a time, which is usually realistic in view of naturally occurring mutation rates as was realized by Maynard Smith [46]. Adaptation can thus be visualized as a continuous path between an initial sequence and an end sequence. Multiple different paths connect the same points. Some paths will be more likely than others to be fol-

lowed by an adapting population, depending on the topology of the fitness landscape. Also heterogenous populations can be considered, which are usually referred to as a quasispecies [47], and can be represented as a cloud in the landscape.

Many theoretical studies have explored the effects of fitness landscape topology on adaptation. In spite of its apparent similarities with an (upside-down) energy landscape, often used in physics and chemistry, the analogy only holds when certain population genetics conditions are fulfilled [48]. Important for the theoretical elaboration of the concept were the contributions by Gillespie [49, 50], where dynamics governing taking steps and the step sizes are studied, those of Levin and Kauffman [51, 52], where the influence of ruggedness is explored in models with varying levels of epistasis (see below). More recent work by Van Nimwegen and Crutchfield [53] has focused on the escape from local (suboptimal) peaks, and the evolutionary dynamics when a neutral region separates two well-adapted states.

Important in the light of fitness landscapes is the concept of epistasis, which refers to the situation where the effect of a mutation depends on the genetic background in which it occurs. For example, the effect of a mutation from a to A can be different depending on the state of another locus, being b or B. The basis of this phenomenon often lies in the physical interactions of the gene products of A and B. The fitness (or phenotype) effect when mutating from a to A in different backgrounds can differ not only in magnitude, but also in sign: ab to Ab can be advantageous, while aB to AB can be deleterious. The latter form of epistasis is referred to as sign epistasis [54]. A more severe form, referred to as reciprocal sign epistasis [55] occurs if there is also sign epistasis for mutations b/B with respect to the background a/A. If (reciprocal) sign epistasis is present in a fitness landscape it tends to increase the ruggedness of the surface. In this way we can see the molecular or biochemical structure of the evolving system being reflected in the topology of the landscape.

An important issue is, obviously, what natural fitness landscapes look like. Recently studies have began to appear that construct fitness landscapes based on experimental data [55]. Attention has been focused on the accessibility of trajectories between evolutionary important starting and ending points: between naturally occurring polymorphisms [56], between ancient and present-day sequences [57, 58], or between sequences that are key in likely evolutionary scenarios [59]. So far these studies have shown that in general a solid fraction of pathways are inaccessible, but that accessible paths are present. The generality of these initial findings needs to be verified in future studies. The inaccessibility of pathways are directly related to adaptive constraints [60]. If no paths are accessible there is a strong genetic constraint in the system under consideration [61]. In that case a biochemical or structural cause can be further inspected. If accessible paths are present, but no adaptation is observed (for example in a laboratory evolution experiment), this might point to a lack of genetic variation. Currently landscape predictions for the distributions of beneficial mutations [62, 63] and for the occurrence of epistasis are being verified [64]. An intriguing and medically rel-

evant example is the finding is that the epochal appearance of new influenza viruses that transiently escape our immune system, can be explained by the viruses' wanderings through neutral regions in the landscape until they can access a beneficial mutation [65, 66].

1.4 Selection in variable environments

Most environments in which organisms live and develop are variable with respect to temperature, amount of light, chemical composition, or the presence of other organisms. This environmental heterogeneity may be spatial, temporal, or a combination of the two. When one shifts focus from selection in a constant environment to selection in variable environments, one immediately faces complications as many new parameters emerge. For example, if the environment is spatially heterogenous, what are the length scales of the variations, what is their distribution, and how much migration occurs between habitats? If the environment is temporally fluctuating, what is the nature of these fluctuations, stochastic or periodic, what is their time scale, and what are the ratios of time spent in each environmental state? For both types of heterogeneity the extent to which the environmental states are contrasting plays an important role.

Due to this multitude of possibilities, no encompassing theoretical framework for the evolution in variable environments has been developed. Nevertheless, the basis for describing dynamics in fluctuating environments is (often, but not always) directional selection in each environmental state, with a changing direction of selection among states. Such a fluctuating selective pressure can be a factor in maintaining genetic variation as was described by Levene [67] for spatial heterogeneity with high migration rates, and for temporal fluctuations by Haldane and Jayakar [68]. At the level of the individual organism, it is important to consider the trade-offs it experiences when environmental states impose contrasting demands. We have seen a clear example above: the heat-adapted organisms from Dallinger's experiment were not able to grow anymore at the much lower temperature their ancestors did and *vice versa*. This is commonly referred to as 'the cost of adaptation'. In general, in a changing environment such cost can arise from two factors. First, while adapting to one environmental state, mutations could accumulate that are detrimental to the fitness in the other state, but are not currently selected against. Second, which is more likely in Dallinger's case, the cost originates from what is referred to as 'antagonistic pleiotropy': a mutation increases fitness in one state, but decreases it in another. Both mutation accumulation, which is a property of the selective process, and pleiotropy, which is an inherent property of the genetic architecture of an organism, can constitute a constraint for adaptation to a variable environment.

The influence of trade-off structure on adaptation was explicitly considered from a theoretical viewpoint by Levins when he introduced his concept of a fitness set [69, 70]. He plotted the fitness in one environment versus the fitness in another for different

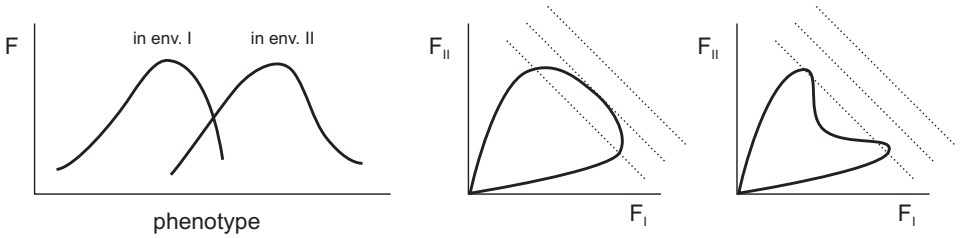


Figure 1.3: Selection in variable environments as represented by Levins [69]. The graph on the left depicts the effect of a certain phenotypic parameter on the fitness (F) in two environmental states I and II. When fitnesses in each state are plotted against each other, a so-called fitness set is obtained (middle and right graphs): the set of fitnesses resulting from different values of the phenotypic parameter. The more contrasting the demands are that the environmental states impose on the organism, the further the two peaks in the left graph shift apart. The amount of overlap determines the shape of the fitness set: more overlap results in a convex set (middle) and strongly contrasting demands result in a concave set (right). When the organism resides equally long in each environmental state, equal average fitness is indicated by the dotted isoclines. It can be seen that under a convex fitness an optimal strategy is to perform moderately well in both environments (middle), and adaptation could be expected to lead to generalists. Strongly contrasting states will favor specialization (right).

values of a phenotypic parameter that determines these fitnesses (fig. 1.3). From the resulting fitness set, one can determine which value(s) of the phenotypic parameter maximizes the overall fitness in the fluctuating environment. Interestingly, depending on the shape of the fitness set, this can correspond to a *generalist* phenotype that performs moderately well in both environmental states, or to *specialist* phenotypes that have a high fitness in one state only. When evolution would optimize this phenotypic parameter, the outcome therefore will critically depend on the strength of the trade-off, which in turn is determined by the contrast between the environmental states.

One way to escape the trade-off is by varying the phenotype in response to the environmental state, a phenomenon that in several biological disciplines is known as phenotypic plasticity [71, 72]. As adaptations to environmental variation, organisms possess regulatory systems that sense and process environmental signals to accomplish altered gene expression, metabolic change, mechanic responses, or even genetic changes. The remainder of this introduction will be primarily focused on an altered gene expression in response to the environment, carried out by gene regulatory networks. As gene regulation is subject to selective forces, the expression of a gene as a function of the relevant environmental parameter should in some way reflect the nature of the environmental fluctuations. Recent experimental work addressed the adaptive optimization of expression levels in terms of cost (spurious expression when not required) and benefit (high expression when needed) [73]. Theoretical [74–78] and some experimental [42] studies have appeared that analyze the environmental condi-

tions under which it is beneficial to employ a regulatory system.

Another line of research into the effects of variable environments are experiments into diversification in microbial ecosystems [79]. Here, controlled spatial and/or temporal variation is imposed and adaptive responses are observed. Fitness trade-offs are inferred and attempts are made to understand the underlying causes for costs of adaptation (e.g. [80, 81]). Although these studies have been successful in demonstrating pleiotropic trade-offs and the effect of several population dynamical phenomena, there is usually a limited access to the molecular level.

To conclude this section, we note that gene regulatory change is commonly believed to be the main factor underlying phenotypic and morphological evolution [82–87]. It is found that regulatory elements often undergo much faster evolutionary change than sequences coding for structural, metabolic, or other proteins [88]. Therefore, in order to study evolution at all levels ranging from immediate phenotypic responses, through developmental control, to speciation, it is necessary to understand selection under variable environments. As will be clear by now, this is a big challenge, but an important one.

1.5 The lactose operon

Most work presented in this thesis is an investigation into (evolutionary) aspects of the *Escherichia coli* lactose operon. *E. coli*, an intestinal inhabitant of warm-blooded animals, has been a model system for studying bacterial physiology and genetics, and is often the organism of choice in laboratory evolution studies. The lactose operon can be said to be the model organism's model regulatory system. Expression of the *lac* operon gene products allows *E. coli* to efficiently metabolize lactose, in the absence of other substrates that would sustain higher growth rates. When a mix of sugars is available, expression of this and other sugar metabolizing operons is regulated in such a way that the bacterium sequentially utilizes the substrates that yield the highest instantaneous growth rate. These and similar observations in other organisms, led Jacob and Monod in the late 1950s to postulate the model of "negative control", where expression of metabolic genes is shut off by an inducible repressor [89]. Immediately after this discovery theoretical models (based on reaction kinetics) began to appear (e.g. [90, 91]), aimed at a quantitative description of the system. Although the *lac* operon became the paradigm for gene regulation, subsequent experimental discoveries and theoretical refinements have continued to date (see also the introduction of chapter 7).

As for the structural parts of this operon, the current knowledge is summarized in figure 1.4. The lactose operon consists of four genes: *lacI*, the repressor that can bind in three locations to operon DNA where it suppresses the expression of the downstream *lacZ*, *lacY*, and *lacA* genes. *lacZ* encodes the catabolic enzyme β-galactosidase that hydrolyses lactose into glucose and galactose. The gene product of *lacY* is a permease for α and β galactosides. Lastly, *lacA* codes for a thiogalactoside transacetylase, whose

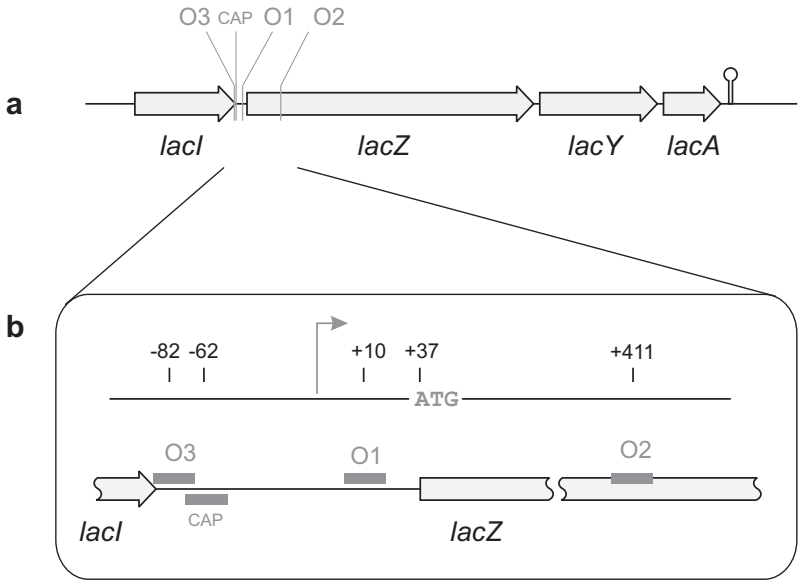


Figure 1.4: Schematic overview of the *lac* operon of *E. coli*. **a**) The location of genes and sites in the operon DNA an their sizes are drawn to scale. *lacI* codes for the *lac* repressor that can bind specifically to three locations, operators O1, O2, and O3, thereby competitively inhibiting RNA polymerase from transcribing the downstream genes *lacZ*, *lacY*, and *lacA*. CAP represents the binding site of the transcriptional activator protein CAP or CRP. A transcriptional terminator is located downstream of *lacA*, determining the boundary of the operon. **b**) The *lac* regulatory architecture in more detail. Distances from the transcriptional start (grey arrow) are given in base pairs. The translational start of *lacZ* is denoted as `ATG`. Note the overlap between operator O3 and the CAP site. Exact locations as given in EcoCyc [92]

function in the *lac* operon is, remarkably, still unknown [93].

The control region of the *lac* operon has been zoomed into in figure 1.4b. Distances in base pairs from the transcriptional start are shown, as well as the binding sites for LacI: the operators O1, O2, and O3. The *lac* operators are (partly) palindromic DNA sequences with a length of 22 base pairs. The symmetry of the operators reflects the symmetry of the *lac* repressor where two monomer subunits bind to one operator. Since the *lac* repressor acts as a tetramer, it can bind two operators at the same time by looping the DNA. Also shown is the binding site for the CAP or CRP activator protein, which partly overlaps with O3. As this region is described and modeled in much more detail in chapter 7 and the references therein, here the description is limited to its essentials. In short, the *lac* control region integrates two signals[1], that enable it to express the metabolic enzymes only when lactose is present and a more optimal substrate, glucose, is not. Lactose, in its modified form of allolactose, binds to the repressor to release it

[1]The role of a third signal, being the nucleoid structuring protein H-NS, is less relevant for the current description.

from the operators (induction). The absence of glucose causes a rise in the cytoplasmic concentration of cAMP, which is a cofactor of CAP and hence results in activation of the *lac* promoter. Recent mechanistic descriptions of the *lac* operon behavior include explicit descriptions of DNA bending due to binding of the repressor and the CAP protein, and include the CAP-LacI competitive binding. They take into account the association of multiple inducer molecules to the tetrameric repressor, the stochastics due to the low number of repressors per cell, and focus on nonlinear characteristics of the system, such as feedback and bistability.

The *lac* system has been elucidated to the present high level of detail as a result of a tremendous research effort in the course of nearly 50 years. Our knowledge of its evolutionary history, however, is still in its infancy, let alone our knowledge of other (including eukaryotic) gene regulatory systems. Many aspects are wide open for exploration. It is almost an empty statement that *lac* regulation somehow evolved in response to an environment where the presence of lactose is fluctuating. As we have seen earlier, evolution in variable environments has many more dimensions (also literally) than evolution in a constant environment. We may ask ourselves for which fluctuation time scales the *lac* system is optimal. And under what conditions the regulatory system is retained, lost, or modified? Moreover, since transcriptional activation is available as an alternative mode of regulation, what conditions favor repression in the case of lactose metabolism [75, 76]? The answers to these questions will depend heavily on the dynamics of *natural* populations of *E. coli*, which in fact is another subject in its infancy.

Notably, the previous questions mainly address the optimality of a regulatory system already in place, and are markedly different from questions about its origin. We may wonder what are the early stages of the development of a regulatory system. Phylogenetic inquiries, also in the case of the *lac* operon, lift a part of the veil: the *lac* repressor is a member of an extended family of transcriptional regulators in *E. coli* and related bacteria [94]. The existence of such families (based on sequence or structural homology) can be explained by either horizontal gene transfer [95] or by duplication and divergence of the constituting genes [59, 96]. Most probably both processes are responsible in part. In any case, having a number of homologous regulators in one organism poses the interesting question how two dedicated parts, the transcription factor and its operator, which function like a key and a lock, can change while retaining their tight binding affinity, when one realizes that evolution usually proceeds by one mutation at a time. Another way of phrasing this question is: does the system exhibit rampant (reciprocal) sign epistasis, or are there evolutionary trajectories where this is avoided? In the case of duplication, cross-talk and possible heterodimerization of the repressor multimers could play important roles. These issues cannot be resolved by looking at the phylogeny alone: a detailed understanding of the system's function and its influence on organismal fitness are essential.

1.6 This thesis

This thesis will present a number of explorations into the evolution of regulation in a temporally varying environment. We have employed a number of approaches and concepts described above.

First, in chapter 2 we review four studies that analyzed evolutionary trajectories in fitness landscapes representing a variety of molecular components in different organisms. These studies have focused on the accessibility of trajectories from one functionality to another, taking as begin and end point ancestral and present day forms, specific clinical isolates, or inferred points from homology analysis. In all cases viable pathways exist, although there were also indications for non-accessible trajectories. Since these studies have access to the molecular, phenotypical, and/or fitness level, all cases revealed interesting features of the selective process. For example a potential adaptive constraint due to key-lock issues governing multi-component evolution could be avoided by either molecular cooption or interactions at the network level.

Chapter 3, which is one of the studies discussed in chapter 2, uses the abundant evidence for gene duplication and divergence being a major creative evolutionary force, as a starting point to investigate regulatory divergence at the level of the operator and repressor's recognition domain. Here the link between the genotype and phenotype is formed by a large collection of measured mutant binding affinities, originating from the group of B. Müller-Hill. Assuming a relevant selective pressure, we investigated the nature of the divergence pathways and how their accessibility is affected by alternative selective conditions. We found that divergence can proceed relatively unconstrained, in spite of a rugged fitness landscape. It appears that also the high dimensionality of this landscape (owing to its large mutational dataset) favorably affects the amount of accessible trajectories.

Subsequently, chapter 4 presents an experimental investigation into the trade-off structure that leads to an effective selective pressure for regulation. Again, the study employs a fitness landscape (here a phenotype-fitness map) that could be fully experimentally determined. By varying the selective stringency, analogous to the situation in fig. 1.3, we could quantitatively control the selective advantage of regulatory phenotypes over non-regulatory phenotypes. We followed the adaptation of randomly mutated *lac* repressors and a regulatory cascade to the imposed fluctuating selective pressure. In this work we combined an evolutionary and synthetic approach to be able to directly observe the molecular basis of regulatory adaptation and its constraints.

Chapter 5 elaborates on the results of the previous chapter. Taking a *lac* repressor with a novel, inverse functionality as starting point, we tried to assess which genetic substitutions with respect to the wild-type *lac* repressor are functional, which are neutral and what is the extent of epistasis between the mutations. The approach followed here is useful in reducing the combinatorial complexity of that arises when one wants to assess the selective importance of genetic substitutions between an ancestral and an evolved sequence. Also we should be able to obtain information on the ruggedness of

an actual repressor fitness landscape.

In chapter 6 we use a laboratory evolution approach to investigate how the natural *lac* regulatory system adapts to a selective regime for which it is not optimal. We were able to map and vary the extent of non-optimality by using both an artificial inducer and an artificial carbon source, thereby separating two features normally incorporated in lactose. In this case genetic variation originates from spontaneous mutation, in contrast with work in the chapter 4. By performing adaptation experiments, we explored the conditions under which *lac* regulation is conserved, lost, or modified, and compare the results to our prior optimality analysis. Information about selective pressures in constant environments is used to interpret the evolutionary outcomes of experiments in alternating environments.

In order to make inferences about the selective pressures governing regulatory adaptation, one has to have a sufficient level of quantitative understanding of the regulatory system, and *vice versa*. For example, the noise properties of molecular systems (due to a small number of constituent molecules), could be a factor that has been optimized by selection, as is proposed e.g. in refs. [97, 98]. In chapter 7 we argue that residual affinity of induced *lac* repressor, which is omitted in recent thermodynamic descriptions of the *lac* operon, is directly relevant for the optimal functioning of *lac* regulation. We worked out a basic thermodynamic model that includes residual affinity, and highlight predictions that are different from models without residual affinity. The results point to an optimal relation between the *lac* repressor copy number and its residual affinity for the operator.

Chapter 8 contains a theoretical exploration of the consequences of reciprocal sign epistasis for the shape of a fitness landscape. It was noted by Weinreich *et al.* [54] that sign epistasis is a necessary and sufficient condition for having inaccessible trajectories in the fitness landscape. Here we conclude that reciprocal sign epistasis is a necessary condition for multiple adaptive peaks in the landscape. However, sufficient conditions of any simple form do not exist.

Biological systems not only reflect their evolutionary history, but in the age of genetic manipulation and synthetic biology, another history is sometimes also important: that of human practice. Appendix B briefly discusses how a widely used expression marker, *lacZ*α became incompatible in combination with plasmids containing pBR322 origin of replication. It is an inheritance from the pre-sequence era. The cloning steps leading to this incompatibility are discussed, and an alternative *lacZ*α was constructed. Finally, along the way it was found that the complementary part of this marker *lacZ*ω, when present in the common genotypic marker ϕ80d*lacZ*Δ(M15) is unexpectedly accompanied of a highly expressed *lac* repressor. That this is never stated in specifications of the genotype is the result of a historic typographical accident (appendix A).

Summarizing, this thesis presents a number of experimental and theoretical explorations into (mainly) evolutionary aspects of gene regulation. As always, each of the followed approaches will have its strong points and its limitations. The idea is that

these and other approaches will contribute to a quantitative understanding of regulatory evolution. It is with the current state of knowledge, experimental tools, and the growing amount of genetic information [99], an exciting field to work in, asking for a multidisciplinary approach. What was true in the time of the synthesis between evolutionary biology and genetics is as true for present evolutionary research, as expressed by G.G. Simpson in 1944 [100]: "The basic problems of evolution are so broad that they cannot hopefully be attacked from the point of view of a single scientific discipline."

CHAPTER

2 Empirical fitness landscapes reveal accessible evolutionary paths

It is much easier for a mouse to get a set of genes which enable it to resist *Bacillus typhimurium* than a set which enable it to resist cats.

J. B. S. Haldane,
Ric. Sci. Suppl. A 19 68-76, 1949

Evolutionary intermediates represented a central preoccupation for Darwin in his case for the theory of evolution. He remarked, for example: '...why, if species have descended from other species by insensibly fine gradations, do we not everywhere see innumerable transitional forms?' [101]. Although Darwin developed a convincing rationale for their absence, he did realize that the lack of intermediates as proof leaves room for criticism. He noted, for instance: 'If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down.' [101]. Indeed, in their opposition to evolution, the proponents of 'intelligent design' have seized on our current ignorance of intermediates.

Building on earlier ideas [22–24, 46], an approach has recently been developed to explore the step-by-step evolution of molecular functions. The central innovation is that all molecular intermediates along multiple putative pathways are explicitly reconstructed. Together with a phenotypic characterization of each intermediate, one can determine whether paths towards a certain novel function are accessible by natural selection. Although others have reconstructed and characterized phylogenetically ancestral forms of proteins [21–23, 102], here the focus is on fitness landscapes [44] in which multiple mutational trajectories can be compared. Fitness landscapes have

been widely studied on a theoretical level (see refs [45, 50, 52, 53] for example), but one can now obtain a glimpse of actual biological landscapes. This view finally allows us to ask which particular evolutionary paths are taken and why. In particular, to what extent do biomolecular properties constrain evolution? Does it matter in which order mutations occur? Are fitness landscapes rugged, with many local optima acting as evolutionary dead-ends, or are they smooth? Is neutral genetic drift essential for a new trait to emerge?

When examining the molecular underpinnings of the evolution of new traits, we distinguish two elementary cases. First, we discuss a single mutable component such as an enzyme. Second, we look at molecular interactions involving two or more mutable components, which is typical for regulatory evolution. The specific features of this broad range of molecular systems will be discussed using the notions of epistasis and fitness landscapes, which we will explain and relate to each other (Box 2.1 and Fig. 2.1).

The tentative picture emerging from the new results is one that emphasizes the possibilities of continuous optimization by positive selection. Although evolution was clearly constrained, as illustrated by many inaccessible evolutionary paths, the studies also revealed alternative accessible routes: a succession of viable intermediates exhibiting incremental performance increases. Although these findings do not address whether natural evolution proceeds in the presence or absence of selection, they do show that neutral genetic drift is not essential in the cases studied. We note that the presented approach starts with naturally occurring sequences, which are themselves the product of evolution, and may therefore yield a biased sample of trajectories. Whether the conclusions are general or not, and whether they break down when the evolved feature becomes more complex, can only be determined through future studies.

2.1 Enzyme evolution

When a well-adapted organism is challenged by a new environment, an existing gene may perform suboptimally. One of the most basic questions one may then ask is: when mutating step-by-step from the suboptimal to an optimal allele, are all possible trajectories selectively accessible? This question depends critically on the stepwise changes in performance, or in fitness, which are governed by unknown physical and chemical properties at the molecular level. When all mutations along all paths yield a fitness improvement, evolution can rapidly proceed in a straightforward incremental darwinian fashion. In this case, the fitness landscape can be portrayed by a single smooth peak (Fig. 2.1a).

Whether this picture is realistic was investigated for the adaptation of bacterial β-lactamase to the novel antibiotic cefotaxime [56]. The central step was to reconstruct and measure all likely intermediates, allowing a systematic study of all possible trajectories. The intermediate sequences can be easily identified, because the (five) mutations that control the cefotaxime resistance phenotype are known, resulting in $2^5 = 32$

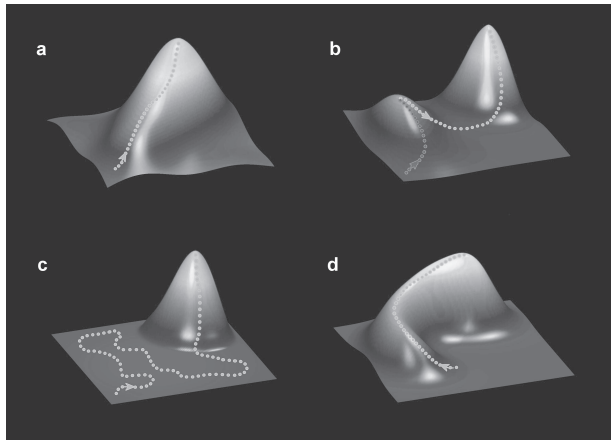


Figure 2.1: Schematic representations of fitness landscape features. Fitness is shown as a function of sequence: the dotted lines are mutational paths to higher fitness. **a**, Single smooth peak. All direct paths to the top are increasing in fitness. **b**, Rugged landscape with multiple peaks. The light path has a fitness decrease that drastically lowers its evolutionary probability. Along the darker path selection leads in the wrong direction to an evolutionary trap [59]. **c**, Neutral landscape. When neutral mutations are essential, evolutionary probabilities are low [53, 103]. **d**, Detour landscape. The occurrence of paths where mutations are reverted [59] shows that sequence analysis may fail to show mutations that are essential to the evolutionary history.

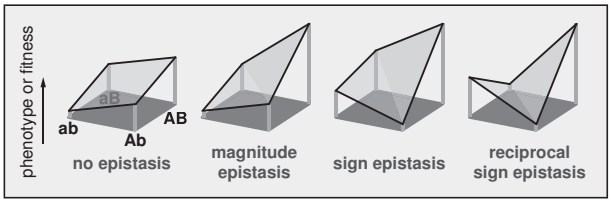
possible mutants. The order in which the mutations are fixed can of course be different, giving rise to 5! = 120 possible direct trajectories between the start and end sequences.

The trajectory analysis showed that the fitness landscape is not as simple as depicted in Fig. 2.1a. A majority of the pathways towards maximum cefotaxime resistance actually shows a dip in fitness (see light path in Fig. 2.1b), or contain selectively neutral steps (as in Fig. 2.1c), resulting in much smaller chances of being followed by natural selection [53, 103]. For 18 paths however, each step appeared to confer a resistance increase, making these trajectories accessible to darwinian selection. The part of the fitness landscape mapped out in this manner therefore does appear to have a single peak, but one that contains depressions and plateaus on its slopes. We stress that such three-dimensional analogies, while useful for conveying basic characteristics, do not rigorously represent the many direct trajectories existing between two alleles. Also note that there may be additional paths that contain detours, involving other mutations that are eventually reverted [59] (Fig. 2.1d).

Interestingly, some mutations yielded either a resistance increase or decrease, depending on the preceding mutations. This phenomenon, called sign epistasis [54] (see Box 2.1), is both a necessary and sufficient condition for the fitness landscape to contain inaccessible paths to an optimum [54]. Some cases of sign epistasis could be understood in terms of competing molecular mechanisms. For instance, a first muta-

tion in the wild-type enzyme increased the resistance by enhancing the catalytic rate, even though it also lowered the thermodynamic stability. This loss of stability was repaired by a second mutation, thereby further increasing the resistance. In contrast, when this 'stabilizing' mutation occurred first in the wild-type enzyme, the resistance was reduced. Such back and forth balancing between structural and functional benefits might well be a more general evolutionary mechanism [104, 105].

Box 2.1. Epistasis and the accessibility of mutational paths. Epistasis means that the phenotypic consequences of a mutation depend on the genetic background (genetic sequence) in which it occurs. In the figure below we distinguish four cases that illustrate paths composed of two mutations, from the initial sequence 'ab' towards the optimum at 'AB'. When there is no epistasis, mutation 'a' to 'A' yields the same fitness effect for different genetic backgrounds ('b' or 'B'), while for magnitude epistasis the fitness effect differs in magnitude, but not in sign. For sign epistasis, the sign of the fitness effect changes. Finally, such a change in sign of the fitness effect can occur for both mutations, which we here term reciprocal sign epistasis. These distinctions are crucial in the context of selection. Mutations exhibiting magnitude epistasis or no epistasis are always favored (or disfavored), regardless of the genetic background in which they appear. In contrast, mutations exhibiting sign epistasis may be rejected by natural selection, even if they are eventually required to increase fitness. In other words, some paths to the optimum contain fitness decreases, while other paths are monotonically increasing. When all paths between two sequences contain fitness decreases, there are two or more distinct peaks. The presence of multiple peaks indicates reciprocal sign epistasis, and may cause severe frustration of evolution (Fig. 2.1b). Indeed, reciprocal sign epistasis is a necessary condition for multiple peaks, although it does not guarantee it: the two optima in the diagram may be connected by a fitness-increasing path involving mutations in a third site.



In a second study [58], the connection between fitness landscape and underlying molecular properties has been explored for the evolution of isopropylmalate dehydrogenase (IMDH, Fig. 2.2a), an enzyme that is involved in the biosynthesis of leucine. As in the previous study, a set of mutational intermediates between different functions were characterized. Here the mutations changed the cofactor binding affinity of IMDH. *In vitro* measurements of enzyme activity did not show epistasis: each mutation gave

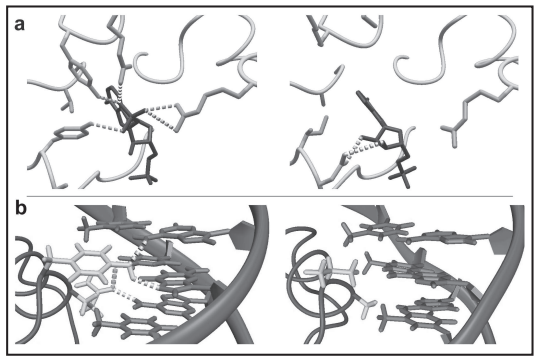


Figure 2.2: Molecular structures in different evolutionary forms. **a**, The left panel shows wild-type *E. coli* isocitrate dehydrogenase [107] (IDH), which is structurally similar to IMDH, with NADP as cofactor. The right panel shows an engineered IDH form with NAD as cofactor [108]. Main chains are shown in grey, cofactor in black, and hydrogen bonds as dashed lines. **b**, The left panel shows a wild-type *E. coli lac* repressor and operator [109]. The right panel shows a *lac* repressor and operator variant, with mutations mimicking the gal system [110]. Binding is tight and specific (despite the absence of hydrogen bonds): these variants bind wild-type partners poorly. DNA backbone and key bases are shown in dark grey, repressor chains in black, key repressor residues in grey, and and hydrogen bonds as dashed lines. Figures prepared with MOLMOL [111].

a fixed catalytic improvement, which was independent of the order in which they occurred. Thus, the 'enzyme activity' landscape is single-peaked.

The story becomes more complete with the following elements. First, the study also considered evolutionary paths from the suboptimal cofactor NADP to the normal cofactor NAD [106]. Second, selection does not act directly on enzyme activity, but rather on the fitness of an organism. As fitness is typically nonlinear in enzyme activity, epistasis is introduced. Therefore, the IMDH mutants were also evaluated *in vivo*, providing a direct measurement of the fitness effect of a mutation. The resulting fitness landscape was shown to contain a depression or valley, rendering the trajectories that pass through it selectively inaccessible. There is an intuitive rationale for a valley here: when the recognition of NADP is reduced, the fitness first decreases, before it rises again when NAD recognition is built up. Interestingly however, some trajectories also exist that avoid the valley by simultaneously increasing NAD, and decreasing NADP recognition. In the end, the genotype-fitness landscape has a single peak, but one that includes a depression on its slope.

2.2 Evolution of molecular interactions

The evolutionary puzzle becomes more complex at a higher level of cellular organization. In the web of regulatory interactions between ligands, proteins and DNA, the

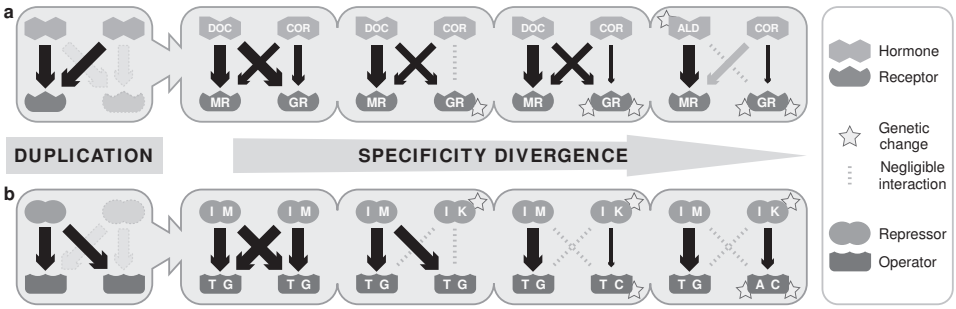


Figure 2.3: Evolution of molecular interactions based on reconstructed intermediates. Arrow thickness denotes measured interaction strengths. DOC, 11-deoxycorticosterone; COR, cortisol; MR, mineralocorticoid receptor; GR, glucocorticoid receptor; ALD, aldosterone. **a**, Pathway towards independent steroid receptors after duplication, via intermediate receptors that remained sensitive to their ligands [57]. A changed mutation order produced a non-sensitive intermediate, making that path inaccessible. The grey arrow indicates that cortisol is absent in MR-expressing tissues. **b**, Pathway towards independent repressor-operator pairs following duplication, taking single-mutation steps without decreases in network performance. Many paths were compared in a landscape based on over 1,000 *lac* mutants [112], covering all substitutions on all key base pairs. For simplicity, the repressor dimer and two operator half-sites are not drawn.

components are strongly interdependent, which might suggest that their evolution is severely constrained. The evolution of molecular recognition has recently been explored by two studies, which also used experimentally reconstructed intermediates. The first examined hormone detection by steroid receptors in the basal vertebrates (Fig. 2.3a) [57]. The second [59] looked at the adaptation of repressor-operator binding, in a large evolutionary landscape based on published mutation data for the *Escherichia coli lac* system [112] (Figs 2.2b and 2.3b). For both studies, the molecular interactions may be thought of as a key fitting a lock. The unifying question is: can a new lock and matching key be formed taking just one mutational step at a time? The adaptation of these components presents a dilemma: if the lock is modified first, the intermediate is not viable because the old key does not fit, and vice versa.

From the evolution of the interactions in the two systems (Fig. 2.3), some interesting parallels are apparent. Both studies start with a duplication event yielding two locks and keys, and then ask how specific interactions can be obtained during mutational divergence. Specificity is clearly vital: two partners must recognize each other, but not recognizing other components is just as important. A major evolutionary challenge is therefore to decrease unwanted interactions, while maintaining desired interactions. Without specific hormone recognition, cortisol regulation of vertebrate metabolism, inflammation and immunity would be perturbed by varying levels of aldosterone, which controls electrolyte homeostasis. Similarly, specific recognition in the *lac* family of repressors allows *E. coli* to consume a wide array of sugars, without the burden of pro-

ducing many unused metabolic enzymes.

Surprisingly, these studies again show that new interactions can evolve in a step-by-step darwinian fashion, despite the mismatching intermediates problem sketched above. In the hormone receptor case, this predicament is overcome by a molecular version of a master key: a putative ancestral ligand, 11-deoxycorticosterone, was found to activate all receptors (ancestral and present-day), allowing the mutational intermediates to remain functional even while the receptors diverged (Fig. 2.3a). The capability to synthesize aldosterone evolved later, finally providing a specific hormone that is recognized by just one of the two receptors. An existing receptor was thus recruited into a new role, as a binding partner to aldosterone, in a process that was termed 'molecular exploitation'. Sign epistasis was again observed: an initial mutation drastically lowered the response to all substrates, but after another mutation, the same mutation improved cortisol response while decreasing the aldosterone response. Thus, just as in the β-lactamase and IMDH cases, at least one selectively accessible evolutionary pathway existed.

In the evolution of the *lac* system, a similar mechanism using a 'master' repressor or operator was not observed. This is illustrated by the transient loss in affinity during the adaptation from one tight repressor-operator pair (IM-TG) to another (IK-AC); see Fig. 2.3b. Between some alleles, all connecting paths transiently reduced the affinity, indicating the presence of multiple peaks in the affinity landscape, which contrasts with the single-peaked landscapes of β-lactamase and IMDH. Multiple peaks indicate a severe kind of sign epistasis, which we here term reciprocal sign epistasis (see Box 2.1). Reciprocal sign epistasis can be intuitively understood for molecular interactions: mutating one binding partner will probably only benefit a new interaction if the other binding partner is mutated first, and vice versa. Interestingly, this means that although sign epistasis does introduce landscape ruggedness and thus perturbs the adaptive search, it can also be valuable because it enables multiple independent lock-key combinations.

If the *lac* repressor-operator affinity landscape is rugged and multi-peaked, how can new recognition evolve in a step-by-step manner? The answer lies in the fact that selection does not act on a single interaction. Instead, multiple interactions in a network determine the regulation, and ultimately organismal fitness. In the *lac* case, deteriorations in one interaction were offset by improvements in another. For example, initial mutations in one repressor duplicate were bad for binding to its designated operator, but good for relieving an undesired cross-interaction (Fig. 2.3b). These results substantiate the suggestion that network robustness [113] may promote evolvability [114, 115]. The observed compensations yielded a smoothened fitness landscape, making the new interactions selectively accessible. In fact, because compensation within biochemical networks is ubiquitously observed [116], we expect that evolution by network compensation constitutes a general mode of regulatory adaptation, molecular interdependence notwithstanding.

2.3 Outlook

The experimental reconstruction of evolutionary intermediates and putative pathways has provided an exciting first look at molecular adaptive landscapes. Although numerous paths appear to be selectively inaccessible, accessible pathways are generally also available. Importantly, various alternative types of fitness landscapes were not observed. The landscapes could have been so rugged and multi-peaked, that accessible paths to optima would not exist, thus requiring, for instance, two or more simultaneous mutations, larger genetic modifications through recombination, or periods of relaxed selection. We have put forward various mechanisms that can reduce landscape ruggedness and improve evolvability. These include the interplay between protein function and stability [56, 58], the exploitation of existing molecules into new roles [57], and compensation within biochemical networks [59].

That only a few paths are favored also implies that evolution might be more reproducible than is commonly perceived, or even be predictable. It is important to note that evolutionary speed and predictability are not determined only by molecular constraints, but also by population dynamics. Population dynamics still presents many open questions, in particular in the context of regulatory evolution and varying environments. The situation in which environmental fluctuations are fast relative to selection timescales has been explored in the repressor divergence study [59]. Recent theoretical considerations [77, 117] may provide promising approaches to address these questions more generally.

The molecular systems interrogated so far represent only a start, but one with great potential to spark further exploration. The analysis of intermediates is generally applicable, which makes finding new candidate systems not difficult. Mutational paths could also be revealed using the directed evolution methodology [118], in which randomly mutated pools are screened. A related approach is the experimental evolution [119] of cells in chemostats [120] or by serial dilution [73, 121]. The advantage of these methods is that more extensive and unbiased evolutionary changes can be explored, although they do not directly reveal why trajectories are chosen. Together, these developments may change the character of molecular evolution research from one that is primarily sequence-based to one that explicitly incorporates structure, function and fitness.

CHAPTER

3 Evolutionary potential of a duplicated repressor-operator pair

Everything not forbidden is compulsory.

T.H. White,
The Once and Future King

Ample evidence has accumulated for the evolutionary importance of duplication events. However, little is known about the ensuing step-by-step divergence process and the selective conditions that allow it to progress. Here we present a computational study on the divergence of two repressors after duplication. A central feature of our approach is that intermediate phenotypes can be quantified through the use of in vivo measured repression strengths of Escherichia coli lac mutants. Evolutionary pathways are constructed by multiple rounds of single base pair substitutions and selection for tight and independent binding. Our analysis indicates that when a duplicated repressor co-diverges together with its binding site, the fitness landscape allows funneling to a new regulatory interaction with early increases in fitness. We find that neutral mutations do not play an essential role, which is important for substantial divergence probabilities. By varying the selective pressure we can pinpoint the necessary ingredients for the observed divergence. Our findings underscore the importance of coevolutionary mechanisms in regulatory networks, and should be relevant for the evolution of protein-DNA as well as protein-protein interactions.

Initially put forward by Stevens in 1951 [122] and later advocated by Ohno in his seminal work [123], gene duplication followed by functional divergence is now seen as a general mechanism for acquiring new functions [124]. Also, regulatory networks are

thought to be shaped significantly by genetic duplication [96]. For instance, sequence analysis of transcription factor families points to various historical duplication events [125, 126]. However, very little is known about the subsequent mutational divergence pathways or about the corresponding stepwise phenotypical changes that are subject to selection. While these issues have not yet been explored experimentally, related generic aspects of mutational plasticity have been addressed theoretically [52, 127–130]. However, a central obstacle in studying mutational pathways through computer simulations remains the unknown relation between the sequence and binding affinity, for which, in general, a rather abstract mapping has to be assumed. To describe the formation of a new regulatory interaction after a duplication event, which is our current aim, such an abstract approach would be particularly speculative.

Here we reason that many characteristics of the adaptation of real protein-DNA contacts are hidden in the extensive body of mutational data that has been accumulated over many years (e.g., [93, 112, 131] for the *Escherichia coli lac* system). These measured repression values can be used as fitness landscapes, in which pathways can be explored by computing consecutive rounds of single base pair substitutions and selection. Here we develop this approach to study the divergence of duplicate repressors and their binding sites. More specifically, we focus on the creation of a new and unique protein-DNA recognition, starting from two identical repressors and two identical operators. We consider selective conditions that favor the evolution toward independent regulation. Interestingly, such regulatory divergence is inherently a coevolutionary process, where repressors and operators must be optimized in a coordinated fashion.

The mere presence of a selective pressure is clearly not a sufficient condition to achieve a new function. Rather, the evolutionary potential and limitations can be seen as governed by the shape of the actual fitness landscape and the evolutionary search within it. Studying these intrinsic limitations to divergence represents the overall aim of this work. Many open questions arise when considering the formation of a new protein-DNA interaction, which may be viewed as the construction of a new lock and uniquely matching key. For instance, how should the protein be modified step-by-step to recognize a new DNA-binding site that also does not yet exist, or vice versa? One would expect that complementary mutations need to occur in the protein and DNA-binding site. Does this mean that temporary losses in fitness must be endured when taking single-mutation steps? And, how many mutations must minimally accumulate before a noticeable new recognition is obtained on which selection can act? The latter is an important point: mutations conferring a selective advantage spread more readily through a population [103], resulting in a drastic increase of the divergence probability. These questions are addressed by exhaustively searching the landscape for optimal pathways, as well as by complementary population dynamics simulations.

Previously it has been shown that *lac* repressor mutants indeed exist that can bind exclusively to mutant *lac* operators [112]. Our simulations reveal that a duplicated repressor-operator pair can readily evolve to achieve such independence of binding,

while monotonously increasing its fitness in a step-by-step process. Moreover, simply following the fittest mutants does predominantly guide the system to the desired global optimum, which indicates funnel-like features in the fitness landscape. A detailed analysis of the subsequent network changes indicates a generic sequence of events, of which we study the underlying mechanisms by varying the applied selective pressure. Next, we show that the trajectories we find in the optimal pathway simulations are not rare exceptions, since similar trajectories are followed using a probabilistic scheme for accepting a mutation. The results further suggest the feasibility of studying regulatory divergence in laboratory evolution experiments, and finally we make a comparison to alternative models for the creation of new regulatory interactions.

3.1 The model

3.1.1 Selective pressure and the fitness landscape

We consider an ecological situation where natural selection would favor independent regulation of two genes X and Y. Regulation is not independent in the initial symmetric network with duplicated components (see Figure 3.1): X and Y have two identical upstream binding sites (O_1 and O_2), which bind two identical repressors (R_1 and R_2) equally strongly. Such a situation will, for instance, arise upon duplication of a repressor that regulates two or more genes. Note that this selective pressure, of course, is not a general outcome of a repressor duplication. A duplication event may arise in the context of a different functional pressure, which could direct the evolution toward a different topological motif [132]. Most often, selective pressures for a new function will be absent, in which case silencing of one of the duplicates is the most probable outcome [124, 133]. However, the rare cases where a selective pressure is present are crucial to developing new functions.

We aimed to define a transparent selection pressure for the divergence of these regulatory interactions. Attributing a fitness value to a network function is non-trivial: unlike for an enzymatic function, network fitness cannot be captured in a single biochemical parameter. Here we propose to assign a fitness value based on the desired input-output relation of the network (see Figure 3.1A and 3.1C). For simplicity, only two concentration levels (high and low) of input and output protein are considered, resulting in four possible input conditions. For each of these input conditions, it follows straightforwardly which repressor-operator interactions should be maximized and which must be minimized. The interaction strength between operator O_i and repressor homo-dimer R_j is expressed by repression values ($F_{O_iR_j}$). This value represents the expression level of a downstream gene in the unrepressed condition divided by the repressed condition and it is obtained directly from measured data (see below and Materials and Methods, section 3.4). Taking the fitness to scale linearly with the repression values, the fitness of

the complete network is denoted by the product of all optimization factors:

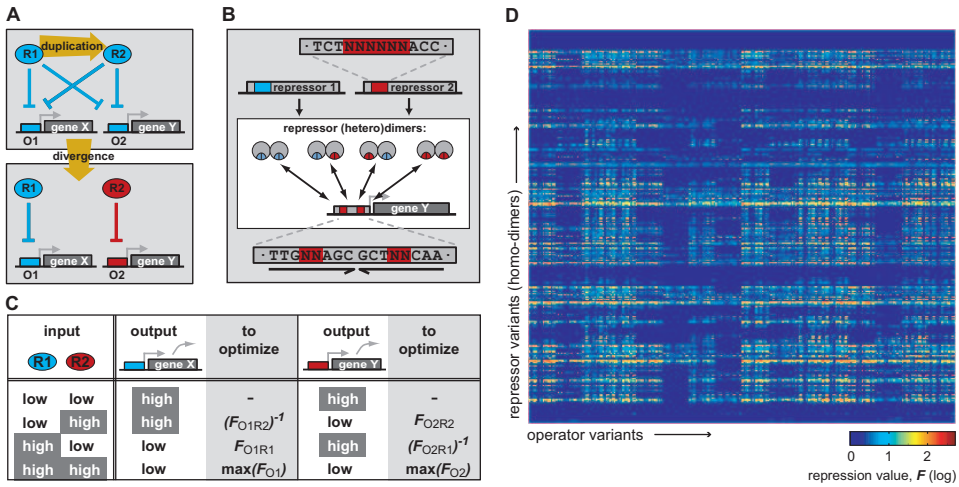
$$\text{Fitness} = \max(F_{O_1}) \frac{F_{O_1R_1}}{F_{O_1R_2}} \max(F_{O_2}) \frac{F_{O_2R_2}}{F_{O_2R_1}} \tag{3.1}$$

In this expression $\max(F_{O_i})$ denotes the repression value of the strongest interaction with O_i, either by homodimers of R_1 or R_2 or the hetero-dimer composed of R_1 and R_2 (see Figure 3.1 and section 3.4).

The fitness definition comes down to a minimum set of two demands for regulatory binding: each operator must bind a repressor tightly ($\max(F_{O_1})$ and $\max(F_{O_2})$ should be large) but also exclusively ($F_{O_1R_1}/F_{O_1R_2}$ and $F_{O_2R_2}/F_{O_2R_1}$ should be large). Prior to divergence the first demand is already met, but the latter is not. The challenge during divergence is therefore to improve binding exclusivity, while maintaining tight binding. Tight and exclusive binding is a core functionality of most regulatory systems, and most pairs of existing transcription factors must therefore score well on the employed fitness definition. Take for instance the LacI and RafR repressors, which regulate enzymes required for growth on lactose and raffinose, respectively. If operator binding would not be tight in the absence of lactose and raffinose, the wasteful expression of the downstream metabolic enzymes would lead to sub-optimal growth speeds [73, 134]. If RafR would also bind to the *lac* operator (and thus bind non-exclusively), the effect on growth speed would also be negative since the mere absence of raffinose would then lead to insufficient β-galactosidase for high lactose concentrations.

One therefore expects a conservative selective pressure that minimally includes binding tightness and exclusiveness, to keep the *lac* and *raf* regulation intact. Important here is that the *lac* and *raf* repressors are in fact related: their origin has been traced to duplication events from a common ancestor [126]. If a conservative pressure keeps their function intact now, it seems a good candidate for the initial divergence pressure as well. Full divergence to the current *lac* and *raf* systems clearly involves many additional developments after duplication. For instance, the divergence of ligand-binding properties [135] might have occurred prior to operator-binding divergence. While these considerations put additional constraints on the entire divergence process, they do not alter the particular operator-binding divergence studied here.

A remaining question still is how the various demands should be weighed in the total fitness. That choice is clearly not general: it will strongly depend on the operons in question and on the changing cell environment. For example, if active RafR is present more than half of the time, then its cross-interaction with the *lac* operator would be comparatively more harmful because it lasts longer. In order to give a uniform presentation we weighed the factors of the four input states equally, which would correspond to an equal contribution of these phases to the overall fitness. However, weighing the factors unequally (e.g., by increasing the power of the tight operator binding, or the cross- interaction factors from 1 to 2) did not alter the main conclusions.



input		output gene X	to optimize	output gene Y	to optimize
low	low	high	-	high	-
low	high	high	$(F_{O1R2})^{-1}$	low	F_{O2R2}
high	low	low	F_{O1R1}	high	$(F_{O2R1})^{-1}$
high	high	low	$\max(F_{O1})$	low	$\max(F_{O2})$

Figure 3.1: Divergence process, fitness criterion, and mutational dataset of repression values. (A) Diagram of the studied divergence process: after a duplication event, a new regulatory interaction can be formed by mutating the two operators, O_1 and O_2, and two repressors, R_1 and R_2. (B) Duplication and divergence yields heterodimers, which can all bind to the operator. The (initially symmetric) operators and repressors are based on the *lac* sequence, as indicated. Base pairs that are key to altering specificity (colored red and blue) can be mutated to arbitrary sequence. (C) The selective pressure for independent regulation follows from four input conditions that contribute to the total fitness. When, e.g., R_1 is high and R_2 low, this implies that X should be low and Y high. Out of all interaction parameters of the network, in this case only $F_{O_1R_1}$ and $(F_{O_2R_1})^{-1}$ are relevant and need to be optimized. When R_1 and R_2 are high, both X and Y should be low, regardless of which repressor-dimer causes repression. Therefore $\max(F_{O_1})$ (the strongest interaction with O_1 by either homodimers of R_1 or R_2 or by the heterodimer of R_1 and R_2) and $\max(F_{O_2})$ need to be be optimized. When both R_1 and R_2 are low, no parameters need to be optimized. (D) Resulting repression value landscape, showing repression values based on actual measurements of mutants.

3.1.2 Mutation data and pathway simulations

In our simulations, the strength of a mutant repressor-operator interaction (as expressed by the repression value F), is assigned using data from mutational analysis [112]. In these experiments, repression values have been determined *in vivo* from the repressed and unrepressed expression levels of a *lacZ* gene, controlled by a mutant *lac* operator and mutant *lac* repressor (see section 3.4). Obviously not all possible *lac* mutants have been constructed. Therefore, a potentially significant limitation of our simulations is the restricted number of base pairs that can be mutated *in silico* and linked to experimental data. At the same time however, while the tightness of DNA binding is the result of the integral protein architecture, surprisingly few base pairs (ten in total) have been found to be important for altering binding specificity [112] (see Figure 3.1B). Focusing on these key base pairs is therefore reasonable for the minimal paths that we are inter-

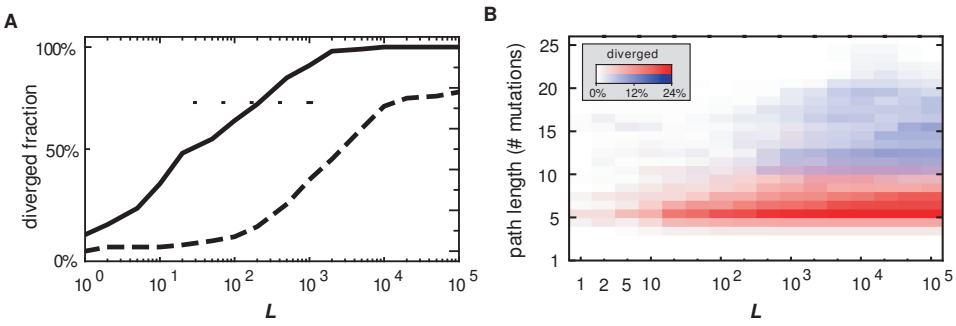


Figure 3.2: Divergence success ratio and path length distributions. (A) Fraction of starting sequences (numbering 132 in total) that successfully diverge, as a function of the number of networks carried to the next round (L). Dashed line, idem, but with the additional requirement of continued tight binding ($F \geq 100$) for both repressors. (B) Distribution of path lengths until divergence. Red color map, optimal co-divergence pathways. Blue color map, pathways with the additional requirement of $F \geq 100$ for both repressors. Note that a vertical summation of the color maps yields the lines in (A).

ested in here. Using measurements on 1,286 mutants, repression values of all variants in these key base pairs could convincingly be determined [93, 112, 136]. These variants thus include all multiple mutants in both the repressor and the operator. Repression values of heterodimers and asymmetric operators are calculated using an additive contribution of the repressor monomers to the dimer-DNA binding [137] (see section 3.4). In total, about $1{\cdot}10^7$ possible repressor-operator combinations are obtained (see Figure 3.1D for the homodimer variants).

Every mutational path starts with the duplicated sequence of a tight binding repressor-operator combination (repression value > 100). These possible starting sequences obviously include wild-type *lac*, but also e.g., the *gal* and *ebg* systems, which are part of the same family of repressors. Their high measured repression values are rather remarkable because the rest of the *gal*, *ebg*, and *lac* sequences have in fact diverged considerably. These observations further indicate that the key base pairs play the central role in specific recognition.

The aim of the simulation method (see section 3.4 for details) has been to reveal the intrinsic possibilities for the divergence of repressor-operator binding, given the measured data and the constraints of single base pair substitutions and no fitness decreases. For this purpose, we search the landscape for optimal paths and study what their limitations and potential are. To trace these optimal paths, all mutants with a single base pair substitution with respect to their parents are evaluated based on the fitness described above, and the best performers are selected for the next round. The number of selected mutants L is varied to assess its effect. We also question whether these optimal paths are not just rare cases, by comparing them with pathways generated by a different simulation method, where a random mutation is accepted with a probability that

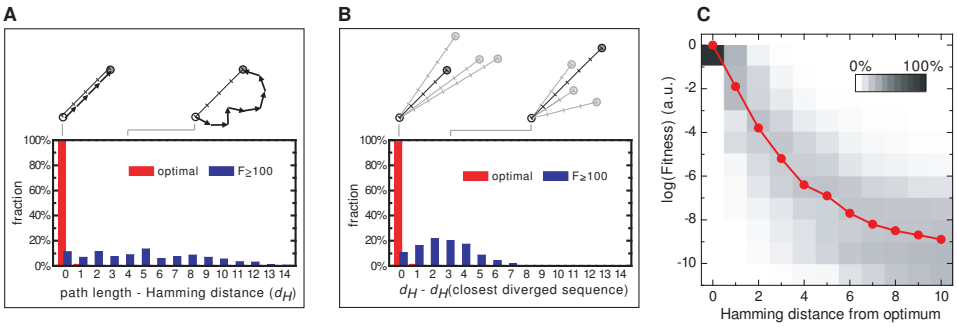


Figure 3.3: Analysis of pathway detours and local environment of fitness optima. (A) Histogram showing the number of detour mutations of the divergence pathways. The Hamming distance d_H of two sequences is defined as the number of positions at which they have different base pairs. Paths that are longer than d_H arrive at an optimum after a detour. (B) Histogram of the Hamming distance between the optimum that is found and the closest optimum. If this measure is zero, a path leads to the closest optimum. (C) Median fitness value as a function of the Hamming distance from a global optimum (solid line). Grey levels indicate the spread of the fitness values.

depends on its associated fitness increase [138] (see section 3.5.1).

3.2 Results

The simulations show that paths to independent recognition are readily found. Even when only the best network is carried to the next round (L = 1), which implies always following the steepest ascent in fitness, some starting sequences can evolve to the highest fitness in the sequence space. In these networks, both repressors bind tightly to one operator ($F_{O_1R_1}$ = 520 and $F_{O_2R_2}$ = 200, respectively), while not at all to the other ($F_{O_1R_2}$ = 1, $F_{O_2R_1}$ = 1). We considered paths to be successful when the fitness value is within an order of magnitude of the highest fitness in the landscape, which is a strict criterion given the fact that the fitness parameter is a product of six factors. The diverged fraction increases for higher L (Figure 3.2A, solid line), which is expected since it allows alternative paths to be explored. More surprising is that successful trajectories can eventually be found from all starting points, but note that paths that can only be followed for higher L are increasingly less probable because they imply more (near) neutral mutations.

Most optimal paths are rather short: 70% require just five to nine mutations for L = 20 (Figure 3.2B). The systems almost exclusively find the nearest diverged state in sequence space (Figure 3.3B) and do so without taking any detours (Figure 3.3A). Notably, despite the fact that the starting points lie in very different areas of the sequence space, a generic sequence of network changes is generally observed (see Figure 3.4 for an example). First of all, one repressor-operator combination remains unchanged, except at the very end, as the other diverges away. This is an example of asymmetric

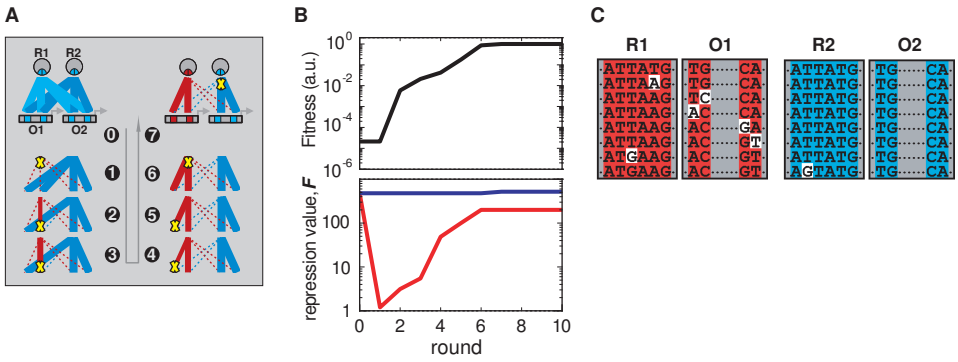


Figure 3.4: Typical divergence pathway: network changes, fitness, and sequence. (A) Evolving interaction network, where line thickness denotes binding strength between repressor monomer and operator-half. Dotted lines denote negligible repression. Yellow crosses indicate repressor and operator mutations, which are positioned at the top and bottom of the interaction lines respectively. (B) Fitness trajectory (black) and corresponding repression of each repressor on its operator (red and blue). Fitness is normalized to the maximum value ($\sim 1 \cdot 10^{10}$). (C) Sequences for each round. Mutated positions are colored white.

divergence due to positive selection, as has also been found in phylogenetic analysis of duplicate genes in eukaryotes [139]. A striking general feature of the pathways is an early reduction in the binding strength of the diverging repressor, brought about by a single base pair substitution (Figure 3.4B, red trace). Such a mutation would be unfavorable for a single repressor-operator pair, but here it can be fitness neutral, partly because the unchanged duplicate repressor ensures a continued repression. At this specific point the diverging repressor is freed from functional constraints and therefore most vulnerable to degenerative mutations resulting in silencing of the gene. The probability of silencing is reduced however, because already at the second mutation and onward, new and unique protein-DNA recognition can be built up. At the sequence level, this phase is characterized by transient asymmetries. The operator must go through non-palindromic sequences because it can only receive one mutation at a time. Heterodimers are the best binders in this phase because of their ability to mirror the non-palindromic operator sequences. Eventually all successful trajectories recover palindromic operators, even as the selective pressure does not explicitly specify this. With all dimer varieties present, a homodimer is available and now binds most tightly to the palindromic operator.

In order to obtain a better insight in the essential ingredients for the observed evolvability, various additional simulations were performed. For instance, we were triggered by the recurrent early knockout of one of the repressors, which is one of the most noticeable features of the mutational pathways. To test for the importance of this step, both repressor-operator pairs were required to maintain a significant repression ($F_{\mathrm{O_1R_1}}$

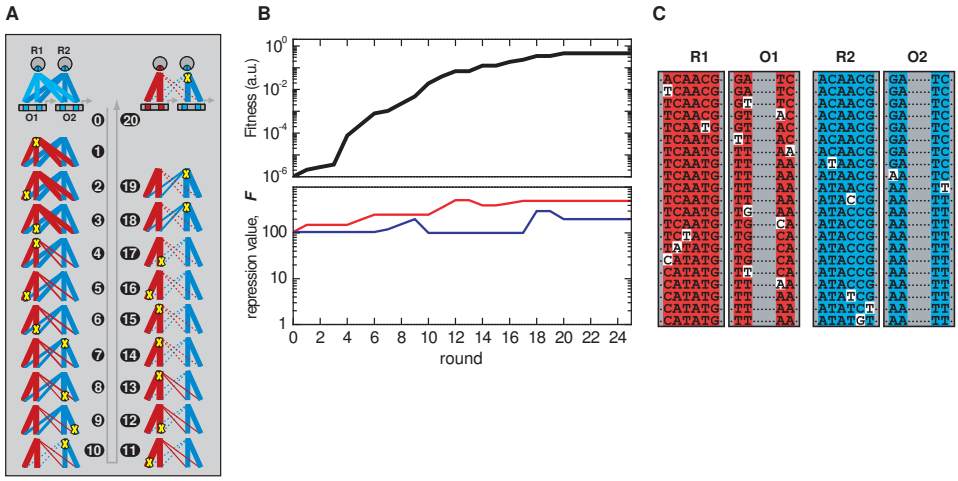


Figure 3.5: Typical divergence pathway, with the additional requirement of continued tight binding of both repressors ($F \geq 100$.) (A) Evolving interaction network, where line thickness denotes binding strength between repressor monomer and operator-half. Dotted lines denote negligible repression. Yellow crosses indicate repressor and operator mutations, which are positioned at the top and bottom of the interaction lines respectively. (B) Fitness trajectory (black) and corresponding repression of each repressor on its operator (red and blue). Fitness is normalized to the maximum value ($\sim 1 \cdot 10^{10}$). (C) Sequences for each round. Mutated positions are colored white.

> 100 and $F_{O_1R_2} > 100$). Divergence is indeed significantly frustrated by these conditions (Figure 3.2A, hatched line). The amount of selected mutants needs to be two orders of magnitude larger ($L > 1{,}000$) for half of the starting sequences to diverge. The saturation of the diverged fraction for very high L, where prolonged neutral drift is allowed, indicates that for 22% of the starting sequences no pathways exist. Moreover, in contrast to the optimal paths, the nearest diverged state in the landscape is generally not found, and the paths contain significant detours (Figure 3.3). The same is seen from the increased path length: 70% of the paths take 11-21 mutations (Figure 3.2B). These paths lack a recurring mutation pattern as observed for the optimal paths and instead show a large variation in the sequence of events. Both repressors and operators are significantly mutated, and the fitness increases slowly or is neutral over multiple rounds (see Figure 3.5 for an example).

Another defining feature of duplicated transcription factors is the heterodimerization of transcription factor monomers. It is not a priori evident whether this constraint on the network topology either promotes or hampers divergence. To assess its effect, simulations were performed where heterodimers are not able to form (data not shown). The results indicated a surprisingly limited effect on the divergence. The paths do initially show a slower fitness increase, but the path length does not appear much affected, nor the success rate of divergence. The other simulation variations we conducted (with

unequally weighted factors in the fitness definition), did not qualitatively alter the main divergence features, such as substantial divergence success without fitness decreases, short paths, and an early repression dip, indicating the robustness of our results.

3.3 Discussion

3.3.1 Duplication and coevolutionary divergence

We obtain a first view on a fitness landscape for regulatory divergence that is based on actual measured data. We show that the landscape allows evolutionary paths toward independent repressor-operator interactions, exhibiting a step-by-step increasing fitness, starting as early as the first or second mutation. Since the possibility of following such paths critically depends on molecular properties, the use of empirical data is essential for such claims. One could also have imagined fitness landscapes where paths to diverged networks do not exist, or where they are very long, involving large detours. Our results contrast with the notion that a number of neutral or even deleterious mutations have to accumulate before a new function can develop (see for a discussion e.g., [140]). Having beneficial mutations available early on is important, since it greatly enhances divergence probabilities [103]. A lack of early selection would result in much higher probabilities of silencing of one of the duplicates by the accumulation of mutations [124, 133].

While the presented systematic search for optimal pathways is useful in revealing necessary conditions for divergence, one may wonder whether paths are not very different in a probability-based fixation process that typifies natural evolution. However, we found that population genetics simulations reveal the same pathway characteristics: a significant fraction of paths are successful with monotonous fitness increases, one repression dip early on, and few neutral mutations are present (see section 3.5.1 and Figure 3.6).

The coevolutionary search for a new and independent recognition, which is relevant for both protein-DNA and protein-protein interactions, comprises fundamental differences with often-considered evolution of ligand-binding and enzymatic activity [141, 142]. While in the latter cases the new evolutionary target is fixed, here it is open-ended: as with locks and keys, many possible combinations are unique matches, and each of those is equally suitable. This large degree of freedom allows the system to choose the solution that is most accessible. Another difference with fixed-target evolution lies in the selective pressure. Binding is already tight to both operators at the start of the coevolutionary scenario, so the initial pressure to change, in fact comes from benefits of not binding another operator. This pressure for unique recognition is characteristic for regulatory interactions but plays much less a role in developing other functions such as enzymatic activity. These characteristics of a coevolutionary mechanism, together with the remarkable plasticity of protein-DNA interactions result in a highly evolvable system.

3.3.2 Fitness landscape funnels

The diversity of molecular architectures is not only constrained by their inherent physico-chemical limitations, but also by the existence of viable evolutionary routes that shape them. For instance, in a population of bacteria there is only a small probability that an advantageous function emerges *if* a temporary fitness decrease is required first. Put differently, the shape of the fitness landscape is critical, and one can readily imagine fitness landscapes where the optima are very difficult to reach. Upon first inspection, the measured landscape we consider indeed contains many potential frustration sources: over 99% of all optima in the landscape are in fact below our divergence criterion. Such local optima represent traps in which the system gets permanently stuck once it encounters one. However, the results show that the system is still guided in the right direction to (near) global optima, which indicates that the fitness landscape contains funnel-like features. Moreover, the optimal paths contain negligible detours (Figure 3.3A) and lead to the nearest optimum (Figure 3.3B), showing that the funneling is efficient and not constrained by ruggedness. A funnel-like organization of the landscape is also supported by the monotonous fitness increases of the probabilistic pathways (Figure 3.6C), as well as by the smooth fitness decrease when stepping away from a global optimum (Figure 3.3C).

The underlying causes for funnels in the fitness landscape may be found at two levels. The first level is that of a single repressor-operator interaction. The surface smoothness that is needed for the funnels may be partly understood from the reported additive contributions of the *lac* amino acids to the binding energy. In mathematical models, additive interactions have been shown to yield smoother fitness surfaces because they can be optimized independently [52].

At a higher level, features of the network topology shape the landscape surface and divergence potential. We found that the tightly interconnected topology, as present after the duplication, does not frustrate divergence but instead promotes it. In contrast to an isolated repressor-operator pair, where a drop in the binding strength decreases the fitness, the same mutation can be neutral in the interconnected topology. Compensation for the decrease in binding strength can be attributed to two features of the topology. First, there is the characteristic pressure to not bind the rival operator: when a mutation decreases an interaction that should be maximized, this negative effect on the fitness is partly balanced by the decrease of an unwanted cross-interaction. A second mechanism is a coevolutionary twist on Ohno's original idea, in which one repressor-operator pair can search for a new recognition, while the other repressor maintains repression on both operators in the very early stages. As we have observed that a drop in the binding strength is necessary for efficient divergence, the ability to compensate for its negative contribution to the fitness is crucial for funneling.

The evolutionary fate of redundant genes has previously been studied primarily using sequence analysis [124, 143]. By using a different dataset and approach, our simulations strengthen recent evidence for a more rapid fixation of mutations in redundant

genes [143] (termed "accelerated evolution"). Our analysis enables a next step in our understanding of this important process: It provides a mechanistic rationale for why such a rapid divergence can indeed occur, in terms of minimal selective conditions bacteria must experience, in combination with independently measured plasticity of protein-DNA interactions. Furthermore it yields a quantitative prediction for the minimum number of essential mutations to achieve divergence.

3.3.3 Suggested experiments

Our results show divergence to be possible with monotonic increasing fitness, which hints at the feasibility of monitoring similar processes in experiments. It has recently been shown that the serial dilution assay, as pioneered by Lenski and coworkers [121], can be employed to adapt bacterial strains to a new condition within weeks [73, 144]. Similarly, one could attempt to evolve a duplicate *lac* repressor/operator copy towards the independent regulation of a second operon. However, this more complex assay does require key modifications: (1) growth and selection of the mutants should occur in alternating media, in analogy to our discussion of multiple input conditions, and (2) a starting network must be engineered that satisfies the conditions for DNA-binding divergence: a duplicate repressor/operator and a selective pressure for tight and independent binding.

In practice, one could place the *lac* operator upstream of the raffinose utilization operon, and construct a lacI duplicate that is sensitive to raffinose. This initial situation is now similar to our simulations: two *lac* repressors bind to the two *lac* operators. The employed fitness definition is also suitable: (1) in media where the two metabolites are both low (supplemented e.g., by another carbon source), the metabolic enzymes should not be expressed. The resulting optimal growth is well represented by positive contributions to the overall fitness by high values for tight binding. (2) When just one metabolite is present, one screens for exclusive binding. In a medium without lactose the lactose-sensitive repressor shuts both operons down if binding is still non-exclusive. Upon mutations that allow this repressor to bind exclusively to the operator of the lactose operon, raffinose metabolic enzymes would be expressed. The resulting faster growth due to raffinose utilization thus correlates well with higher values for exclusive binding. The pressure for a correct behavior under multiple conditions prevents the fixation of trivial solutions that would just work under one condition.

3.3.4 Other network growth scenarios

For biological regulatory networks to grow, not only new components are required, but also new and independent interactions. Next to the coevolutionary duplication-divergence scenario for network growth, alternative models for the creation of new regulatory interactions have been proposed. In the first alternative, a new operator must emerge upstream of the regulated gene in an effectively random DNA sequence [145].

This scenario has mainly been considered for eukaryotes with large upstream regulatory regions and short binding sites. For longer operators in prokaryotes, this scenario requires many neutral mutations before improvements can be selected for (see section 3.5.2), which represents a major evolutionary obstacle.

Another possible source for new regulatory interactions is lateral gene transfer, which is thought to be the source of many paralogs found in prokaryotes [95]. In this scenario divergence would occur while two genes each reside in different organismal lineages (essentially being orthologs at that stage) and each experiencing different selective constraints. Lateral gene transfer unites the diverged genes, resulting in immediate contributions to fitness by both homologous genes. Although examples of this scenario have been found for enzymes [146], transcription factor-operator interactions are a special case, as there is no obvious internal or external selection pressure for their interaction to diverge by itself. Our results illustrate the feasibility of coevolutionary divergence of two transcription factors within a single organismal lineage. These findings are supported by the lack of evidence for horizontal transfer of the *lac* system in *E. coli* [147]. However, this is not to say that lateral gene transfer and duplication-divergence are mutually exclusive. Summarizing, the coevolutionary divergence studied here differs from alternative models of network growth by providing both a high probability of selective advantageous point mutations and a rationale for a divergence pressure.

Finally, it is of interest to consider different selective pressures within the same duplication scenario. While the pressure for independent regulation seems to be a dominant one, as evidenced by the many independent transcription factors that are paralogs, duplications also have yielded other network motifs. An interesting example is the UxuR/ExuR pair of repressors. Like the case studied in the present work, they have originated by duplication and share two operators (see section 3.5.3). However, they seem to have diverged under a different selective pressure, since their cross interaction was not eliminated, but instead has been retained, forming a socalled *bi-fan* motif [132].

This work describes how regulatory network connections can be formed and broken after a duplication event. Our quantitative approach takes the selective conditions and molecular adaptability explicitly into account, and opens up a new angle on the duplication-divergence question that is complementary to existing approaches. Evolution of network connections is treated more abstractly in numerical studies of biological network growth, which have recently received much attention [129, 148, 149]. The use of experimental data will help to perform such studies on a more realistic footing. Finally, the promising new field of experimental network engineering [150–152] and evolution (see e.g., [153]) will also benefit from the quantification of network adaptability.

3.4 Materials and methods

Mutational dataset. In this work we used an extensive dataset of binding affinities of *lac* repressor and operator mutants, obtained by B. Müller-Hill and coworkers. In these experiments, repression values $F_{O_iR_j}$ have been determined in vivo as the ratio of the unrepressed and repressed expression levels of a β-galactosidase (*lacZ*) reporter gene, controlled by a mutant *lac* operator O_i and mutant *lac* repressor R_j. This was done using the standard assay by Miller [154]. Since the β-galactosidase synthesis is proportional to the fraction of free operator (see e.g., [90]), we find for the repression value $F_{O_iR_j} = 1 + [R_j]/K_D$, where K_D is the equilibrium dissociation constant and $[R_j]$ is the concentration of active repressor R_j. The dataset contains repression values of base pair substitutions leading to changes in amino acid residues 1 and 2 of the recognition helix of the *lac* repressor (Y17 and Q18) and base pairs 4 and 5 of the symmetric *lac* operator [155]. These residues and base pairs were found to be most important for altering repressor operator-binding affinities [112]. The dataset covers a considerable fraction of all possible substitutions involving a homodimeric repressor and a symmetric operator (1,286 out of a total of 6,400). Part of this raw data is published in Lehming *et al.* [112]; the full dataset is found in [136]. The contributions of the two repressor amino acids to the repression value were found to be additive. With this knowledge, repression values could convincingly be assigned to all mutants, including those that were not measured [112]. In the present study we use these assigned repression values, all of which are given in [112]. Moreover, we extend the dataset to include heterodimeric repressors and non-palindromic operators (see below), to obtain the complete mapping between sequence and repression values for all possible mutants ($1{\cdot}10^7$) in the key repressor residues and operator base pairs.

Repression values of heterodimers and non-palindromic operators. We consider the repressors to act as dimers. After their duplication, once the repressors genes are mutated, this leads to heterodimerization of distinct monomers. While heterodimer binding strengths (F_{He}) have not been directly measured, they can be derived from the two corresponding homodimer repression values (F_{Ho_1} and F_{Ho_2}), measured on a palindromic operator. The heterodimer binding energy ΔG_{He} is the sum of the monomer-monomer and the dimer-operator binding energy. Simple equilibrium considerations lead to the following expression, where [R] in this case is the total concentration of repressor subunits:

$$F_{He} = 1 + [R]^2 e^{-\Delta G_{He}/kT} = 1 + \sqrt{(F_{Ho_1})(F_{Ho_2})} \tag{3.2}$$

With this equation, repression values involving non-palindromic operators are also automatically taken into account: each dimer-operator interaction is built up additively [137] from two interactions between a monomer and an operator-half. In this derivation the dimerization free energy was assumed to be fixed, since it does not directly affect the specificity by which the repressors recognize their operators. The het-

erodimer repression value then becomes independent of the dimerization energy.

Optimal pathway simulations. Each repressor monomer is represented by six base pairs (two amino acid residues), and each operator by four base pairs, which are key to specific binding. The complete network with duplicates is thus represented by 20 base pairs. Each simulation run starts with the duplication of a tight binding repressor-operator pair, having a repression value of 100 or higher. Out of all possible repressor-operator combinations (homodimers and palindromic operators), there are 132 fulfilling this condition. Changing this threshold did not significantly alter the outcome of the simulations. In order to avoid any bias due to codon usage of the starting repressor, separate simulations were run starting from each of its synonymous codon versions. These simulations were averaged to produce the presented results.

In order to determine the optimal mutational pathways in the fitness landscape, an evolutionary algorithm was employed. Beginning with one of the starting sequences, each round we generated all mutants that differ by one base pair (60 in total). Of each mutant network, the strength of all eight possible interactions was determined (see Figure 3.1B where four possible interactions are schematically shown between the repressor dimers and one of the two operators). Interactions between repressor homodimers and palindromic operators were directly assigned from the published repression values [112]. Other interactions were calculated from the measured data as described above. Next, we selected the best L networks to the next round based on a fitness parameter that is directly calculated from the interaction strengths (see equation 3.1). The next round started by again generating all single base pair mutants of the L selected networks. The effect of L was assessed by varying it between 1 and 10^5. Decreasing fitness steps were not allowed, and in case of equal fitness, parents were ranked above their offspring. These rules make divergence harder because they constrain the space that can be explored. The evolutionary cycle was repeated until the fitness could not be further improved. Pathways were considered to be successful when the fitness came within a factor 10 of the highest fitness in the landscape.

3.5 Appendix

3.5.1 Simulation of mutational pathways incorporating probabilistic population dynamics

Here results are presented of a second simulation method, where mutations are fixed with a probability that is based on the associated fitness increase (see methods below). Compared to the optimal pathway simulations, this probabilistic approach does not search the landscape as systematically, but it is arguably closer to natural evolution, in that the fixation chance of mutations with no or lower fitness increases is more well-defined [103].

We find that the key characteristics of the probabilistic pathways are very similar

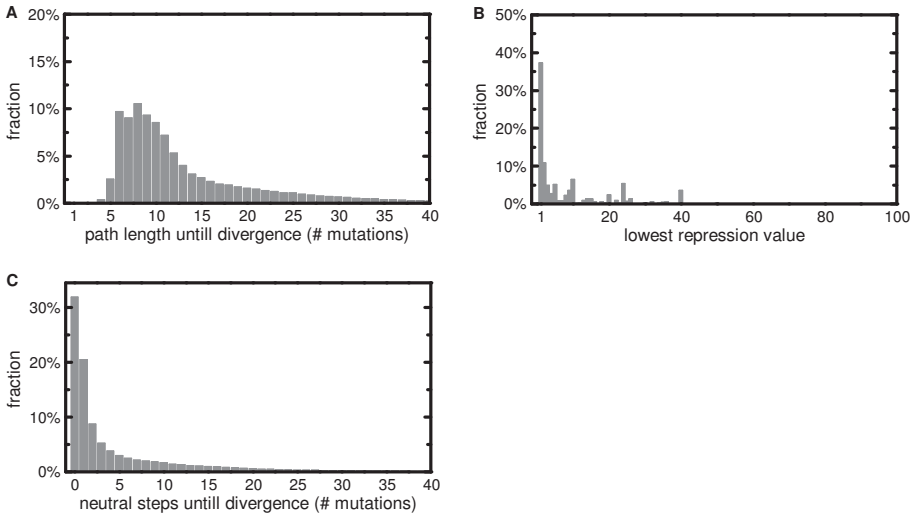


Figure 3.6: Qualitative features of successfully diverging paths in the probabilistic pathway simulations. Simulations were performed with a 5% growth advantage of a diverged network over the initial duplicate network, and a population size of 10^5. Of all traced paths, 17% successfully diverged, despite the strict requirements that promote trapping in local optima (fitness cannot decrease). Relaxing these conditions would lead to larger divergence probabilities. (A) Histogram showing the number of base mutations until divergence for the successful pathways. (B) Histogram showing the lowest repression values of each repressor on its operator during the successful divergence pathways. (C) Histogram showing the number of neutral mutations that occur until the pathways successfully diverged.

to those of the optimal pathways. From every starting condition it appears possible to diverge towards independent binding while the fitness increases monotonously along the way (17% of all paths). The success rate logically differs for the different starting sequences, but all of them can yield successful trajectories. Looking at the probabilistic paths in more detail (Figure 3.6), we see that they are somewhat longer, but half of them still diverge within 10 mutation steps. And although the sequential network changes are not as uniform, the paths are still characterized by few neutral mutations (0 or 1 neutral steps for 50% of the paths) and an early reduction in repression of one repressor (F<5 for 50% of paths).

Our probabilistic model allows us to vary the amount of drift present in our pathways by varying both the effective size of a population (N) and the growth advantage that diverged networks have over un-diverged networks ($s_{\max}$). Taking a conservative growth advantage of 5% [73] we simulated probabilistic pathways for population sizes ranging from 10^3 to 10^8. At high population sizes pathway characteristics remain very similar. Only at population sizes below 10^4 we start to see a strong effect of genetic drift: more neutral mutations, and hence longer pathways. However, the fraction of diverged pathways and the reduction in repression of one of the repressors does not significantly

change.

In our probabilistic model we do not allow disadvantageous mutation to be fixed. An important finding we present in this work is that for divergence to occur, fitness drops are actually not necessary. (Note that if one would allow drops in fitness, all pathways would reach divergence eventually, since trapping in local optima would then not be possible.) Even if disadvantageous mutations would be explicitly modeled, we would not expect to find other pathway characteristics for population sizes above 10^4, as these mutations have very low fixation chances compared to the readily available beneficial mutations present in our system.

Probabilistic pathway simulations. To model the effect of genetic drift in the evolutionary pathways, probabilistic simulations were performed. Thousand such pathways were traced for each of the starting sequences. In each simulation step 60 different single base pair substitutions are possible (the 20 base pairs can each mutate to 3 different bases), which were assumed to occur with equal probability. The fixation probability of each specific mutation depends on its associated fitness increase, and was calculated with a standard and simple population genetics model [103, 138] (and see below). In each simulation step, the fixation probabilities of all 60 possible single base pair substitutions were calculated, and one substitution was randomly chosen according to its share in the total probability. Each path was continued until a (possibly local) optimum was reached, so that the fitness could not be further improved. The purpose of this Monte Carlo-like scheme was to check whether biased random walks show similar features as the ones generated by our optimal pathway simulations.

Mutations that decrease the network fitness were assumed not to be able to fix, while those that keep the fitness constant have a fixation probability of $1/N$, where N is the effective population size. Mutations that do increase the fitness have a fixation chance of $2\Delta s$, where Δs is the selective advantage that this fitter mutant has over its parent. In our simulations we let an increase in the fitness parameter by a factor of 10 correspond to a Δs of 1%. In this way, a successfully diverged network has a selective advantage of 5% ($s_{\max}$) over the initial duplicated state (fitness rises from 10^{-6} to 10^{-1}, see main text), which matches typical experimentally observed growth advantages [73, 134].

In this probabilistic scheme a mutation conferring a selective advantage Δs will have $2N\Delta s$ times more chance to be accepted than a neutral mutation. Therefore, both the effective population size and the defined selective advantage influence the effectiveness of selection. By either lowering N or $s_{\max}$ the amount of genetic drift in the model increases. We typically simulated an effective population size $N = 10^5$, together with a fixed $s_{\max} = 5\%$. The population size needed to be lower than 10^4 before genetic drift substantially increased the number of neutral mutations in successful paths.

3.5.2 Comment on neutral mutations required for the emergence of a new operator

Here we consider an alternative mechanism for creating a new regulatory interaction, where a new operator must emerge in an effectively random DNA sequence. In a typical prokaryotic case, like e.g. the *lac* system, a 20 base pair operator has to emerge somewhere in a roughly 100 base pair region in order to effectively block RNA polymerase binding. Then there are 10 expected prior matching base pairs for the best binding site within the promoter region. However, experimental data suggests that at least 15 base pairs need to match before appreciable binding is achieved [156, 157], from which point further mutations can be positively selected for. This means that more than 5 base pairs need to be optimized without selection, while the coevolutionary pathways can be selected for almost immediately.

3.5.3 Alternative selective pressures and the *Escherichia coli* regulatory network

Here we briefly comment on the possibility of alternative selective pressures after a duplication event. In the main text we have considered a selective pressure for independent regulation of two operator sites by two repressors. But other selective pressures could result in different network topologies, or motifs [132, 158]. In the case of the regulators of the Lac/Gal family it is clear that cross-interactions between the operons should be eliminated: each operon should be transcribed only if the relevant carbon source is present. In other cases the cross-interactions might in fact be desirable and therefore not be selected against. For instance the so-called *bi-fan* motif (see ref [132] and Figure 3.7) might then originate from a duplication of a gene that regulated multiple genes before duplication.

In order to see whether this scenario could have materialized in the *E. coli* regulatory network, one would need to search for *bi-fan* motifs where homologous transcription factors share their operators. Interestingly, these cases are readily found. An example could be the system of repressor genes *uxuR* and *exuR*. The UxuR/ExuR repressors are the regulators of genes involved in the transport and catabolism of fructuronate and glucuronate [159–161]. They are highly homologous (see Figure 3.9 and [125]) and indeed are very likely to bind to shared operator sites, with different affinities (see [162, 163] and Figure 3.8). Moreover, they are found to be able to form heterodimers [164] and can partially substitute for one another [162, 164]. As the enzymes from the UxuR/ExuR regulons have overlapping functions, there is a rationale for retaining the cross-interactions [161].

Other selective pressures for different topologies after duplication could be imagined, for example where one cross-interaction is eliminated (see Figure 3.7). In the case where global regulators become duplicated, one could expect homologs to be present in the so-called *Dense Overlapping Regions*. Alternatively, when more than one dupli-

cation event has taken place, a '*tri-fan*' (see Figure 3.7) would seem a possible outcome. Remarkably, in the *E. coli* regulatory network all of these topologies can be found, with the additional conditions that the transcription factors are homologous and their operators are shared.

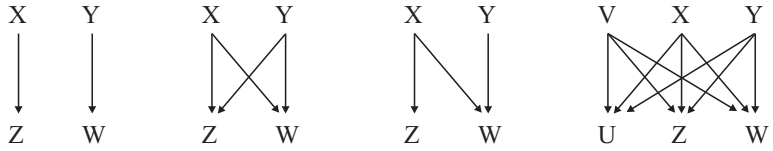


Figure 3.7: Examples of network topologies that might arise after duplication of a regulatory gene. From left to right: independent regulators, the *bi-fan*, a *bi-fan* where one connection has been lost, and a '*tri-fan*'.

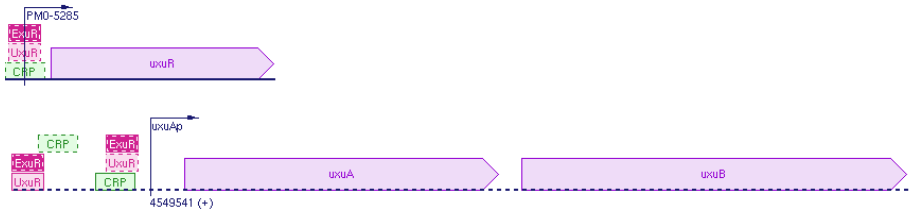


Figure 3.8: Overlapping operators and the homologous repressors ExuR and UxuR (from EcoCyc, [92]).

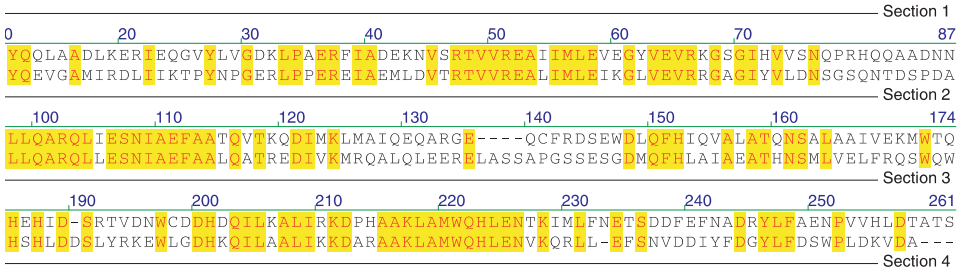


Figure 3.9: FAST alignment of the amino acid sequences of ExuR and UxuR shows 44% homology.

CHAPTER

4 Adaptive landscapes of gene regulatory systems in variable environments

I knew their tremendous possibilities, and I have no doubt I could have speeded up their evolution, perhaps by some millions of years. But for what good?

G.G. Simpson,
The Dechronization of Sam Magruder

Adaptation to variable environments is a fundamental issue in evolutionary biology. We determined the phenotype-fitness landscape for mutant LacI repressors in alternating environments, using an operon with tunable cost and benefit. We found that non-linearities in the relation between expression and growth critically affect adaptation: they alter the competition between specialists, regulated and non-regulated generalists, and can result in weak selection on regulation despite strong alternating pressures. Using random mutagenesis, we showed that LacI adapts to an unfavorable alternating environment according to the predicted landscape, resulting in novel inverse LacI phenotypes. We identified a local adaptive constraint by mapping intermediate phenotypes on the fitness landscape. The adaptation towards more complex regulatory functions was demonstrated using a small genetic network. This study shows that a functional insight into phenotypic responses is central to understanding adaptation in variable environments.

Although regulation is central to cellular behavior, the mechanisms by which regulatory systems evolve remain poorly understood [69, 165–167]. A central obstacle is the complex and unknown relation between environment, phenotype and fitness [55, 56, 58]. In temporally varying environments, fitness depends on multiple environmen-

tal states and the phenotypic changes they induce, but also on the timescales of variation, and the strength of trade-offs experienced across environments [42, 69, 168, 169]. Moreover, it can be challenging to distinguish regulated from non-regulated generalists [170], because a functional understanding of many phenotypes is lacking.

These complications have made it difficult to quantify the selective pressure driving regulatory adaptation, leaving many open questions about its constraints [36, 37, 61]. For instance, do regulatory phenotypes emerge under varying selective pressures, or are specialists maintained, possibly as a balanced polymorphism [68]? When stasis is observed in the phenotypic response to environmental variation, either in nature or in laboratory experiments, is this due to constraint or rather to weak selection on regulation?

To study the relation between gene regulation phenotypes and fitness, one may measure the overall growth rate of regulatory mutants in alternating environments, as has been reported for two repressor mutants [42]. However, this method is less suited for a comprehensive mapping between phenotype and fitness. Here, we employed an operon with a tunable cost and benefit of expression, which allowed us to determine phenotype-fitness landscapes for various temporally alternating environments, and to study adaptation towards novel regulatory functions and its potential constraints.

The operon, harbored by *Escherichia coli*, consists of three co-regulated genes (Fig. 4.1A): expression of *sacB* from *Bacillus subtilis*, confers a cost in the presence of sucrose, expression of *cmR* confers a benefit in the presence of chloramphenicol (cm), while *lacZα* allows measurement of the operon expression level (see sections 4.1 and 4.2.1). The fitness (growth rate) depends both on the concentration of selective agent S (sucrose or cm) and on the operon expression level E, and is described by the function $G(E,S)$ (Fig. 4.1B). Operon expression was initially controlled by the native *lac* repressor LacI, as described by the function $E(I)$, where I is the concentration of the inducer isopropyl-β-D-thiogalactopyranoside (IPTG). Importantly, the ability to vary I and S independently enables the experimental determination of $G(E,S)$. In a second set of experiments, where LacI is mutated, $E(I)$ may change while $G(E,S)$ remains fixed. Operon expression was controlled by a small network composed of the *tet* and *lac* repressors in the last experiments, making E dependent on two input signals, IPTG and doxycycline (Dox). Because cost and benefit are generic aspects of gene expression [73, 134], the selective pressures applied by the operon are relevant for regulatory adaptation in general[1].

To quantify the selective pressure on operon expression, we measured the growth rate as a function of expression level by induction with IPTG (Fig. 4.1). Media containing sucrose yielded a growth rate that gradually decreased to negative values for higher expression (Fig. 4.1C). Increasing sucrose concentrations led to sharper growth decreases. The dependence of growth on expression could be described by a reaction

[1] For many repressible catabolic operons the expression cost is significantly smaller than the benefit [75]. However, when the catabolite is only rarely present the total (time-integrated) cost and benefit may be of similar order, as is the case in our system.

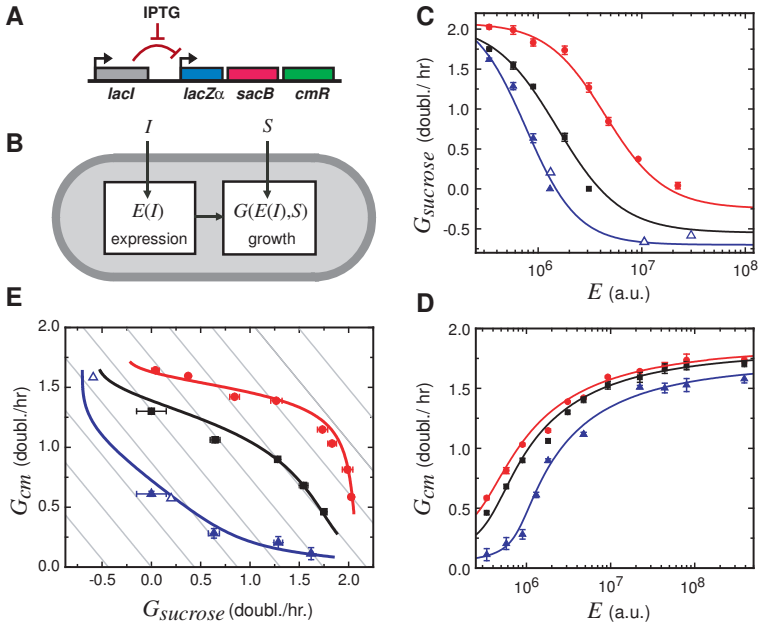


Figure 4.1: (A) Schematic of operon with a tunable cost and benefit of expression. Expression of *sacB* reduces growth in sucrose media (cost), expression of *cmR* facilitates growth in chloramphenicol (cm) media (benefit), and *lacZα* allows expression levels to be quantified. All genes are initially co-regulated by the wild-type *lac* repressor (B) Functional representation of the system. The function $E(I)$ describes the dependence of expression on the IPTG concentration I. The function $G(E(I),S)$ describes the dependence of the growth rate on $E(I)$ and the concentration of selective agent S (sucrose or cm). (C) Measured growth rates as a function of expression level, in the presence of 0.15% (red), 0.25% (black), and 0.40% (blue) sucrose (w/v). Open triangles: negative effective growth rates determined by relative performance assays (section 4.1). (D) Measured growth rates as a function of expression level, in the presence of 25 μg/ml (red), 40 μg/ml (black), and 80 μg/ml (blue) cm. Curves in (C) and (D) represent a growth model based on reaction kinetics (section 4.2.1). (E) Growth rate in cm media as a function of growth rate in sucrose media, for the same expression levels in both media. Depending on the sucrose and cm concentrations the trade-off curves can be either concave (red, black) or convex (blue). Diagonal lines are isoclines indicating constant average growth rate for an alternating environment with equal periods.

kinetics model that incorporates sucrose import and sucrose polymerizing activity of levansucrase (section 4.2.1). Media containing cm exhibited the opposite effect, with growth rates increasing for higher operon expression (Fig. 4.1D). Increasing the cm concentration led to sharper growth increases. This data was similarly modeled by chemical rate equations.

Having determined the relationships between expression and fitness, we could predict the relative performance of LacI mutants in a constant environment on the basis of their altered expression. For instance, in a medium with 0.40% sucrose but without

IPTG, the growth rate of a constitutive LacI mutant that yields a high expression level, is lower by about 2.4 doublings/hour (Fig. 4.1C) compared to wild-type, which has a low expression in this medium. This suggests that in a mixed population containing both variants, the latter will be enriched by a factor of $2^{(6x2.4)} = 1.8{\cdot}10^4$ during 6 hrs of growth. Such differences in growth rate are here referred to as the selective pressure. Note that the maximum growth rate difference (for high and low expression), does not depend strongly on the sucrose concentration. Rather, increasing the sucrose concentration pushes the favorable expression levels (conferring high growth rates) to lower values, which we will denote as more stringent selection. Similarly, media with increased cm concentrations favor higher expression levels, and are referred to as more stringent.

In environments that alternate between sucrose and cm, the selection of non-responsive (fixed expression) phenotypes is governed by a trade-off: high expression results in rapid growth in the presence of cm, but slow growth in the presence of sucrose. Conversely, low expression yields slow growth in the presence of cm, but rapid growth in the presence of sucrose. Plotting the growth rate in one medium versus the growth rate in the other, for each expression level, yields a so-called trade-off curve (Fig. 4.1E), which is analogous to the notion of Pareto optimality[2] as used in economics and in engineering. The expression level that confers maximum growth can be determined using the diagonal isoclines that indicate the average growth rate when residing equally long in each of the two environments.

The graphical method described above was originally introduced by Levins [69] and has since been widely used in evolutionary theory (e.g. [171, 172]) to explain two possible strategies: specializing to one environment is optimal when the trade-off curve is convex, whereas a concave curve favors generalists that do moderately well in both environments. Here the trade-off curves can be determined experimentally (Fig. 4.1E) and their shape can be rationalized: stringent selection yields convex curves because in the two media the favorable expression levels are well separated (low in sucrose and high in cm), while less stringent selection yield concave curves because the favorable expression levels partially overlap (Fig. 4.1CD). These trade-off characteristics may well be more general, since they originate from generic non-linearities of the underlying reaction kinetics.

The constraint delineated by the trade-off curve can be overcome, if cells are able to adjust their expression level in response to the environment. In Fig. 4.1E, this implies escaping from the trade-off curve towards optimal growth under both conditions, as represented by the upper right corner. Note that responsiveness is analogous to 'phenotypic plasticity' as generally employed in evolutionary ecology [169, 173]. When al-

[2] Interestingly, the trade-off curves shown here are conceptually similar to the economical and game-theoretical notion of Pareto-optimality, which describes the solutions in a multi-object optimization where no criterion can be improved without simultaneously degrading an other (thereby pointing at a constraint in the system). In our system the non-responsive phenotypes can improve their performance in one environment only by decreasing it in the other, and therefore constitute a Pareto optimal front that can only be overcome by developing regulation.

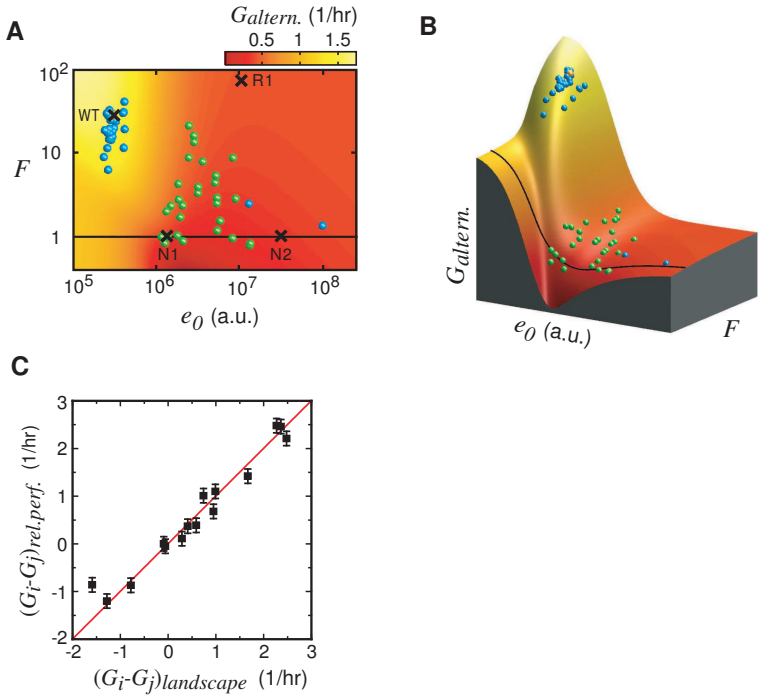


Figure 4.2: (A) and (B) Fitness landscape for gene regulatory phenotypes in a variable environment. An environment is considered that alternates between two media (medium 1: 0.40% sucrose; medium 2: 80 μg/ml cm with 1 mM IPTG), leading to opposing expression demands. Phenotypes are characterized by their basal expression level e_0 in medium 1, and the fold change in expression F when shifting to medium 2. Displayed is the average growth rate G_{altern} (fitness) that is derived from $G(E,S)$ (Fig. 4.1CD), as a function of e_0 and F. In this environment, the wild-type LacI phenotype (WT) performs near optimally and is located on the peak. Other crosses are mutants R1, N1, and N2 that are used with WT in relative growth assays. The spheres are randomly chosen isolates from the following populations: randomly mutated LacI phenotypes prior to selection (green), and after one cycle of sucrose and cm+IPTG selection (blue). (C) Growth rate differences $(G_i - G_j)$ of LacI mutants (WT, R1, N1, and N2) as determined by relative growth assays versus predictions based on the fitness landscape.

ternating between two media, all phenotypes can be represented by two parameters: e_0, the basal expression level in one environment, and F, the factor by which expression changes in the other environment. A fitness landscape is obtained by plotting the average growth rate for the two media as a function of e_0 and F (Fig. 4.2AB, for 0.40% sucrose and 80 μg/ml cm). For simplicity the growth time in each medium is equal and much longer than the system response time, though other cases may be considered (Fig. 4.17).

The landscape exhibited one peak (low e_0/high F), two plateaus (low e_0/low F, and

high e_0), and two optima for non-responsive phenotypes (low and high e_0 at F=1). The optimal responsive phenotypes perform significantly better than the best specialists (low e_0 at F=1), although a large area of responsive phenotypes is seen to perform worse. Upon reducing the selection stringency (0.15% sucrose and 25 μg/ml cm) the peak broadens and partially overlaps with the F=1 line, resulting in a single optimum for non-responsive phenotypes (Fig. 4.16C, intermediate $e0$ at F = 1) , and a reduced advantage of responsive over non-responsive phenotypes. The latter is also seen in Fig. 4.1E, where the concave red trade-off curve is closer to the optimal responsive phenotype (upper right corner) than the convex blue curve. Interestingly, this illustrates that the selective pressure on regulation in alternating media can be weak, even though the selective pressure on expression is large in each medium separately (the growth rate difference for high and low expression is large).

We tested the phenotype-fitness landscapes (Fig. 4.2AB) using two responsive (WT, R1) and two non-responsive mutants (N1, N2), as obtained by random mutagenesis. When the cm medium is supplied with inducer (1mM IPTG), but not the sucrose medium, wild-type LacI exhibits the appropriate response and is thus positioned on the peak, while the other mutants have lower fitness (Fig. 4.2A). Several experiments were performed, in which two mutants were grown together for 6 hours, while the change in their relative abundance was monitored by plating. The data were in good agreement with the fitness landscape predictions (Fig. 4.2C). The measurements also confirmed that responsive phenotypes do not always outperform non-responsive ones. For instance, N1 out-performed R1 by 0.47 doublings/hour when grown in 0.15% sucrose followed by 25 μg/ml cm. However, increasing the selective stringency (0.40% sucrose and 80 μg/ml cm) reversed these roles, favoring R1 over N1 by 0.23 doublings/hour. This dependence on selection stringency is directly apparent from changes in the phenotype-fitness landscapes (Fig. 4.16B and D).

We investigated alternating selection on a diverse population, using a pool of $\sim 5 \cdot 10^6$ random LacI mutants, which have 3 base substitutions on average (section 4.1). A randomly chosen sample of 35 mutants appeared well separated from wild-type LacI, having higher e_0 and lower F, with some non-responsive to IPTG (F=1) (green spheres, Fig. 4.2A and 2B). The population was grown for 6 hours in the sucrose medium and for 6 hours in the cm+IPTG medium, totaling ~24 generations at the maximum growth rate. Much faster alternation would lead to dominant transient effects, effectively averaging the two media [42], while much slower alternation may lead to a loss of population diversity and specialization. Guided by the trade-off analysis, we chose stringent selective conditions that result in a large selective pressure on regulation (0.40% sucrose and 80 μg/ml cm). After selection, the population clustered in the e_0-F plane around a point that co-localized with the fitness optimum, illustrating the accuracy of the adaptive landscape, as well as efficient enrichment (Fig. 4.2AB).

So far we studied the principles of alternating selection on regulatory responses that differ only in magnitude to wild-type LacI. Organisms may also be confronted with en-

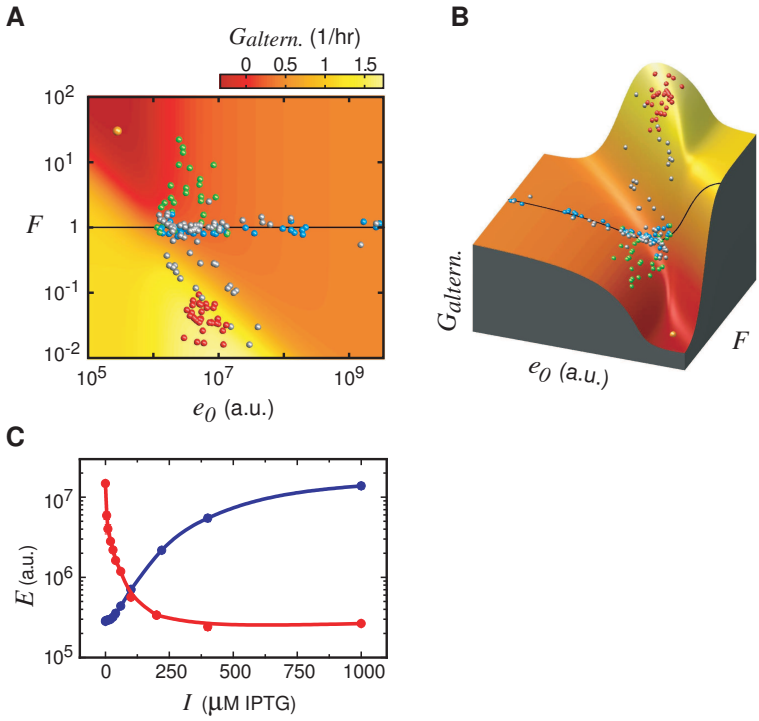


Figure 4.3: (A) and (B) Adaptation towards a novel regulatory phenotype. An alternating environment is considered that is maximally unfavorable for wild-type LacI phenotypes (medium 1: 0.40% sucrose with 1 mM IPTG; medium 2: 80 μg/ml cm). Phenotypes have basal expression level e_0 in medium 2, and a fold change in expression F when shifting to medium 1. Displayed is the average growth rate G_{altern} (fitness) that is derived from $G(E,S)$ (Fig. 4.1CD), as a function of e_0 and F. Note that the landscape has been rotated in panel B for a clear view. In this environment, the wild-type LacI phenotype (WT) is located in the valley. The spheres are randomly chosen isolates from the following populations: randomly mutated LacI phenotypes prior to selection (green), after a first (blue spheres), second (grey spheres) and third (red spheres) cycle. Each cycle involves random mutation of *lacI*, 6 hours of growth in the sucrose+IPTG medium, and 6 hours in the cm medium. Adaptation occurred in accordance with the predicted fitness landscape, leading to the optimum, and resulting in inverse LacI phenotypes (C) Measured induction curves of the wild-type regulatory phenotype (blue), and an adapted inverse LacI phenotype (red).

vironments that demand qualitatively different responses. Here we investigated a maximally unfavorable alternating environment for wild-type LacI: sucrose with IPTG (1 mM), and cm without IPTG. e_0 now indicates expression in the cm medium, and F is the fold change in expression when shifting to the sucrose+IPTG medium. In the corresponding phenotype-fitness landscape (Fig. 4.3A and 3B), wild-type LacI is now positioned in a valley, while the fitness optimum is seen at F values below 1. The phenotype at that optimum would have to achieve tight repression with IPTG, and high expression

without IPTG. IPTG would thus act as a co-repressor instead of an inducer. Within the LacI family of transcriptional regulators [94], such a function has been adopted by the purine repressor, with guanine acting as a co-repressor [174].

After a single cycle of sucrose+IPTG and cm selection on a population of random LacI mutants, the isolates showed no inverse function, but instead clustered along the F=1 line in the e_0-F plane (Fig. 4.3AB). After a second cycle of LacI mutagenesis and selection in both media, some isolated phenotypes did appear below F=1 (Fig. 4.3AB, grey dots). These improvements, together with the position of the cluster at F=1, suggests a local constraint [37] due to a limited access to F <1 phenotypes in combination with a high probability of generating non-responsive mutants. Some isolates already outperform the best specialist phenotypes (F=1, low e_0), which have appreciable fitness and might have presented an adaptive challenge. After a third cycle, the fitness optimum was reached (Fig. 4.3AB, red spheres), yielding inverse LacI phenotypes with expression ratios of around 100 (F ~0.01, Fig. 4.3C). Note that we did not observe phenotypes on the low e_0 flank of the peak, which may indicate a constraint in increasing the repressor-operator interaction strength.

The inverse *lacI* sequences showed a substantial diversity (section 4.2.2), revealing several genetic solutions to the same evolutionary challenge. Mutations were spread over the complete *lacI* coding sequence, although none specifically affected DNA-protein interaction. No direct indications were found for the molecular mechanism, which may be based on increased aspecific binding [175] or on altered allostery (see e.g. [174, 176]). However, all sequences contained mutations at the interface between repressor monomers. One recurring mutation (Ser97Pro) did not yield an inverse LacI phenotype in isolation [177], which hints at epistasis in the system, although a mutational bias cannot be excluded at the moment.

Cells can integrate multiple environmental signals using regulatory systems composed of multiple regulatory proteins. Whether such increased network complexity leads to adaptive constraints [55, 59] was studied using a *lacI* and *tetR* repressor network controlling the expression of the operon (Fig.4A). After construction we measured the operon expression function as a function of the two inducers Dox and IPTG (Fig. 4.4 B)[3]. The complete network, including regulatory and coding sequences, was subsequently randomly mutated followed by selection for a novel expression function, which involves growth in four media with different combinations of selective agent and inducers (Fig. 4.4C and D). The earlier trade-off analysis (Fig.1E) remains relevant: to achieve selection for regulated phenotypes, the favorable expression levels in the sucrose and cm media must be well separated.

After two cycles of mutation and selection, the two resulting phenotypes (Fig. 4.4E and 4F) were similar to the target expression pattern of Fig. 4.4C and D. The regulatory circuits thus adapted in accordance to the selective pressures of all four media, reveal-

[3] As has been observed previously for similar networks [153, 178], the measured input-output relation for the network as constructed did not match the one expected from the topology, indicating some limitations to the rational design of networks.

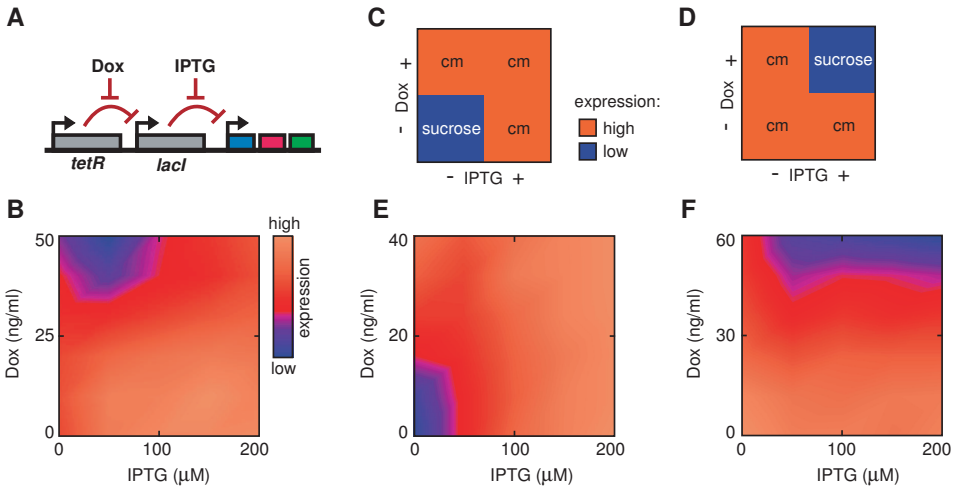


Figure 4.4: Adaptation of a regulatory circuit. (A) Schematic of the *tetR* and *lacI* repressor circuit, which controls operon expression as a function of inducers doxycycline (Dox) and IPTG. (B) Measured expression level in color, as a function of Dox and IPTG concentrations, for the circuit as constructed. Expression range: $5.0{\cdot}10^7$ to $2.7{\cdot}10^8$. (C and D) Schematic representation of two experiments in which the environment alternates between four media. The composition of the four media are indicated, as well as their corresponding favored expression level in color (low or high). In the first environment an 'OR' relation between Dox and IPTG is favored. In the second a 'NAND' relation is favored. (E) Measured expression level in color, as a function of Dox and IPTG concentrations, for a circuit adapted to an OR-environment. Expression range: $8.5{\cdot}10^6$ to $1.4{\cdot}10^8$. (F) idem, for a circuit adapted to a NAND-environment. Expression range: $1.1{\cdot}10^8$ to $3.1{\cdot}10^8$.

ing no insurmountable adaptive constraints. None of the observed mutations were located in regulatory sites on the DNA, which are often considered to be a main source of regulatory network plasticity [179]. The results emphasize the possibility of generating novel regulatory functions via structural changes in transcription factors.

This study underscores the value of obtaining a network-level, functional understanding of phenotypes when studying their adaptation. It allows one to identify the phenotypic parameters on which selection acts, to measure their relation to fitness, and to disentangle causes of constraint. This approach can be applied to a wide range of other open issues, such as the adaptation in spatially heterogeneous environments [180] and the role of network topology in adaptation [59]. The controlled shaping of regulatory networks by evolutionary methods also provides a complementary method to their rational design [181–183].

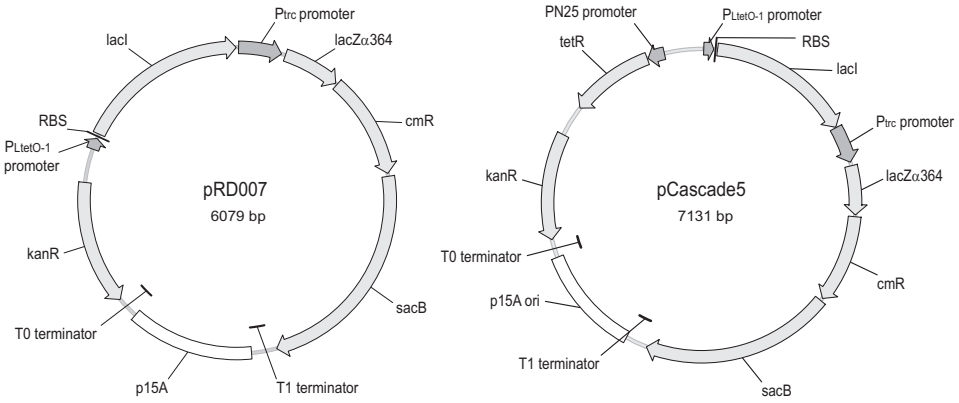


Figure 4.5: Plasmids pRD007 and pCascade5, in which the selection module is under control of *lacI* or a *tetR-lacI* regulatory cascade. The selection module consists of lacZα364 (expression marker), cmR (chloramphenicol resistance), and sacB (levansucrase).

4.1 Materials and methods

Strains. In all selection experiments *Escherichia coli* K12 strain MC1061 [184] was used, which carries a deletion of the complete *lac* operon. Genotype of MC1061:
F$^-$ Δ*lacX74 mcrB1 e14*$^-$(*mcrA0*) *rpsL150*(Str^R) *galE15 galK16* Δ(*ara, leu*)7697 *ara*Δ139 λ$^-$ *hsdR2*(r_k^-, m_k^+) *spoT1*
This strain was obtained from Avidity LLC, Denver CO, USA, as electrocompetent strain EVB100 (containing an additional chromosomal birA gene).

For colony counting purposes, after the relative performance assay (see below), strain DH10B [185] was used. Genotype of DH10B:
F$^-$ *ϕ*80d*lacI*^{q}*Z*Δ(M15) Δ*lacX74 deoR recA1 endA1 mcrA* Δ(*mrr hsdRMS mcrBC*) *nupG rpsL*(Str^R) *galU galK* Δ(*ara, leu*)7697 *ara*Δ139 λ$^-$

Plasmids. We constructed two plasmids based on the pZ vector system [186] in which the expression of the selection module is either regulated by *lacI* (pRD007) or by the *tetR-lacI* regulatory circuit (pCascade5) (see Fig. 4.5). The selection module consists of the co-ex-pressed genes *lacZα364*, *cmR*, and *sacB*, under control of the Ptrc promoter from pTrc99A [187] (which is amplified until base pair -300 before start).

The lactose repressor gene *lacI* is PCR amplified from pTrc99A [187]. Tet repressor gene *tetR* and the constitutive promoter PN25 were amplified from the chromosome of DH5*α*Z1 [186]. A functionally random spacer (originating from *D. melanogaster* kinesin coding sequence) of 277 base pairs was inserted between the diverging promoters, to minimize potential transcriptional interference.

Reporter gene *lacZα364* (see also appendix B) consists of the first 364 base pairs of *lacZ*, amplified from the chromosome of strain MG1655 [188] (CGSC stock center).

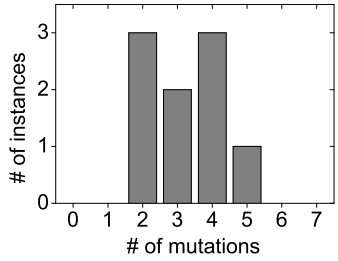


Figure 4.6: Mutations per 1093 base pairs for a random sample of mutagenized isolates.

Chloramphenicol resistance gene *cmR* originates from the pZ vector system [186]. The levan sucrase coding sequence *sacB* was amplified from plasmid pKNG101, obtained from the BCCM/LMBP Plasmid and DNA Library Collection (Belgium), accession number LMBP 5246.

Two reporter plasmids (pRepLacZω and pReplacZ) were created for measuring expression either in cis or in trans, respectively by deleting pTrc99A for *lacI* and Ptrc and inserting a constitutive PlacI^q-*lacZ*ω fragment, or by deleting pTrc99A for *lacI* and Ptrc and inserting the MG1655 Plac-*lacZ* fragment.

Media. All growth and expression measurements, as well as the selection and relative performance experiments were performed in Defined Rich medium (Teknova, Hollister, CA, USA, cat. nr. M2105), with 0.2% glucose as carbon source, and supplemented with 1 mM thiamine HCl.

Mutagenesis. Mutants were created in a mutagenic polymerase chain reaction using the Stratagene Genemorph II Random Mutagenesis kit. Mutation rates can be controlled by varying the amount of template DNA in the reaction. Mutagenized product was restricted and ligated into the (non-mutated) selection vector. Transformation into MC1061 was carried out by electroporation. A control of pool size was performed at every transformation. Pool sizes were routinely between $5{\cdot}10^5$ and $1{\cdot}10^7$.

In order to determine the mutation rate, a random sample of mutants was sequenced after one mutagenesis round. The number of mutations per sequence (length 1093 bps) are given in the histogram in Fig. 4.6. Under the used conditions for mutagenesis on average around 3.2 mutations per 1093 base pairs are applied, which implies a mutation rate of roughly 0.003/bp.

Determination of β-galactosidase activity. To determine the β-galactosidase activity (and thus the expression level) of mutant pools and clones in our experiments, we used the fluorogenic substrate fluorescein di-β-D-galactopyranoside (FDG), which al-

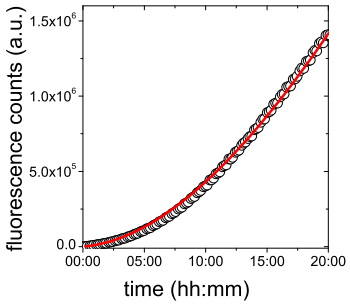


Figure 4.7: Typical fluorescence trace, fitted with equation (4.3) (fit in red)

lows for an accurate determination of the LacZ activity over at least 4.5 orders of magnitude. FDG contains two galactose groups that both have to be cleaved in order to release the fluorescein.

$$\text{FDG} + E \rightarrow \text{FMG} + E + \text{galactose}$$
$$\text{FMG} + E \rightarrow \text{fluorescein} + E + \text{galactose}$$

In [189] an extended model for the FDG-FMG hydrolysis is proposed. In our concentration range of LacZ and FDG, the increase in fluorescence is given by (eq. 7 in [189]):

$$\frac{d}{dt}F = k_2 E \frac{S_0}{K_m + S_0}(\alpha_P + (\alpha_M - \alpha_P)e^{-Rt}) \tag{4.1}$$

where R is the relaxation constant (time scale to reach maximum fluorescence rate), E is the (total) concentration of enzyme, k_2 is the catalysis rate constant of FDG to FMG, and the α's are proportionality factors between product and fluorescence, in the paper given as $F = \alpha_P P + \alpha_M M$ (P is product (fluorescein) and M is FMG). K_m is the Michaelis-Menten constant for FDG and S_0 is the initial FDG concentration. We can see that at time t=0 as well as at large t's the rate with which the fluorescence increases is proportional to E, though with different proportionality constants (first α_M, then α_P).

The paper gives measured values for α_M= 5.3 μM $^{-1}$ and α_P= 150 μM $^{-1}$. Although assigning arbitrary units to the fluorescence counts, they are relevant as relative quantities between FMG and fluorescein. At t=0, equation (4.1) reduces to

$$\frac{d}{dt}F = \alpha_M \; k_2 E \frac{S_0}{K_m + S_0} \tag{4.2}$$

In order to determine the enzyme concentration per cell, fitted slopes should be divided by the cell density: we use here $\varepsilon \propto E/\text{OD}_{600}$, where ε is the LacZ concentration per cell.

Integration of equation (4.1) leads to

$$F(t) = c_1(\alpha_R t + \frac{(\alpha_R - 1)}{R}\text{e}^{-Rt}) - \frac{c_1(\alpha_R - 1)}{R} \tag{4.3}$$

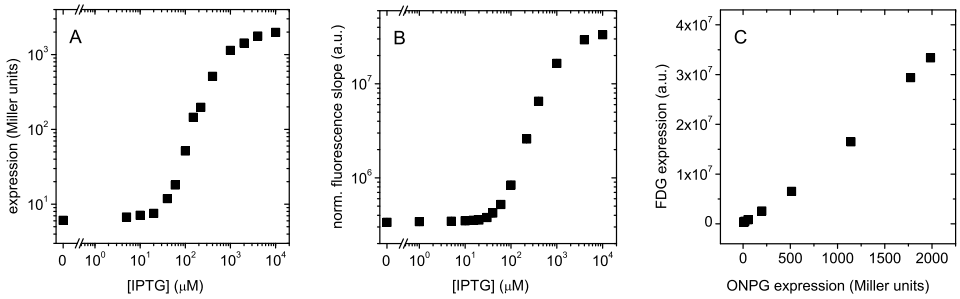


Figure 4.8: Comparison of Miller assay (A) and FDG assay (B) in determining the induction curve of wild type LacI (expressed from pRD007). 'norm. fluorescence slope' indicates the slope of the fluorescence trace at t=0, divided by the optical density at 600 nm. (C) FDG expression versus Miller expression.

where $c_1 = k_2 E \frac{S_0}{K_m + S_0} \alpha_M$, and $\alpha_R = \frac{\alpha_P}{\alpha_M}$. Fig. 4.7 shows a fit to a typical fluorescence trace.

FDG expression measurements were compared to the standard Miller assay for β-galactosidase activity [154], We measured an induction curve of wild-type LacI (as expressed from plasmid pRD007), both by using the Miller assay (Fig. 4.8A) and the FDG assay described above (Fig. 4.8B).

Since β-galactosidase also has affinity for IPTG, inductive IPTG in the medium will competitively inhibit the hydrolysis of FDG by LacZ. Therefore, FDG assays that have to be compared, have to be performed at the same final level of IPTG: we added inhibitive IPTG directly before fixation of the cells and measuring expression.

In order to quantify the IPTG inhibition of FDG hydrolysis, a concentration range from 0 to 10 mM of inhibitive IPTG was added to MC1061 cells harboring pRD007 and pRepLacZ that had been growing without IPTG (Fig. 4.9). The resulting data points are fitted according to the following model. Around t=0, the development of the fluorescence is entirely due to FDG to FMG hydrolysis:

$$\text{FDG} + E \rightarrow \text{FMG} + E + \text{galactose}$$

This yields a simple time derivative of the fluorescence proportional to the concentration of enzyme-FDG complex ($E * \text{FDG}$)

$$\frac{d}{dt} F = \alpha_M k_2 [E * \text{FDG}] \tag{4.4}$$

In the absence of inhibitive IPTG, we recover equation (4.2), using the equilibrium relations for binding of FDG to LacZ. When IPTG is present, the effective concentration of the enzyme-FDG complex is decreased due to titration of enzyme with the competitive binder IPTG.

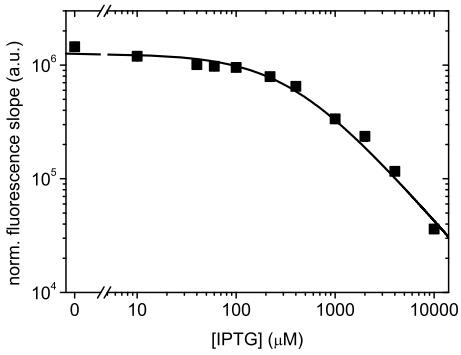


Figure 4.9: The effect of competitive inhibition by IPTG on FDG hydrolysis. 'norm. fluorescence slope' indicates the slope of the fluorescence trace at t=0, divided by the optical density at 600 nm. MC1061 cells containing plasmids pRD007 and pReplacZ were grown without IPTG; the indicated concentrations of IPTG were added just prior to fixation.

The $E*$FDG and the $E*$IPTG concentrations are coupled via the conservation of enzyme

$$[E]_{\text{tot}} = [E]_{\text{free}} + [E*\text{FDG}] + [E*\text{IPTG}] \tag{4.5}$$

where E_{free} is the free concentration of enzyme is solution. We can solve the relevant equilibrium equations to yield:

$$\frac{d}{dt}F \propto [E*\text{FDG}] = \frac{[E]_{\text{tot}}}{1+\frac{K_m}{[\text{FDG}]}(1+\frac{[\text{IPTG}]}{K_{\text{dIPTG}}})} \tag{4.6}$$

where K_{dIPTG} is the equilibrium dissociation constant for IPTG. This equation, for large enough IPTG concentrations ([IPTG] $\gg K_{\text{dIPTG}}$) reduces to

$$\frac{d}{dt}F \propto \frac{[E]_{\text{tot}}}{1+\frac{K_m}{[\text{FDG}]}\frac{[\text{IPTG}]}{K_{\text{dIPTG}}}} \tag{4.7}$$

The data points in Fig. 4.9 are fitted with equation (4.6), using [FDG] = 109 μM (see assay conditions below), and the literature value for the Michaelis-Menten constant for FDG K_m = 18 μM, as given in [189]. This yields a value of 49.6 μM for K_{dIPTG}.

Assay conditions for the determination of β-galactosidase activity were as follows. A reporter plasmid expressing lacZ or *lacZω* was cotransformed into the mutant pool or clone that is to be assayed. Cell cultures were grown at 37°C in a Perkin & Elmer Victor^3 plate reader, at 200 μl per well in a black-clear bottom 96 well microtiter plate (NUNC 165305). Medium was EZ Rich Defined medium + glucose (Teknova, Hollister, CA, USA, cat. nr. M2105), supplemented with 1 mM thiamine HCl and the appropriate antibiotics. Optical density at 600 nm was recorded every 4 minutes, and every 29 minutes 9

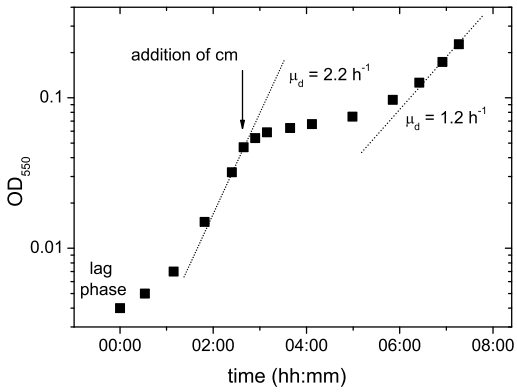


Figure 4.10: Example of a growth curve (optical density at 550 nm) of chloramphenicol selection on a pool (dilution at OD_{550}~0.1 is not shown here).

µl sterile water was injected to each well to counteract evaporation. When not measuring, the plate reader was shaking the plate at double orbit with a diameter of 2 mm. Cells were fixed and stained after the cultures had reached an optical density of at least 0.015 and at most 0.07 (in the plate reader, which corresponds to an OD_{600} of 0.05 to 0.23), by adding 20 µl FDG-fixation solution (109 µM FDG (MarkerGene Technologies Inc, Eugene, OR, USA, cat. nr M0250), 0.15% formaldehyde, and 0.04% DMSO in ddH_2O). Fluorescence development was measured every 8 minutes, as well as the optical density at 600 nm. Shaking and dispensing conditions as above. Note that, as described above, when cells are induced with IPTG, directly before or after fixing and staining, an appropriate amount of inhibitive IPTG was added.

Growth conditions during the selection and the measurement of the fitness landscape. Growth was performed at 37^{o}C in 100 ml erlenmeyer flasks, under vigorous shaking. Culture medium was 20 or 40 ml EZ Rich Defined medium + glucose (Teknova, Hollister, CA, USA, cat. nr. M2105), supplemented with 1 mM thiamine HCl, the appropriate antibiotic, and IPTG when needed. Selective compounds (chloramphenicol, sucrose) were added after 3 hours of pre-selection, after which the cultures were grown for 6 hours. This duration of selective growth was chosen to obtain significant enrichment factors (of up to 10^4), while still maintaining diversity in the population (which starts off at about 10^6). Optical density was monitored at 550 nm (see the example below) and whenever an OD_{550} of 0.1 is reached, a dilution was made into fresh prewarmed selective medium. After selection, cultures were washed, and flash frozen. When transferred to the next environment (without mutagenesis), a threshold dilution was applied, which sets the minimum growth rate for mutants to effectively increase in

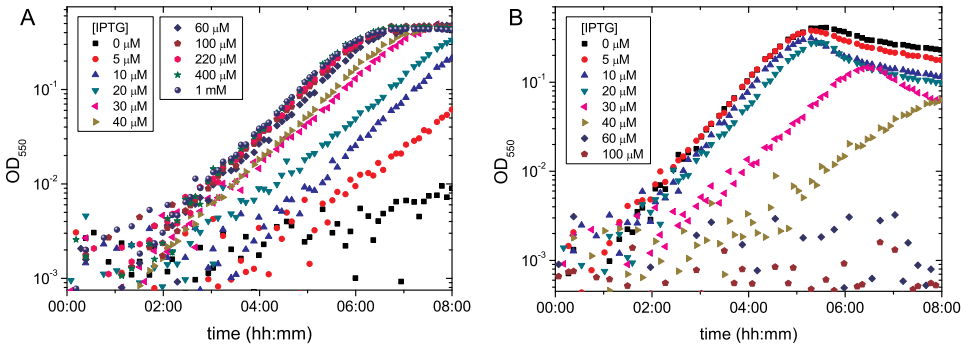


Figure 4.11: Examples of measured growth curves in selective media. (A) Medium containing 25 μg/ml chloramphenicol. (B) Medium containing 0.15% sucrose. Cultures with more than 60 μM IPTG did not increase their OD.

number in the previous environment. This minimum growth rate is typically 1.0 doublings h^{-1}, which implies a dilution of around 500x after each environment (2^6 for 6 hours of selective growth plus a factor 2^2 for the pre-selection period).

In order to measure growth rates for determination of the fitness landscape, wells of a 96-well plate containing 200 μl of Defined Rich glucose medium with the appropriate amount of IPTG were inoculated with a $2{\cdot}10^4$ x dilution of an O/N (LB) culture, and grown for three hours (pre-selection) until an OD_{550} of around 0.0005 (in the plate reader). Since this OD is too low to be determined directly, in the same plate 6 wells were inoculated with a mere $5{\cdot}10^2$ x dilution, which reached a measurable OD of around 0.02 at the same time. At that moment sucrose or chloramphenicol was added[4].

Optical density at 550 nm was recorded every 4 minutes, and every 29 minutes 9 μl sterile water was added to each well to counteract evaporation. When not measuring, the plate reader was shaking the plate at double orbit with a diameter of 2 mm. From the measured growth curves the growth rate was obtained by determining for each well what the increase in cell density was at t=6 hours. From this the effective growth rate was obtained according to $\mu = \frac{\log(OD_{t=6h}/OD_{t=0})}{\log 2}/6$, in doublings per hour. In order to accurately determine the OD at t=0, the plate also contained wells with cells without selective compounds. In case the growth rate was high and stationary phase was reached within 6 hours, the slope of the growth curve was taken directly, since in the selective experiments the cultures were always diluted before reaching stationary phase.

For the 1-to-1 relative performance assay two mutants were mixed in a known ratio and subjected to selective environments. After 6 hours of growth in each environ-

[4]It was checked that the found growth rates do not vary much when the incubation parameters vary within reasonable bounds (a lower dilution upon inoculation (up to $5{\cdot}10^3$), or a shorter pre-selection incubation (2 hours), did not change the results).

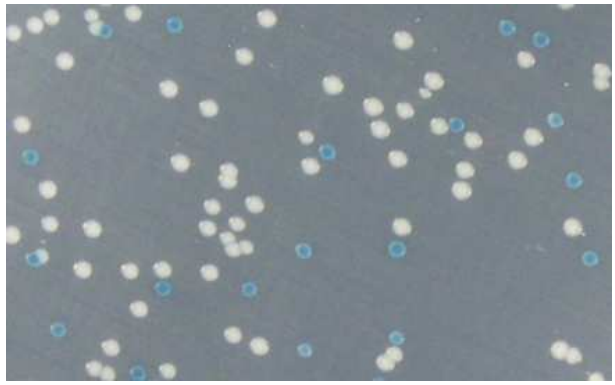


Figure 4.12: Part of a typical counting plate for the 1-to-1 relative performance assay. Cells with high LacZ activity form blue colonies on plates containing Xgal. Image is twice real size.

ment (in which the initial inoculation was such that an OD_{550} of just under 0.1 was reached), cultures were washed, and allowed to grow to stationary phase in LB medium. A DNA extraction was performed on the whole pool for each culture, of which subsequently around 0.1 ng was electroporated into BioRad EP-Max10B electro-competent cells (cat.no. 170-3330), and directly plated on agar containing Xgal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) and/or IPTG. As our selection module contains a *lacZα* gene, complementation with the chromosomally expressed *lacZω* allowed for discrimination between the mutants and determination of their ratio (Fig. 4.12).

4.2 Additional material

4.2.1 Interpolation of expression-growth curves using growth models

In order to interpolate the measured points on the expression-growth relations, we use models for the selective action of chloramphenicol and sucrose.

chloramphenicol growth

In the presence of a certain concentration of chloramphenicol acetyl transferase (CmR), the internal concentration of chloramphenicol (cm) is reduced and determined by the equilibrium between influx through the cell membrane and acetylation ('inactivation') by CmR. As such, we model the action of cm by comparing the situation with growth under sublethal concentrations of cm. The most basic equation relating growth to the concentration of an inhibitive substance is derived from the Monod form for nutrient limited growth [190], $\mu \sim \frac{KX}{X+K}$, where μ is the growth rate, X is the concentration of nutrient and K is a constant determining the nutrient concentration that allows half-maximum growth rate. Interestingly, this is the same functional form as the fraction

of substrate bound enzyme under Michaelis-Menten kinetics. Now, for sublethal concentrations of cm, whose action is to block protein synthesis upon binding to the ribosomes, it would not be unreasonable to expect the the growth of the cell (as a first-order approximation) to be proportional to the *unbound* fraction of ribosomes, which is given by $\frac{K}{X+K}$. Therefore we adopt the following simple functional form for growth in the presence of chloramphenicol

$$\mu([\mathrm{cm}]_{\mathrm{ext}}) = \frac{\mu_0}{c_1\,[\mathrm{cm}]_{\mathrm{int}} + 1} \tag{4.8}$$

where c_1 is a constant, μ_0 the growth rate in absence of cm, and $[\mathrm{cm}]_{\mathrm{ext}}$ and $[\mathrm{cm}]_{\mathrm{int}}$ respectively the cm concentrations outside and inside the cell.

To obtain a relation between the internal and external cm concentration, we express the equilibrium between influx and acetylation of cm by

$$C_{\mathrm{bar,cm}}([\mathrm{cm}]_{\mathrm{ext}} - [\mathrm{cm}]_{\mathrm{int}}) = r_{\mathrm{acet,cm}} \tag{4.9}$$

Here the influx of cm is either diffusion limited or limited by the permeability of the membrane, which does not matter for the functional form of the equation, and can be expressed as a constant $C_{\mathrm{bar,cm}}$ times the concentration difference between inside and outside. The acetylation rate $r_{\mathrm{acet,cm}}$ is given by

$$r_{\mathrm{acet,cm}} = k_{\mathrm{cat,cm}}[E*\mathrm{cm}] = k_{\mathrm{cat,cm}}\frac{E_{\mathrm{tot}}}{1+\frac{K_{\mathrm{m}E\mathrm{cm}}}{[\mathrm{cm}]_{\mathrm{int}}}} \tag{4.10}$$

where $k_{\mathrm{cat,cm}}$ is the catalysis rate constant for the acetylation reaction, and $K_{\mathrm{m}E\mathrm{cm}}$ is the Michaelis-Menten constant for CmR. Solving for $[\mathrm{cm}]_{\mathrm{int}}$ in

$$C_{\mathrm{bar,cm}}([\mathrm{cm}]_{\mathrm{ext}} - [\mathrm{cm}]_{\mathrm{int}}) = k_{\mathrm{cat,cm}}\frac{E_{\mathrm{tot}}}{1+\frac{K_{\mathrm{m}E\mathrm{cm}}}{[\mathrm{cm}]_{\mathrm{int}}}} \tag{4.11}$$

now yields the expression for the growth rate as a function of the external chloramphenicol concentration (here abbreviated as [cm]), being

$$\mu([\mathrm{cm}]) = \frac{\mu_0}{1+\frac{c_1}{2}([\mathrm{cm}] - K_{\mathrm{m}E\mathrm{cm}} - E_{\mathrm{tot}}k_{\mathrm{cat,cm}} + \sqrt{4[\mathrm{cm}]K_{\mathrm{m}E\mathrm{cm}} + ([\mathrm{cm}] - K_{\mathrm{m}E\mathrm{cm}} - E_{\mathrm{tot}}k_{\mathrm{cat,cm}})^2})} \tag{4.12}$$

Expression-growth data for media containing cm were fitted with this equation.

sucrose growth

Sucrose selection is based on the formation of sugar chains (levan) in the periplasmic domain of gram-negative bacteria [191]. The enzyme catalyzing this polymerization reaction is levansucrase (SacB) from *Bacillus subtilis*. In the gram-positive *B. subtilis*

Figure 4.13: Lysis of *E. coli* MC1061 cells as a result of the build-up of levan chains in the periplasm. The interval between the pictures is roughly 7 minutes. Cells were often observed to suddenly adopt a spherical shape prior to complete lysis. Scale bar is 1 μm.

the enzyme is exported through the inner membrane, where it constitutes a protective poly-sugar layer outside the cell wall. In gram-negative bacteria, which have a second cellular membrane, the enzyme is not exported through the second membrane and therefore accumulates levans in between the cellular membranes, which decreases the cellular growth rate. High expression of the protein in the presence of sucrose is lethal and leads to lysis of the cells (see Fig. 4.13).

Thus, the rate of levan formation is the factor influencing cell growth. In contrast to chloramphenicol, which is a bacteriostatic, high levan production leads to lysis of cells, and in a population average this can give rise to a negative growth rate. Therefore the growth as a function of levan formation rate cannot directly be described by the Monod form. However, expecting that the relevant parameter for the toxic effect is the levan formation rate (r_{levan}) *relative* to the instantaneous growth rate[5], we can write a modified Monod form

$$\mu([\text{sucrose}]) = \frac{\mu_0}{c_0 \frac{r_{\text{levan}}}{\mu([\text{sucrose}])} + 1} \tag{4.13}$$

where and c_0 is a constant. This can be solved to yield

$$\mu([\text{sucrose}]) = \mu_0 - c_0\, r_{\text{levan}} \tag{4.14}$$

In the same way as for the chloramphenicol selection, we can write down the rate of levan formation and the equilibrium governing the transport of sucrose through the outer membrane, yielding

$$r_{\text{levan}} = k_{\text{cat,sucr}}[E * \text{sucrose}] = k_{\text{cat,sucr}} \frac{E_{\text{tot}}}{1 + \frac{K_{\text{m}E\text{sucr}}}{[\text{sucrose}]_{\text{int}}}} \tag{4.15}$$

and

$$C_{\text{bar,sucr}}([\text{sucrose}]_{\text{ext}} - [\text{sucrose}]_{\text{int}}) = k_{\text{cat,sucr}} \frac{E_{\text{tot}}}{1 + \frac{K_{\text{m}E\text{sucr}}}{[\text{sucrose}]_{\text{int}}}} \tag{4.16}$$

[5] A stronger effect of sucrose was indeed observed when the basal growth rate is lowered (e.g. growth with glycerol as a carbon source instead of glucose)

We can solve equation (4.16) for $[\text{sucrose}]_{\text{int}}$, substitute this into equation (4.15), which in its turn can be substituted into equation (4.14) to obtain the growth rate as a function of the external sucrose concentration and the expression of *sacB*.

However, all measured expression-growth characteristics for sucrose show a steeper dependency on enzyme concentration than can be obtained by this form. Indeed, SacB mediated formation of levan is a process that needs a levan seed in order to proceed [192, 193]. Most probably seed formation is also dependent on the enzyme and sucrose concentration. Therefore we phenomenologically alter the equation for the growth rate as a function of levan formation rate into

$$\mu([\text{sucrose}]) = \mu_0 - c_0\, r_{\text{levan}}^n \tag{4.17}$$

The obtained function provides good fits for the low-enzyme regime of the expression-growth data.

However, at the high $[E]$ end, we observe a saturation at higher growth rates than equation (4.17) can account for. There are at least three saturation effects (see also [192, 193]) coming into play at high rates of levan synthesis (apart from potential feedback on protein production in 'struggling' cells):
1) Since the levans are (possibly branching) chains, the autocatalytic seed-effect (see above) of the reaction decreases: attaching a fructosyl-group to an existing long chain does not increase the number of fructosyl-acceptors.
2) At high levan production rates, there is a high concomitant production of glucose, that has an inhibitory effect on levan formation in two ways:
2a) The fructosylation reaction by the $E*S$ (levansucrase-sucrose) complex branches between levan elongation and fructosylation of glucose (which re-forms sucrose).
2b) The competitive inhibition of E to S binding by glucose. Due to levan formation, the internal sucrose concentration decreases, and the glucose concentration increases.
3) The formed levans themselves act as an inhibitor at higher concentrations.
Since it is at this stage impossible and will not yield further insight to adapt the model to account for the saturation at high enzyme concentration, we opt for a more phenomenological description. For fits over the complete concentration range of SacB enzyme, we used

$$\mu([\text{sucrose}]) = \frac{\mu_0 + \mu_{\text{sat}}}{c_0\, r_{\text{levan}}^n + 1} - \mu_{\text{sat}} \tag{4.18}$$

4.2.2 Mutant sequences

Here mutant sequences are given for some of the *lac* repressors after wild-type selection, for the inverse repressor phenotypes, as well as for some of the regulatory circuits.

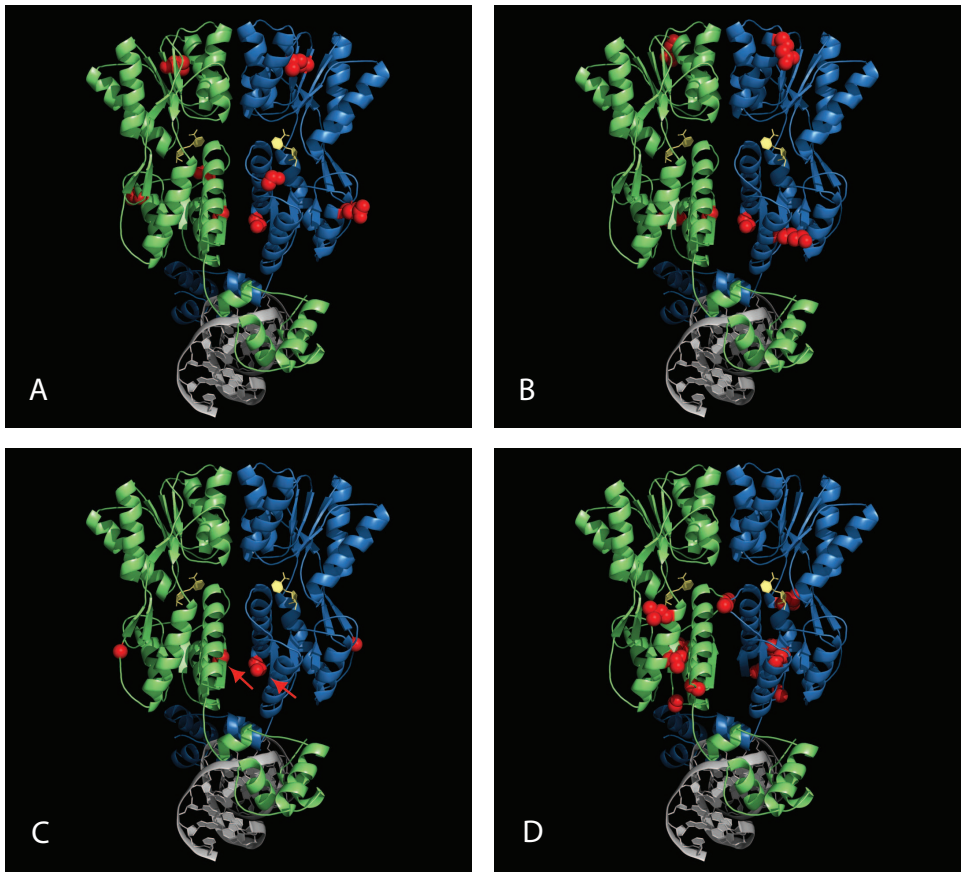


Figure 4.14: Structure of the C-terminal deleted but otherwise wild-type *lac* repressor (PDB 1JWL); residues that are mutated in inverse repressors are rendered as red space-filling residues. DNA sequences are given on page 76. (A) mutant FPM400m21 (B) mutant 3BII5 (C) mutant M32alt; red arrows indicate the recurring Ser97Pro mutation (D) mutant FPM399m22

***lac* mutants after wt selection**

5B-1		5B-4		5B-9	
C828T	silent	C828T	silent	T269A	Leu90Gln
				T663C	silent
				T702A	Asn234Lys

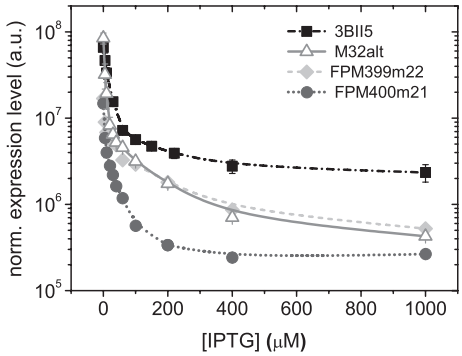


Figure 4.15: Induction profiles of mutants FPM400m21, 3BII5, M32alt, and FPM399m22. The inverse profile of mutant FPM400m21 is also the one shown in Fig. 4.3.

inverse *lac* repressors

Shown below are sequences for isolated inverse repressors from different lineages. Figure 4.14 indicates mutant amino acids as red space-filling residues in the wild type (but dimeric) *lac* repressor in complex with operator DNA and the ligand ONPF (PDB ID code 1JWL). Repressor monomers are shown in blue and green, DNA in grey, and ligand in yellow. Note that the structure in PDB 1JWL is a dimeric C-terminal deletion mutant: it lacks the tetramerization domain, residues 330-360. Images were created using PyMOL [194]. Measured induction profiles are shown in Fig. 4.15.

FPM400m21		3BII5		M32alt		FPM399m22	
C206A	Ser69Tyr	G171T	silent	T289C	Ser97Pro	G123A	silent
T289C	Ser97Pro	T289C	Ser97Pro	G944A	Gly315Asp	G213A	silent
A392C	Gln131Pro	A324T	Lys108Asn	C1016A	Pro339His	C215T	Ala72Val
G723T	silent	G432A	silent			C275T	Ala92Val
G726A	Met242Ile	T573C	silent			G320C	Cys107Ser
G813A	silent	G705T	Glu235Asp			T360A	Ser120Arg
C952T	silent	C963G	silent			C382T	silent
T1037A	Leu246stop	C1054G	Gln352Glu			C883A	Leu295Met
						T902C	Val301Ala
						C1001T	Thr334Met
						T1037A	Leu346stop

Distribution of amino acids substitutions in inverse repressors.
Overall there seems no clear clustering of substitutions in a particular domain of the repressors. However, no isolates were recovered with substitutions in the domain that makes direct contact with the DNA. Furthermore, all inverse isolates contain substitutions in the interface between the monomers, among which particularly often Ser97Pro, which is marked with red arrows in Fig. 4.14C.

regulatory circuit sequences

OR phenotype

tetR mutations

C132T	silent
C145G	Arg49Gly
T269 del	
A270 del	deletion Leu (90)
C271 del	
T355A	Phe119Ile
G549T	Glu183Asp

promoter mutations

A–36G	In PN25 promoter (36 bps before transcriptional START)

lacI mutations

G285A	silent
G325T	Ala109Ser
C900T	silent

NAND phenotype

tetR mutations

G128C	Trp43Ser
C132T	silent
C149G	Ala50Gly
G214T	Gly72Trp
C568A	Leu190Ile

promoter mutations

T-136C	In insert between PN25 and PLtetO1 (136 bps before PLtetO1 transcriptional START)

lacI mutations

T289C	Ser97Pro
T401C	Ile134Thr
A466G	Ile156Val
G531A	silent
C1001T	Thr334Met
G1064 del	Arg355His
	Leu356Trp
	Glu357Lys
	Ser358Ala
	Gly359Asp
	Gln360Asn
	stop361Asn
	362Thr
	363Trp
	stop

4.2.3 Alternative selective pressures and fitness landscapes

Here we consider alternative selective pressures than used in the main text.

comparison between stringent and less stringent selection & relative growth of regulatory mutants

We can construct the fitness landscape for alternating environments, comprised of any combination of media described by the curves in Fig. 4.1C and D. Especially for the relative growth assay of the mutants depicted in Fig. 4.2B (mutants R1 and N1), the fitness landscapes clearly illustrate the different outcome in stringent and less stringent environments.

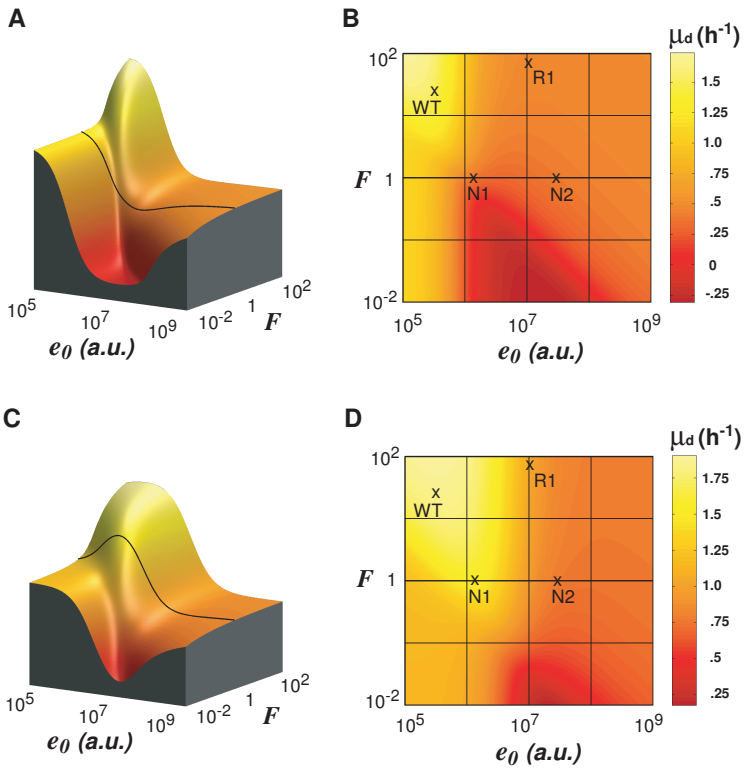


Figure 4.16: (A) and (B): Fitness landscape for environments alternating between 80 μg/ml chloramphenicol + 1 mM IPTG and 0.4% sucrose. In (B) the mutants are depicted of which we assayed their 1-to-1 performance. In these environments responsive mutant $R1$ outperforms non-responsive mutant $N1$. (C) and (D): Fitness landscape for environments alternating between 25 μg/ml chloramphenicol + 1 mM IPTG and 0.15% sucrose. Here the landscape predicts that mutant $N1$ outperforms mutant $R1$, which was indeed found in relative growth assay. Growth rates μ are in doublings h^{-1}. Black lines in (A) and (C) indicate non-responsive phenotypes: F=1.

non-equal dwelling times in alternating environments

In principle the expression-growth relations from Fig. 4.1C and D can also be used to predict the fitness landscapes for regulation when the dwelling time in the alternating environments is unequal. In the figure below we show the resulting fitness landscapes when the dwelling time in the cm environment is four times longer than the dwelling time in the sucrose environment. We can see that the resulting selective pressure to conserve wild-type regulation has effectively decreased, since large regions in the e_0-F plane have emerged that do not have wild-type regulation, but do have a near-optimal fitness.

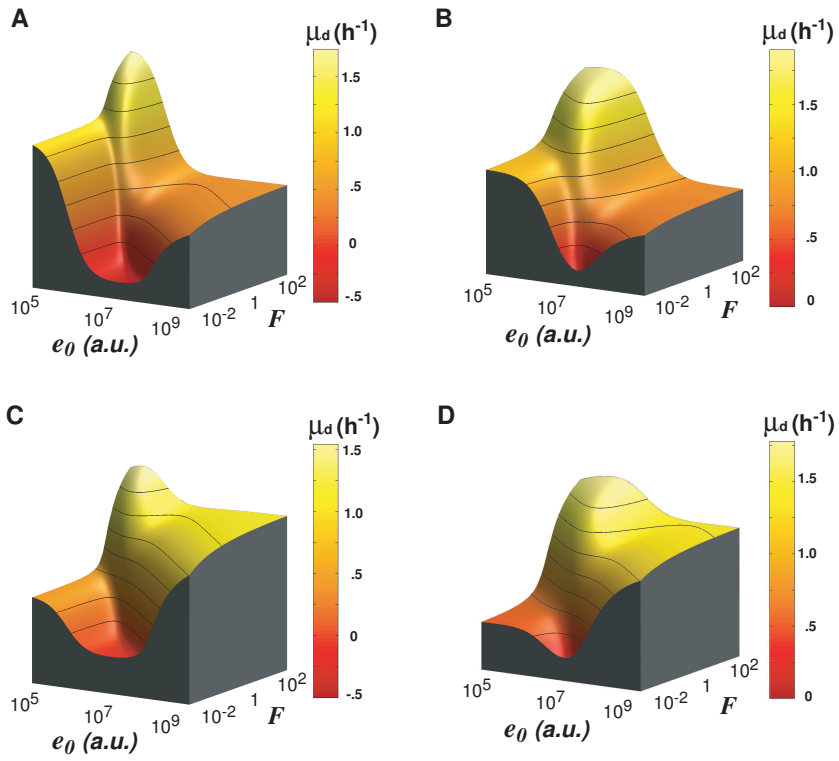


Figure 4.17: (A) and (C): Fitness landscape for environments alternating between 80 μg/ml chloramphenicol + 1 mM IPTG and 0.4% sucrose. In (A) dwelling times in both environments are the same, while in (C) dwelling time ratio between cm + IPTG and sucrose conditions is 4:1. (B) and (D): Fitness landscape for environments alternating between 25 μg/ml chloramphenicol + 1 mM IPTG and 0.15% sucrose. In (B) dwelling times in both environments are the same, while in (D) time ratio between cm + IPTG and sucrose conditions is 4:1. Growth rates μ are in doublings h^{-1}.

4.2.4 Simple simulation of mutant pools and direction of selection

simulation of mutant pools in the $e_0 - F$ plane

We performed a simple simulation of differential enrichment in a mutant population. A pool of mutants with a certain distribution $N_0(e_0, F)$ is passed through two subsequent selective environments, yielding growth rates of $G_1(E,S)$ and $G_2(E,S)$ respectively. In case the selection times in both environments are equal ($t_1 = t_2 = t$), we may write for the new distribution

$$N(e_0, F, t) = N_0(e_0, F)\mathrm{e}^{\ln 2\,(G_1(E,S)+G_2(E,S))\,t} \tag{4.19}$$

Fitness landscapes as depicted in Fig. 4.2AB and 4.3AB are now represented by the term $G_1(E,S)+G_2(E,S)$. Assuming an initial Gaussian e_0,F-distribution with the same center of mass as the experimentally measured mutated pool (green spheres in Fig. 4.2 and 4.3), end distributions are calculated on the basis of a discrete pool of mutants. Figure 4.18 shows for wild-type selection the assumed Gaussian mutant pool before selection, the calculated pool after selection in the sucrose medium, and the calculated pool after selection in the cm medium. After each environment a 500x dilution is applied. The location of the end-distribution (Fig. 4.18, right) corresponds well to the experimentally observed distribution (Fig. 4.2AB).

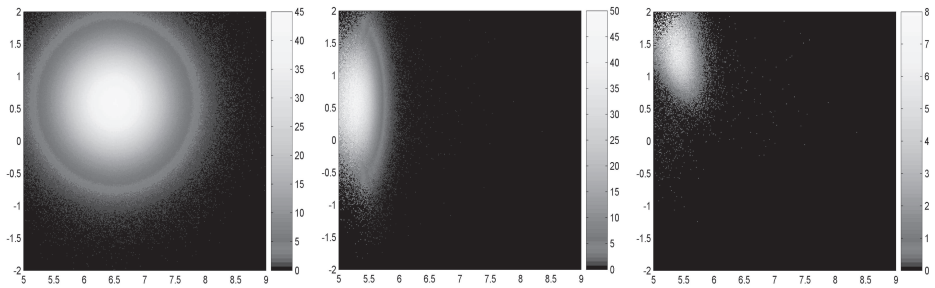


Figure 4.18: Initial distribution before selection (left), calculated pool after sucrose selection (middle), and calculated pool after sucrose and cm selection (right). Initial mutant distribution is here assumed Gaussian, with a total of 10^6 mutants. Axes are the same as in Fig. 4.2A. Grey values indicate the number of individuals per 'unit surface' in the fitness landscape.

The same is done for one round of selection for inverse repressor phenotype (Fig. 4.19). While the initial mutant distribution is of course an assumption that heavily influences the distribution after selection, the calculation does show how this distribution would change, were there no genetic constraint. Upon comparison with the measured pool after the first selection round (blue spheres in Fig. 4.3), we can infer that the clustering of the measured pool around the line $F = 1$ is not dictated by the phenotype-fitness mapping, but due to the underlying genetic architecture, which determines the mapping from genotype to phenotype.

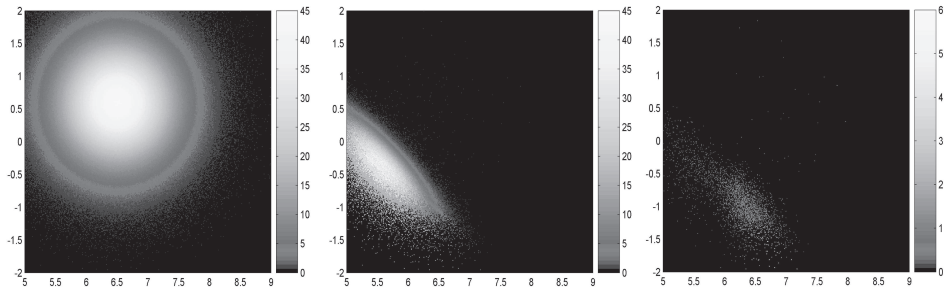


Figure 4.19: Initial distribution before selection (left), calculated pool after sucrose selection (middle), and calculated pool after sucrose and cm selection (right). Initial mutant distribution is here assumed Gaussian, with a total of 10^6 mutants. Axes are the same as in Fig. 4.3A. Grey values indicate the number of individuals per 'unit surface' in the fitness landscape.

calculation of the direction of selection

Using a simple calculation we can demonstrate how the net direction of the selective pressure in an alternating environment is towards the right upper corner in a trade-off diagram, like the one shown in Fig. 4.1E. Depending on how strong the trade-off is, this is generally in the direction of responsive phenotypes. Indeed this is the case for the trade-off curves shown in Fig. 4.1E. Note that what will follow assumes a regime of environmental fluctuations that is not so rapid that the cells effectively experience an average environment (see [42]), but also not so slow that clonality is obtained during the dwelling time in one environment. We start from a general distribution $P_o(\mu_1,\mu_2)$ of individuals with growth rates μ_1 and μ_2 in environment 1 and 2 respectively (see Fig. 4.20).

For every individual, after remaining a time t_1 in environment 1 and a time t_2 in

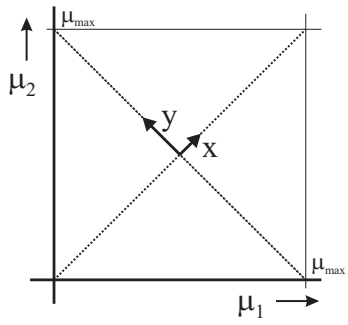


Figure 4.20: Trade-off diagram of growth in two environments 1 and 2, where mutants exhibit growth rate μ_1 and μ_2 respectively. Indicated are the axes x and y of the alternative coordinate system.

environment 2, its number will have increased by a factor of $2^{\mu_1 t_1} \cdot 2^{\mu_2 t_2}$. Therefore the distribution will have changed towards

$$P_s(\mu_1, \mu_2) = \frac{2^{\mu_1 t_1} \cdot 2^{\mu_2 t_2} \, P_o(\mu_1, \mu_2)}{N_s} \tag{4.20}$$

where N_s is the normalization factor accounting for the total growth of the population. We now change the coordinate system of the growth rates to the perpendicular axes x and y (see Fig. 4.20), locating the origin in the center, using $\mu_1 = \frac{\mu_{max}+x-y}{2}$ and $\mu_2 = \frac{\mu_{max}+x+y}{2}$. We can rewrite the mutant distribution after selection as

$$P_s(x, y) = \frac{2^{\frac{(\mu_{max}+x-y)t_1}{2}} \cdot 2^{\frac{(\mu_{max}+x+y)t_2}{2}} P_o(x, y)}{N_s} \tag{4.21}$$

which in case of equal $t_1 = t_2 = t$ gives

$$P_s(x, y) = \frac{2^{(\mu_{max}+x)t} P_o(x, y)}{N_s} = c \, 2^{xt} P_o(x, y) \tag{4.22}$$

from which we can see that by selection in these alternating environments the distribution shifts in the direction of x, towards the right upper corner of Fig. 4.20.

CHAPTER

5 Identification of functional mutations and epistasis by reverse neutral evolution

I returned and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all.

Ecclesiastes 9:11

In recent years a major focus in evolutionary research has been on the molecular basis of adaptation. An exciting new development is the reconstruction of evolutionary intermediates between an ancestral and evolved sequence and investigate, assuming a relevant selective pressure, which evolutionary pathways are accessible. However, when the total number of mutational differences between the ancestral and evolved sequence rises, the combinatorial complexity soon surpasses our capacity to reconstruct all intermediates. In that case pathway reconstruction and statistical analysis will be incomplete. Often though, a considerable fraction of the mutations between ancestor and evolved sequence will be functionally neutral, and our understanding of the evolving system may not depend on them. In this work we apply a PCR-based technique to remove neutral mutations from an evolved inverse lac repressor and screen the obtained pool of intermediate sequences for conservation of function. We analyze correlations between mutations in the selected pool, after correcting for correlations due to the PCR procedure. Based on this data we attempt to deduce functional information and decide which subsets of mutations are interesting for further investigation. In ongoing work selected loci are further analyzed by creation and measurement of all mutational intermediates and inspection of genotype-phenotype and genotype-fitness landscapes.

For a long time the main source of detailed information about evolutionary processes has been standing genetic variation. Current polymorphisms and phylogenetic analysis have been used to uncover the signature of natural selection, and in some cases to infer evolutionary pathways from ancestor to present-day DNA sequences and proteins. However, even if phylogenetic data on a certain system is sufficiently complete, this type of analysis does not provide information at the level of phenotype or fitness, which leaves an important area of fundamental questions inaccessible. For example, without recourse to data on fitness, many issues surrounding the repeatability or the predictability of evolution, or the prominence of adaptive constraint cannot be resolved. Based on earlier ideas, and helped by the development of molecular biological techniques, recent studies have been very successful in filling this gap [20, 55]. By reconstructing ancestral sequences and possible intermediates towards the present-day sequence, and often analyzing them in the context of a fitness landscape, one can obtain information about key events in evolutionary history at the molecular level.

However, the amount of mutational differences between ancestor and evolved sequences easily becomes so high that the number of possible intermediates increases beyond regular molecular screening techniques. Two sequences that are polymorphic at L locations, have $2^L - 2$ possible intermediates. To recreate these intermediates can by itself already be a challenge, even apart from assaying relevant phenotypes and fitnesses. Often, a subset of the L mutant loci (note that in this chapter 'loci' will usually refer to positions *within* a gene) will have no or negligible effect on a present phenotype or fitness. In these cases, one may consider intermediates consisting of a subset of the L loci and still capture the essential information about the evolutionary process. On the basis of sequence information alone, however, it is impossible to know which of the loci are neutral and can be disregarded in the analysis of the evolutionary trajectories.

In this work we apply a PCR technique that is followed by selection for conservation of function to sieve out non-functional mutations and a statistical analysis for functional relevance. The PCR reaction accomplishes *in vitro* recombination of mutant loci on a DNA sequence conceptually similar to the DNA shuffling method by Stemmer [195], but does not require reassembly of fragmented sequences. A comparable PCR approach has been described [196] but without a statistical functionality analysis. Due to the nature of the PCR reaction, or recombination more in general, distance correlations will arise between the occurrence of mutations. Our analysis aims to separate these correlations from those that can be an indication for functionality or epistatic interactions. For this analysis it is not necessary that all neutral mutations have been removed in the selection process: we compare observed correlations between loci with their expected average correlation based on their distance on the DNA.

We follow this approach, using an (artificially) evolved inverse *lac* repressor (chapter 4). This inverse repressor has an opposite response to its ligand isopropyl-β-D-thiogalactopyranoside (IPTG) compared to the wild-type *lac* repressor: it represses in the presence of IPTG and abolishes repression in its absence. We found earlier that

there is a large diversity of genetic changes in the wild-type repressor that can accomplish this inverted functionality. Here we focus on one such mutant, containing 8 base pair substitutions compared to wild-type LacI. We discuss another statistical test (ANalysis Of VAriances, ANOVA) with respect to our data which in principle is powerful to infer functionality and epistatic interactions, but has limitations when the data is not complete. For subsets of polymorphic loci that seem functionally important we are currently in the process of constructing and measuring phenotypes of all possible intermediates. By tracing mutational paths in a fitness landscape, we expect to find ample evidence for epistatic interactions governing the inversion of function in the *lac* repressor. Some implications and limitations of the followed approach will be discussed.

5.1 Methods

5.1.1 PCR procedure and selection

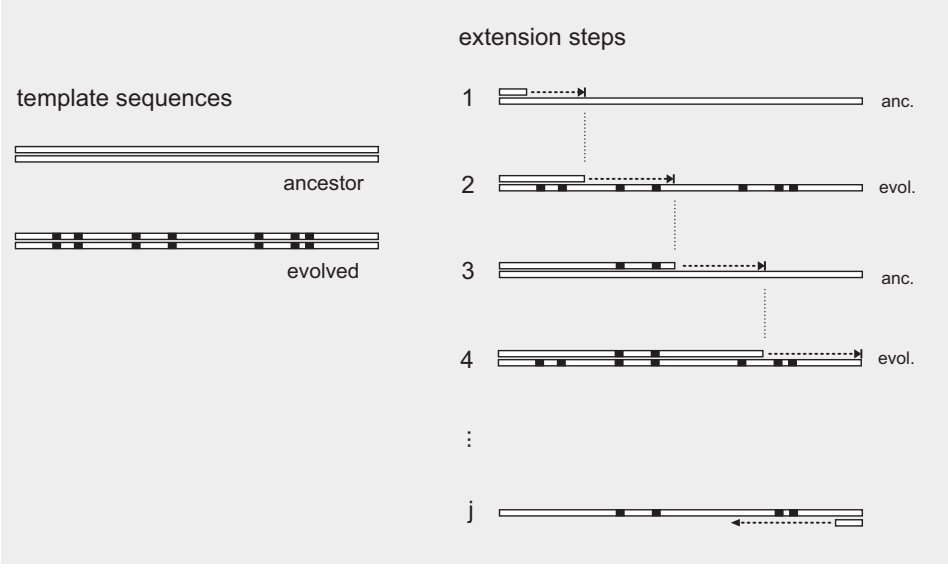


Figure 5.1: Template switching during the dilution PCR procedure. Two DNA templates differing in a limited number of base pairs are mixed. 1) After an annealing step, a primer is extended, replicating an non-mutated (ancestral) template strand. Because of short extension time steps, elongation does not proceed to the end of the template. The elongated primer is released in the subsequent melting step (not shown). 2) The elongated primer anneals to a mutant (evolved) strand, and by further extension incorporates a subset of the mutations. 3) and 4) the procedure is repeated until a full-length complementary strand is formed. j) When an extension is completed, the resulting DNA strand containing a subset of the mutant loci can serve as a template for a reverse primer.

The procedure of the 'dilution PCR' is only slightly modified from a standard PCR

reaction. There are two main differences. First, the template is a mix of two DNA sequences ('ancestor' and 'evolved') that are similar, except for a number of polymorphic loci. Second, the elongation steps in the PCR procedure are drastically shortened. A standard PCR protocol typically consists of 30 cycles of melting (order 30s at 94^{o}C), primer annealing (order 30s just below melting temperature for primers), and elongation (60s per 1000 base pairs at 72^{o}C). A dilution PCR reaction shortens the elongation step (in the present work to 20s for 1200 base pairs), and increases the number of cycles (here 99 instead of 30). In this way incomplete elongation is accomplished in the extension steps, and the nascent DNA strands are able to switch template in subsequent steps (see fig. 5.1). Care should be taken that the primer annealing temperature is sufficiently low, otherwise the effective elongation step is much longer than the time spent at 72^{o}C. Parameters that influence the switching between templates are the elongation time and the concentration ratio of the two templates (which here was chosen to be one). How these parameters should be tuned, depends on how strong the dilution of mutations should be, which in turn depends on the purpose of the experiment and the pool sizes that can be screened. Generally, to remove unwanted linkage between loci, the switching between templates should be as often as possible. However, control over this parameter is limited, since the temperature steps are never real step functions, and the polymerase will also work at non-optimal temperature ranges. Moreover, at very short elongation steps, the product yield will be unpredictable.

After creation of the recombined DNA sequences, we can inspect their phenotype. The creation of the pool consisting of mutation diluted sequences is followed by selection for conservation of function. The selection procedure is described in detail in chapter 4. The procedure is set up such that the evolved inverse LacI phenotype has a high enough fitness to not overly favor potential fitness improvements with respect to this phenotype.

5.1.2 Identification of correlated loci

In order to determine whether there is a meaningful correlation between polymorphic loci in sequences after selection, we have to remove correlations that are due to the PCR procedure. The nature of the PCR procedure will make that closely neighboring loci have a higher chance to be correlated, since they might originate from a single elongation event. Of course, in later stages of the PCR reaction, mixing has already taken place, and the same elongation event may then include polymorphic loci from different origin (ancestor or descendant). These latter events will reduce the average *effective* run length. To know the characteristic length scale at which correlations decay due to the dilution PCR, we need to know the average effective run length.

determination of run length

To estimate this average effective run length, we cannot simply inspect the resulting amplified sequences. Since most base pairs of the two template sequences will be identical, information about the run length can only be obtained from inspection of the polymorphic loci, and the correlation between consecutive polymorphic loci. For example, if two consecutive polymorphic loci are uncorrelated among amplified sequences, the probability is high that the average run length is smaller than the distance between these loci. We now derive an expression for the probability that two polymorphic loci share the same origin (are both ancestor, or both descendant) as a function of average run length. To this end, we consider a Poisson process that alternates between two states, 0 and 1, for simplicity with equal characteristic switching rates Δ for both states. Starting in one of the states, here 0, we calculate the probability $P_0(\kappa|0)$ that after a certain propagation κ (which here is the length over the DNA sequence, and is considered a continuous variable) we observe the same state 0 again.
This probability is given by the sum

$$P_0(\kappa|0) \;=\; P_{0\to 0}(\kappa|0) \;+\; P_{0\to 1,1\to 0}(\kappa|0) \;+\; P_{0\to 1,1\to 0,0\to 1,1\to 0}(\kappa|0) \;+\; \cdots \tag{5.1}$$

where $P_{0\to 0}(\kappa|0) = S(\kappa) = (1/\Delta)\mathrm{e}^{-\kappa/\Delta}$ is the survival probability density of the individual state. The other terms can be calculated by considering the example

$$\begin{aligned} P_{0\to 1,1\to 0}(\kappa|0) \;&=\; \int_0^\kappa d\kappa_2 \int_0^{\kappa_2} d\kappa_1 \; S(\kappa_1)S(\kappa_2-\kappa_1)S(\kappa-\kappa_2) \\ &=\; \frac{1}{\Delta^2}\,S(\kappa)\int_0^\kappa d\kappa_2 \int_0^{\kappa_2} d\kappa_1 = \frac{1}{2}\Big(\frac{\kappa}{\Delta}\Big)^2 S(\kappa) \end{aligned} \tag{5.2}$$

When all terms are taken into account, the result is

$$P_0(\kappa|0) = \sum_{n=0}^{\infty} \frac{\big(\frac{\kappa}{\Delta}\big)^{2n}}{2n!}\mathrm{e}^{-\kappa/\Delta} = \cosh\Big(\frac{\kappa}{\Delta}\Big)\mathrm{e}^{-\kappa/\Delta} \tag{5.3}$$

In this calculation it was assumed that upon switching, the state changes from 0 to 1 or from 1 to 0. In the case of template switching, however, this is not necessarily so: the subsequent annealing may be again on a 'same state' template. In other words, it may flip from state 0 to state 0, which results in a rescaling of the decay rates. We do not explicitly perform this scaling, however. Apart from the fact that it will yield the same results in what follows, we are here interested in extracting the average *effective* run length (see also 'caveats' below).

Using equation (5.3), we can now perform a regression procedure to obtain the average run length Δ. Assuming we have N back-crossed sequences, each with L polymorphic loci, we can compare the expected and the real correlation for all consecutive loci. In order to find an estimate for the average run length, we have to minimize the

following sum[1]

$$\chi^2 = \sum_{n=1}^{N} \sum_{l=1}^{L-1} (E^n_{l,l+1})^2 \tag{5.4}$$

where $E^n_{l,l+1} = 1 - P_0(\kappa|0)$ if the consecutive loci l and $l+1$ have the same origin (both 0 or both 1), or $E^n_{l,l+1} = P_0(\kappa|0)$ if they have a different origin (0 and 1, or 1 and 0). The value for κ in the calculation of the $P_0(\kappa|0)$ terms is given by the distance between loci l and $l+1$ in base pairs. The average run length Δ is that which minimizes equation 5.4.

p-values for correlations

Having obtained the effective run length Δ, we can look at pairs polymorphic loci and decide whether they are more correlated or anti-correlated than can be expected on the basis of their distance. The null hypothesis –no functional correlations between pairs of loci– would result in a binomial distribution of equal states of the loci, with an intrinsic probability $P_0(\kappa|0)$, using the found run length Δ. For a certain pair of loci, given that ν out of a total N (number of sequences) share an equal origin (both states are 0, or both are 1), we can perform a two-tailed test to assess whether the null hypothesis should be rejected. The p-value is given by

$$\begin{aligned}
p &= \sum_{n=\nu}^{N} \binom{N}{n} P^{N-n}(1-P)^n && \text{if } \nu > 2PN \\
p &= \sum_{n=\nu}^{N} \binom{N}{n} P^{N-n}(1-P)^n + \sum_{n=0}^{2PN-\nu} \binom{N}{n} P^{N-n}(1-P)^n && \text{if } 2PN \geqslant \nu > PN \\
p &= 1 && \text{if } \nu = PN \\
p &= \sum_{n=0}^{\nu} \binom{N}{n} P^{N-n}(1-P)^n + \sum_{n=2PN-\nu}^{N} \binom{N}{n} P^{N-n}(1-P)^n && \text{if } 2PN - N \leqslant \nu < PN \\
p &= \sum_{n=0}^{\nu} \binom{N}{n} P^{N-n}(1-P)^n && \text{if } \nu < 2PN - N
\end{aligned} \tag{5.5}$$

where the probabilities P are given by $P_0(\kappa|0)$. Mind that some of the summations strictly can only be performed when PN is an integer. If not, either a continuous approximation of the distributions and integration can be used, or the differences $2PN-\nu$ should be rounded to the closest integer value away from PN. Here we do the latter.

[1] Mind that although the quantity in equation (5.4) plays the role of a χ-square, critically considered it is different, since its error terms are not normally distributed.

Note further that since we perform $M = \sum_{l=1}^{L-1} l$ pairwise checks, we should adjust the significance level accordingly, in order to minimize false positives. If we have a significance level α for each pairwise comparison, the chance that we wrongly reject the null hypothesis in at least one of the cases is given by $1-(1-\alpha)^M$. Therefore, to perform a strict hypothesis testing, one should require $1-(1-\alpha)^M < 0.05$, which yields $\alpha \approx 0.05/M$. Mind that with this stringent criterion we might assume incorrectly that the null hypothesis is not rejected.

caveats

Some caveats are in place. First, the value of Δ can only be determined accurately if enough sequences of the PCR products are obtained. What 'enough' is, depends also on the number of polymorphic loci: if there are few of these, Δ will also be less accurate. Further, the determination of Δ should in principle be done on the basis of unselected sequences. However, since the interesting *functional* correlation will only show up after selection, more data will probably be gathered from selected sequences. We expect that, as long as functional correlations are not too prominent, the above determination of the run length Δ will be reasonably accurate. Third, one may argue that the elongation steps are not well described by a Poisson process. One could indeed expect that if primer annealing is fast compared to the elongation, then the run length will be more directly determined by the duration of the temperature step. This would cause the distribution of Δ to be more peaked around a certain value. On the other hand, replication of a DNA strand using already mixed strands as a template (as will happen more often towards the end of the PCR reaction), will tend to randomize the run lengths again. The in the end (semi-)random nature of the process was our rationale to model it as a Poisson process and we expect that the influence of a deviation from Poisson statistics on the detection of functional correlations is limited.

That the average effective run length is not simply determined by the duration of the elongation steps is clear: from our estimated run length of 199 (see below), using a annealing time step of 30s and an elongation time of 20s, and given the measured rates of DNA replication rates of 10^2-10^3 nucleotides per second [197, 198], we can conclude that the 72^{o}C time steps cannot directly be the run length determining factor. To assess which factors exactly determine the run length is less important for the present work. Apart from the earlier mentioned replication of already mixed templates, run lengths could be influenced by association rates of primers to their complementary strand, association rates of DNA polymerase to the primer-template complex, or the initiation rate of the replication process.

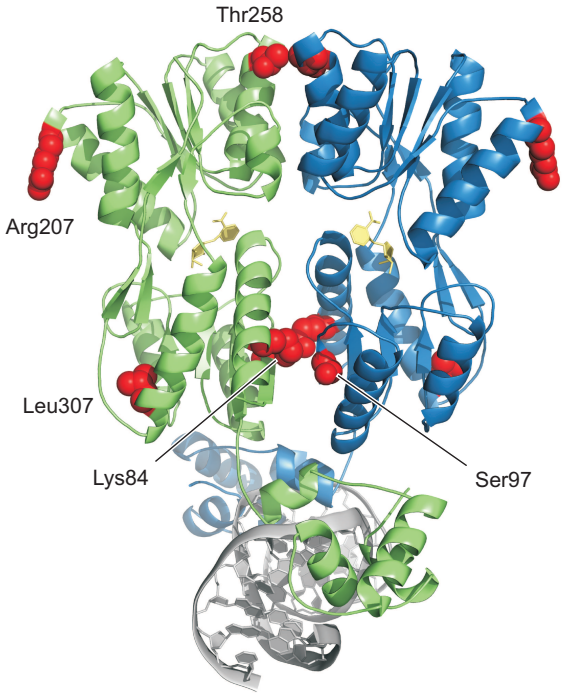
Lastly, here we only look at pairwise correlations, but not at higher order interactions, which could be important, but will be missed by the above treatment. The higher order correlations could be approached in a similar way, but to do this meaningfully, we would have to have many more sequenced recombinants available than we have (here order 20).

5.2 Results

Figure 5.2 shows the wild-type (dimeric) LacI protein structure. Indicated in red space fills are the amino acids that are mutated in the inverse repressor that is studied in the present work. This repressor contains 8 mutations (see table), that give rise to 6 amino acid substitutions, as 2 base pair substitutions are synonymous. The repressor phenotypes in this work are described by two parameters: the basal expression (no inducer present) e_0, and the regulation factor F, that gives the fold change in expression in media that do contain inducer (here 1 mM IPTG). Hence, repressors with an F value smaller than 1 exhibit an inverse phenotype, in which case IPTG acts as a co-repressor rather than as an inducer. The initially evolved inverse repressor has an F value of 0.022 and a basal expression level e_0 of $8.8{\cdot}10^7$, whereas the wild-type *lac* repressor in our assay has an F of 27.4 and an e_0 of $3.5{\cdot}10^5$.

We performed a mutation-dilution PCR on a 1:1 mix of wild-type and evolved sequences, and subsequently created a pool of $\sim 1{\cdot}10^6$ *E. coli* cells carrying a plasmid-borne mutational intermediate. Using a selection operon developed earlier (chapter 4), we selected for inverse functionality (F values smaller than 1), which here is a functionally conservative (purifying) selection. Mutants before and after selection are isolated, and their phenotypes are assayed (fig. 5.3). In figure 5.3 wild-type phenotype (black star) is indicated as well as the inverse repressor containing all 8 base pair substitutions (white star). From the location of the mutation-diluted pool before selection (open circles), one can see that the removal of mutations yields phenotypes that are also largely intermediate between wild-type and inverse repressor. After selection (solid squares) phenotypes again surround the evolved inverse repressor phenotype. Interestingly, the low expression level of the measured isolates (which is e_0 for F>1 and $e_0 F$ for F<1) seems to be very similar for most isolates, being slightly below $1{\cdot}10^6$. This seems to suggest that there is no locus among the 8 base pair substitution that by itself tightens the binding of the repressor to the DNA. This in turn suggests that the modus by which the inverse repressors acquire their functionality is not the kinetic effect proposed in ref [175].

From a subset of isolates before and after selection the *lacI* sequences are determined (table 5.1). We can see that most selected sequences indeed contain only a subset of the 8 mutations. In order to obtain a first indication about which loci are important for the inversion of the *lac* repressor functionality, we show p-values the occurrences of each of the 8 polymorphic loci among sequences that have a F value lower than 0.2 (that are given in table 5.2). The p-values are calculated on the basis of a binomial distribution that assumes equal chances for the presence and absence of the mutations. Also given for comparison is the result of a multi-factor ANOVA with unequal replication [199,200], on the basis of all available sequences. Note that due to the limited number of obtained sequences we cannot perform a ANOVA test that includes interaction terms. Would we have had all 2^8 possible combinations of the 8 polymorphic loci, then an ANOVA test would have given us all direct influences of loci on the



bp	aa
A30T	V9s
C39A	A13s
A251T	K84I
T289C	S97P
G620T	R207L
A772G	T258A
T920A	L307H
T1046C	L349P

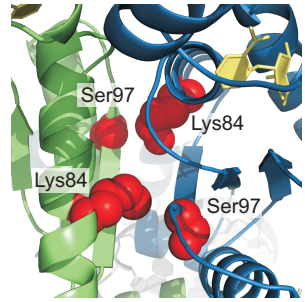


Figure 5.2: Structure of the C-terminal deleted, but otherwise wild-type *lac* repressor (PDB 1JWL) with bound ligands (yellow). Amino acids that are mutated in the inverse repressor are rendered as red space-filling residues. The inset on the right shows the central part of the repressor, tilted forward to provide a clearer view on residues 84 and 97. The table on the right shows the base pair substitutions (left), and the corresponding amino acid changes (right) in inverse repressor versus wild-type. A30T and C39A are silent. Mutation Leu349Pro is located in the C-terminal tetramerization domain of the protein, and is therefore not shown in the structure.

inverse phenotype, as well as all interaction terms.

A first inspection of table 5.2 learns that both ways of analysis (binomial and ANOVA) indicate a high functional significance to mutation Ser97Pro. From the last column we see that its effect is to lower the F value. Earlier work showed that this mutation also arose in other, independent, lineages where a selection for inverse phenotype was performed. Remarkably, however, this *amino acid* substitution was always accomplished by the same *base pair* mutation T289C. This in principle could be a hint towards a mutational bias rather than a functional substitution. Moreover, it was found that this substitution in isolation does not result in an inverted phenotypic response [177]. On the basis of these considerations we could expect that Ser97Pro interacts epistatically with some of the other mutations. On the other hand, the ANOVA test does give an indication for a direct effect of Ser97Pro. Reconstruction of a mutant with Pro97 should

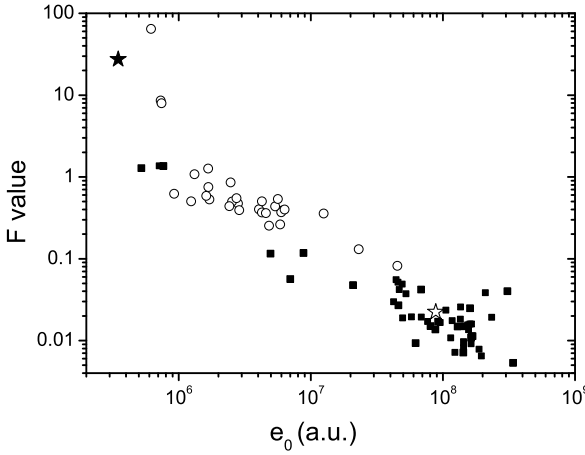


Figure 5.3: Measurement of the phenotypic parameters F and e_0, for a sample of mutation-diluted sequences before selection (open circles), and after selection (closed squares). Wild-type (ancestor) phenotype is indicated with a filled star, evolved inverse phenotype (descendant) is shown as a white star.

decide what its role is.

Both tests also seem to agree on the significance of substitution Leu307His, although less pronounced, to promote inversion of the *lac* response. Interestingly, this locus in isolation (fifth sequence in table 5.1, taking into account that the first two loci are neutral), on the contrary yields a higher F value than wild-type. From chapter 7 we can see that this suggests a destabilization of the repressor-DNA interaction (we argue there that a reduced binding affinity in an overexpressing system leads to a down-shift of the IPTG concentration that is necessary to induce the system, and hence to a higher expression at 1 mM IPTG used here). Again an epistatic interaction is suggested here: substitution Leu307His alone increases the F value, but occurs significantly more often among the sequenced inverse phenotypes (low F value) than expected . Substitution Leu349Pro also seems to occur significantly more often than expected in the selected sequences, but the ANOVA test disagrees here, which is caused by the fact that in the selected sequences Leu349Pro only once occurs in the absence of Leu307His. Their occurrence thus seems correlated in the selected sequences, to which we will come back below. Finally, table 5.2 shows that the synonymous base pair substitutions indeed have no effect on the phenotype (values in the 'effects' column around 0). Their distribution seems perfectly random over the selected sequences (8 out of 17). However, note that the ANOVA test can not assign a p-value, which here is an indication that these base pair substitutions always occur paired (and no significance to individual effects can be assigned).

e_0	$2.48{\cdot}10^6$	$4.05{\cdot}10^6$	$2.31{\cdot}10^7$	$1.71{\cdot}10^6$	$6.18{\cdot}10^5$	$1.68{\cdot}10^8$	$1.43{\cdot}10^8$	$1.65{\cdot}10^8$	$1.64{\cdot}10^8$	$6.21{\cdot}10^7$	$1.60{\cdot}10^8$	$4.95{\cdot}10^6$	$1.29{\cdot}10^8$
F	0.856	0.402	0.131	0.532	64.5	0.0113	0.00714	0.0114	0.0109	0.0093	0.0249	0.116	0.0149
V9s	0	0	0	1	1	0	0	1	1	0	1	1	0
A13s	0	0	0	1	1	0	0	1	1	0	1	1	0
K84I	0	1	0	1	0	1	1	1	1	0	1	0	1
S97P	1	1	1	1	0	1	1	1	1	1	1	1	1
R207L	0	1	0	1	0	0	0	0	0	1	1	0	0
T258A	0	1	0	1	0	0	0	1	1	1	1	1	0
L307H	0	0	1	0	1	1	1	1	1	1	1	0	1
L349P	1	0	1	1	0	1	1	1	1	1	1	1	1

e_0	$6.83{\cdot}10^7$	$6.99{\cdot}10^6$	$8.78{\cdot}10^7$	$4.61{\cdot}10^7$	$1.47{\cdot}10^8$	$8.07{\cdot}10^7$	$5.23{\cdot}10^5$	$9.18{\cdot}10^7$	$3.40{\cdot}10^8$	$8.81{\cdot}10^6$	$7.67{\cdot}10^5$	$2.09{\cdot}10^7$	$3.08{\cdot}10^8$
F	0.042	0.0566	0.0137	0.0271	0.015	0.015	1.28	0.017	0.00534	0.118	1.36	0.0477	0.0401
V9s	0	0	0	1	1	0	0	0	1	0	1	1	1
A13s	0	0	0	1	1	0	0	0	1	0	1	1	1
K84I	0	0	1	1	1	0	1	1	0	0	1	1	1
S97P	1	1	1	1	1	1	0	1	1	1	0	1	1
R207L	0	0	0	1	0	1	1	0	1	1	0	1	0
T258A	1	0	0	1	0	1	1	0	1	0	0	1	1
L307H	1	1	1	1	1	0	0	1	1	1	0	1	1
L349P	1	1	1	1	1	0	0	1	1	1	0	1	1

Table 5.1: Sequences of 26 mutation-diluted isolates. A '0' denotes wild type (ancestor), and a '1' means evolved sequence. First 5 are from the pool before selection, rest from after selection. Measured basal expression levels e_0 are given as well as F values.

	occurrence	p-value	ANOVA F	p-value	effect
V9s	8	1.00	0	–	0.13
A13s	8	1.00	0	–	0.13
K84I	12	0.14	3.42	0.080	-0.49
S97P	17	0.000015	7.65	0.012	-1.37
R207L	6	0.33	0.09	0.77	0.21
T258A	10	0.63	1.59	0.22	-0.28
L307H	16	0.00027	4.25	0.053	-0.97
L349P	16	0.00027	0	0.97	-1.20

Table 5.2: Left two columns: occurrences of mutations in sequences with a measured F value of less than 0.2 and the associated p-values based on a binomial distribution (p=0.5,N=17). Right two columns: Multi-factor ANOVA (with unequal replication) test statistic F and its p-value (only linear terms taken into account) on the basis of all 26 sequences plus wild-type. The ANOVA test is performed with respect to the logarithm of the F value. The last column contains an expression for the effect of the mutation, $\overline{\log F_{\mathrm{mut}}} - \overline{\log F_{\mathrm{wt}}}$, being the difference between the averages of the logarithmic F values for sequences that have and do not have the mutation.

	V9s	A13s	K84I	S97P	R207L	T258A	L307H	L349P
V9s	1.00	0.0000019	0.26	0.66	0.50	0.012	0.66	1.00
A13s		1.00	0.26	0.66	0.50	0.012	0.66	1.00
K84I			1.00	0.50	0.19	0.66	0.12	0.26
S97P				1.00	0.19	0.82	0.00040	0.000040
R207L					1.00	0.12	0.19	0.078
T258A						1.00	0.66	1.00
L307H							1.00	0.000040
L349P								1.00

Table 5.3: Overview of p-values stating the significance of the deviation from the expectation of pairwise correlation without taking PCR effects into consideration. The p-values based on a binomial distribution (p=0.5,N=17).

Next we will focus on the analysis of functional correlations between the mutations. So far we have been able to only discuss direct effects of the substitutions, or had to refer to additional information to speak about possible epistatic interactions. Here we will concentrate on the significance of correlations in the occurrence of substitutions directly. We will compare a naive analysis of correlations to the analysis developed in section 5.1.2. Table 5.3 lists p-values expressing the significance of the pairwise presence or absence of substitutions in the selected sequences. The p-values are calculated on the basis of a binomial distribution assigning equal chance to a pair of loci having the same or a different origin. This table states a highly significant correlation between the two synonymous mutations. As can be seen from table 5.1, they indeed only occur in pairs, but their overall effect on the phenotype is negligible (table 5.2). In principle when this occurs, this could be an indication for (reciprocal) sign epistasis (see chapter

	V9s	A13s	K84I	S97P	R207L	T258A	L307H	L349P
V9s	1.00	0.63	0.50	0.50	0.50	0.012	0.66	1.00
A13s		1.00	0.50	0.50	0.50	0.012	0.66	1.00
K84I			1.00	0.0083	0.18	0.66	0.12	0.26
S97P				1.00	0.18	0.82	0.00040	0.000040
R207L					1.00	0.50	0.18	0.075
T258A						1.00	0.17	0.83
L307H							1.00	0.0037
L349P								1.00

Table 5.4: Overview of p-values stating the significance of the deviation from the expectation of pairwise correlation due to the distance between the loci. Recombinant sequences are included, when their F value is below 0.2. Dark grey cells indicate significant (positive) correlation. Lighter grey cells indicate potential (anti-)correlations.

2). In this case we know that the individual effects of the two synonymous mutations will also be negligible. The correlation will most probably be caused by the fact that the two mutations lie only 9 base pairs apart.

In order to remove these distance effects we applied the analysis from section 5.1.2 to the sequences of the selected pool. We obtained an effective average run length Δ based on all sequences of 199 base pairs. In figure 5.4, we show how the quantity χ^2 depends on Δ, and that the minimum is a clearly defined value. Based on this run length we calculated now the p-values of pairwise correlations, corrected for the distance effects (table 5.4). Note that these correlations address the *occurrences* of the mutations, and are not a direct measure for the interaction between loci in their effect on phenotype (e.g. if two loci additively affect the phenotype, they will also show up in the correlation measure here). The purpose of the correlation measure as developed here is to signal potentially functionally interesting loci, that affect function either directly or through epistatic interactions.

Comparing table 5.4 with table 5.3, we indeed see that the high correlation in the presence or absence of the two synonymous base pair substitutions is entirely due to distance effects: their p-value changes from highly significant when distance effects are not taken into account to highly non-significant. Further we see that correlations between Ser97Pro and Leu307His, as well as between Ser97Pro and Leu349Pro remain significant. As we saw earlier, the effect of the former pair is probably epistatic. What is further interesting from table 5.4 is that there is an anti-correlated pair on the border of significance: Lys84Ile and Ser97Pro. Although the occurrence of Lys84Ile seems to be random and its correlations with Ser97Pro seem to be negligible if distance effects are not taken into account (table 5.3), based on the proximity of Ser97Pro it should occur more often than it does. Interestingly, in the *lac* structure (fig. 5.2 we can see that the residues 84 and 97 spatially lie very close to each other (both within and among dimers), and their interaction is not unimaginable.

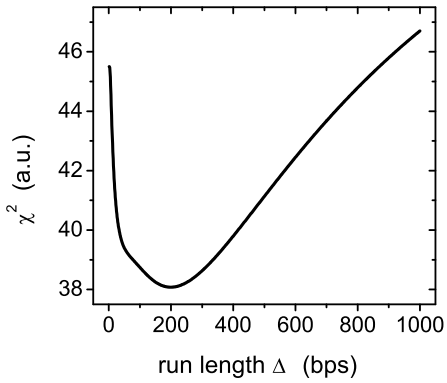


Figure 5.4: χ^2 (equation (5.4)) as a function of run length Δ, on the basis of 26 recombined sequences recovered, from before and after selection. A clear minimum is found for Δ = 199 base pairs.

On the basis of our analysis, for the fitness landscape analysis of a subset of mutations we will focus on combinations of Ser97Pro, Leu307His, and Leu349Pro. This is ongoing work. Among our sequenced isolates, we also found a strongly inverse phenotype that has the substitutions Ser97Pro, Arg207Leu, and Thr258Ala. Intermediates of these will also be further investigated.

5.3 Discussion and Outlook

In this work we developed a general method that together with the reconstruction of specific subsets of intermediate mutants can be applied to acquire insight in evolutionary pathways between an ancestor and evolved sequence or protein. When there are many mutational differences between ancestor and descendent, it might seem impossible to extract information about the potential evolutionary trajectories. However, in cases where there are many functionally neutral loci, attention can be focused on a subset of mutations that are most important for changes in function. Here we applied the method to an evolved inverse *lac* repressor. As a proof of principle, we showed how our analysis correctly discards the distance correlation between two closely located synonymous mutations. Targeted construction of potentially interesting intermediates is ongoing work. In principle, since the total number of amino acid substitutions here is not that large, a complete reconstruction of the phenotypic landscape might also be considered. A similar analysis could be applied to naturally occurring polymorphims, resurrected ancestral molecules, and artificially evolved systems alike, as long as we have access to phenotype or fitness. The approach should prove especially interesting in cases where loci affect phenotype or fitness in mutually additive clusters, within

which their effects are epistatically linked.

When attempting to infer the evolutionary process based on a reduced dataset, some essentials might be missed. For instance, one implicitly makes the assumption that loci that have no influence on the present-day phenotype or fitness, were also neutral when they occurred in evolutionary history. This is not necessarily so. And *vice versa*, mutations that now have a clear influence on the present function, might not have increased fitness when they occurred. Both scenarios may occur in the process of 'molecular cooption', or 'molecular exploitation' (e.g. [57]), where a molecule takes on a new functional role and thus will experience a different selective pressure than before. More broadly, in assaying the characteristics of an ancient molecule or system, one often has to assume that it evolved under a selective pressure similar to the one it experiences at present. Interesting counterexamples can be found where a punctuated adaptation proceeds via intermediate wanderings over so-called neutral nets [45, 66]. In these cases adaptive mutations are not accessible without taking detours via neutral substitutions. In general, it will always remain important to work on the basis of plausible assumptions about the past selective pressures. Nevertheless, the approach should be valuable in studying adaptation and the role of constraints.

In the present work, we obtained clear indications of the functional importance of mutations and for epistatic interactions, based on a reasonably low amount of data. At least three mutations in the evolved repressor seem to cooperate to yield an inverse response, though partly disjunct sets seem to be able to accomplish this. Together with one substitution that is possibly anti-correlated to the functionally important Ser97Pro substitution, it seems that epistasis is very prominent in this system. Apart from information about the evolutionary process this might also yield structural insight into the allosteric changes in the repressor protein. When combined with information obtained from inverse repressors from different lineages (chapter 4), information can be obtained about the structural plasticity of transcription factors to obtain novel responses to their ligands. We expect that reverse neutral evolution might be in itself an interesting tool to elucidate the structural basis of protein functions.

CHAPTER

6 Maintenance and loss of gene regulation in experimental evolution

Degeneration is a much commoner phenomenon than progress.

J.B.S. Haldane,
The Causes of Evolution

The evolution of gene regulation is a major open question in biology. Regulatory systems not only allow organisms to respond to a variable environment, but are themselves shaped by evolution under a variable selective pressure. When adaptation is approached as an optimization process, a variable environment adds many degrees of freedom to the search space compared to adaptation in a constant environment. How environmental variation selects for different modes of regulation, or in which cases other strategies than regulation are favored, such as bet-hedging, is presently under intense debate. These issues cannot be addressed at a theoretical level alone, and require information about the evolutionary plasticity and potential functional constraints of actual biological systems. In the present work we experimentally follow regulatory adaptation starting from a non-optimal regulatory response. We focus on regulation of lactose metabolism in Escherichia coli, which arguably exhibits a near-optimal relation between the amount of lactose in the environment and the level of expression of the lactose metabolic genes. By using separate compounds for induction and metabolism, we dislodged the lac regulatory response from its optimum, and predicted new optima. We followed adaptation in several constant and alternating environments. We find some cases of fast adaptation to the predicted optimum. In a number of instances adaptation occurred, but the predicted optimum was not reached. This may be due to a diminishing return of further optimization, or to the existence of a functional or genetic constraint.

All living organisms are equipped with mechanisms that enable them to sense their environment and respond to it. In many cases the response consists of regulating gene expression. For bacteria the link between sensing and response is often formed by a small network of interacting proteins and regulatory sites on the DNA controlling the expression of downstream genes. Such a gene regulatory system can be characterized by its regulation function (or regulatory profile, or induction curve), that specifies the relation between the environmental signal and the expression level of the regulated genes.

Just as the adaptation of catalytic properties of enzymes or protein expression levels may be viewed as an optimization process [43, 73, 144], so may be the adaptation of regulation. However, in the case of regulation, the system is shaped by varying selective pressures in an environment that is fluctuating, which makes that there are many more potential parameters to optimize. Optimality can, for example, concern the nature of the environmental fluctuations (regular or stochastic), their time scales, the strength of selection in each environmental state, and how variable selective pressures constitute trade-offs experienced by the adapting organism. Moreover, similar regulation profiles can be accomplished by different modes of regulation, for example by employing a repressor (negative control), or an activator (positive control), and several theoretical studies have addressed their optimality in an ecological context [74–76]. Alternatively, in the case of slow and unpredictable environmental variation, it might be optimal not to employ regulatory systems, but to stochastically switch phenotypic states [77]. Such issues have mainly been been approached theoretically and experiments have been lagging behind. As a result, we lack essential information on regulatory plasticity and the potential constraints that hamper reaching an optimal regulatory response.

In this work we will consider the lactose operon of *Escherichia coli* from the viewpoint of optimality and explore adaptation of the regulatory response to new environments by experimental evolution [119].

6.1 Optimality of gene expression

Following ref. [73], we describe the growth rate of a population of *E. coli* cells as a function of expression of metabolic genes and carbon source (here lactose) in the environment in terms of the cost and benefit of gene expression

$$g = g_0 - \eta(Z) + B(Z, L) \tag{6.1}$$

where g_0 is the basal growth rate, set by compounds other than lactose in the environment. $\eta(Z)$ is the decrease of growth rate due to the burden of producing *lac* operon gene products LacZ, LacY, and LacA [134]. $B(Z, L)$ is the growth advantage due to lactose metabolism, which depends on both the expression level of the *lac* gene products (in particular LacZ), and the concentration of lactose in the environment.

This gives rise to an optimal expression level for each concentration of lactose in the environment $Z = Z_{\text{opt}}(L)$. At low levels of lactose the cost term will dominate the

benefit term, and the optimal expression level will be low or zero. Conversely, at high lactose concentrations the optimal expression level will be high.

The purpose of catabolic regulation is to sense the external concentration of the catabolite and to vary the expression level of metabolic genes as a function of this concentration: $Z = Z(L)$. Selection will drive a regulatory system towards the following optimality relation

$$Z_{\mathrm{opt}}(L) = Z(L) \tag{6.2}$$

implying that the system establishes a connection between the catabolic and inductive properties of lactose. Indeed, for *lac* regulation there are strong indications [73, 97] for this relation to hold. It is important to note that this criterion for regulatory optimality only concerns the relation between expression levels and catabolite concentrations. The regulatory system may well have to be optimized for response times, structural architecture, robustness, or otherwise.

In the present work we are interested in the evolutionary plasticity of the regulation profile $Z(L)$. In order to study regulatory adaptation, the system should be dislodged from its optimum, so that selective pressures arise that are directed towards a new optimum. This can be done in several ways. For example, one may change the kinetic parameters or the demands on the downstream regulated genes, which results in different cost and benefit terms and hence in different expression optima. Another approach, which we present here, is to decouple inducer and carbon source and allow the regulatory system to adapt to a new relation between the two, which is imposed by the experimenter. This approach mimics a situation in which an organism is confronted with a novel carbon source that has a different relation between induction and catabolic benefit.

For the *lac* system, a large number of artificial compounds have been synthesized [93], that interact with the gene products in a different way than lactose. The decoupling between *lac* signal and metabolism can be made by using isopropyl-β-D-thiogalactopyranoside (IPTG), and phenyl-β-D-galactoside (Pgal). IPTG is a gratuitous inducer; it binds to the *lac* repressor and relieves repression, but cannot be hydrolyzed by β-galactosidase. Pgal, on the other hand does not induce LacI, but is hydrolyzed by LacZ, releasing galactose (for further metabolism) and phenol. Now the optimality criterion reads

$$Z_{\mathrm{opt}}(P) = Z(I) \tag{6.3}$$

where P and I are independent variables (the Pgal and IPTG concentrations in the medium). In the present work we experimentally determined $Z(I)$ and the growth rate

$$g = g_0 - \eta(I) + B(I,P) \tag{6.4}$$

with which we can make a prediction for the selective pressures on the regulation for combinations of IPTG and Pgal concentrations. Subsequently we determined how regulation changed during experimental evolution, in which growing cultures of *E. coli* were serially passaged in batch cultures for around 800 generations.

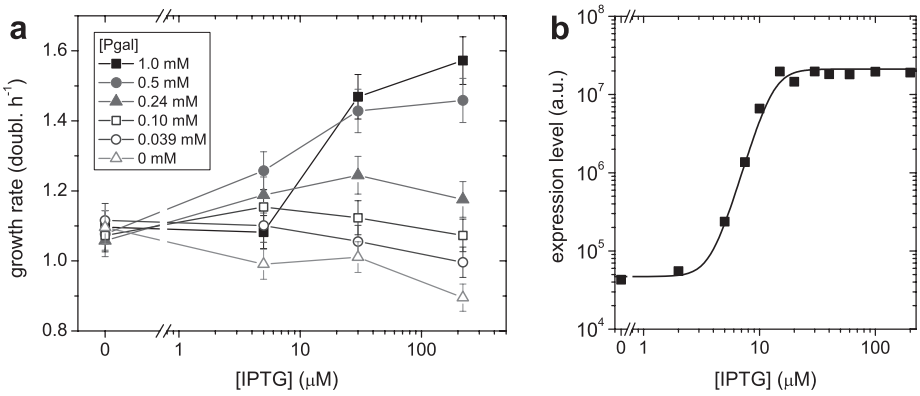


Figure 6.1: a) Measured growth rates as a function of Pgal and IPTG in a minimal M9 medium consisting plus casamino acids (see Materials and methods, section 6.3). At low Pgal concentrations inducing *lac* operon products will yield an expression cost that is larger than the benefit, resulting in a decrease of growth rates. For higher Pgal concentrations the benefit will dominate the cost. b) The induction profile in the absence of Pgal.

We first investigated adaptation of the *lac* system to a constant environment

$$Z_{\text{opt}}(P_1) = Z(I_1) \tag{6.5}$$

which in principle can be attained without regulation, as it only requires the optimization of one expression level.

A selective pressure for a regulatory response can be applied when cells experience multiple environmental conditions that impose opposite expression demands. The system is then confronted with a trade-off that can only be overcome by developing an appropriate regulatory response. To this end we performed evolution experiments in an environment that alternated between two states (P_1, I_1) and (P_2, I_2), so that the optimality criterion reads

$$Z_{\text{opt}}(P_1) = Z(I_1) \qquad \text{and} \qquad Z_{\text{opt}}(P_2) = Z(I_2) \tag{6.6}$$

with different optimal expression levels $Z_{\text{opt}}(P_1) \neq Z_{\text{opt}}(P_2)$.

6.2 Results and discussion

We determined growth rates of *Escherichia coli* MG1655 ('wild-type') cells [188] carrying the *lac* operon, as function of IPTG and Pgal in a minimal medium where casamino acids set the basal growth rate g_0 to be 1.09 generations h^{-1} (see Fig. 6.1). A measured wild-type induction profile is also shown. We observed that when the medium does not contain a carbon source ([Pgal]= 0 mM), induction leads to a cost. The growth rate decrease is 0.20 doublings h^{-1} per hour for full induction.

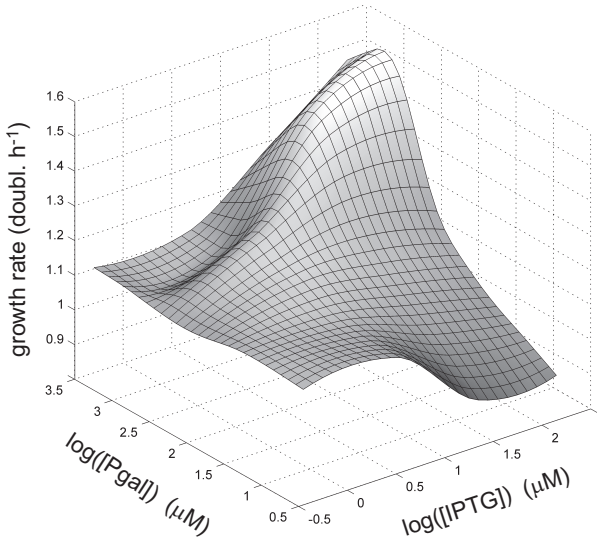


Figure 6.2: Interpolated and smoothened growth data from figure 6.1, providing an impression of the functional form of equation (6.4). The decoupling of inducer and carbon source is visualized: addition of Pgal when not expressing *lac* operon genes (low IPTG concentration) will not result in growth rate increases, and addition of IPTG without Pgal will lower the growth due to an expression cost. The ridge in the landscape is caused by anti-induction of the *lac* repressor by Pgal, when present at a high concentration (see section 6.4). For low Pgal concentrations we used a functional relation for the cost of expression fitted to the data (section 6.4).

At higher concentrations of Pgal cost and benefit are balanced for intermediate inducer concentrations: growth in the presence of 0.10 mM and 0.24 mM Pgal is maximized for IPTG concentrations near 5 μM and 30 μM respectively. For higher Pgal concentrations the maximum observed growth rates lie at inducer levels of 200 μM or higher. We observed that high concentrations of Pgal have an anti-inductive effect due to the competitive binding of IPTG and Pgal to the repressor (see section 6.4). Since the affinity of IPTG for the repressor is much higher than that of Pgal (a K_D of $1{\cdot}10^{-6}$ M versus $1{\cdot}10^{-3}$ M [201]), this effect can be neglected for sufficiently low concentrations of Pgal. We see the effect of anti-induction in the growth data for medium containing 1 mM Pgal, which for 5 μM IPTG has a lower growth rate than medium containing 0.1-0.5 mM Pgal.

In figure 6.2 we show an interpolation and smoothening of the growth data to give an impression of the optimality relations between IPTG and Pgal for the *lac* operon. For low concentrations of IPTG (low expression level) we recover the basal growth rate g_0 independent of the Pgal concentration, while for high IPTG concentrations (high expression level) a benefit only occurs in the presence of a high enough concentra-

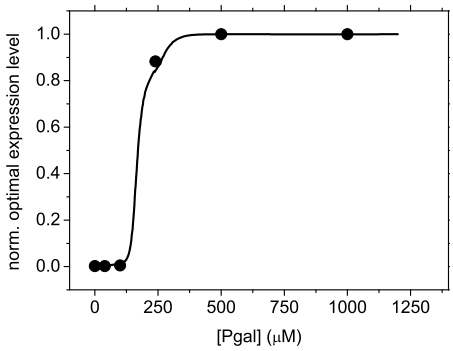


Figure 6.3: Optimal expression levels of *lac* operon genes as a function of Pgal, obtained from the landscape in figure 6.2. The circles represent the optimal expression levels obtained from inspection of the raw growth data in figure 6.1.

tion of Pgal. From these smoothened growth data, we recovered the optimal expression levels for LacZ using the induction profiles that we measured for different concentrations of Pgal (section 6.4). This optimality relation is given in figure 6.3, together with the optimal Pgal concentrations as obtained directly from the growth data in figure 6.1. Although the optimal level shows a very sharp Pgal dependence, this does not mean that the growth difference for optimal and non-optimal expression are necessarily large. We can see from the landscape in fig. 6.2 that for the Pgal concentrations around the inflexion point of fig. 6.3 (~150 μM), the growth rates for high and low expression are very similar. This implies that non-optimality at these Pgal concentrations will not result in high selective pressures.

We modeled the cost and benefit aspects of our system given by equation (6.4) in a similar fashion to what was done in refs. [202] and [73]. Since in our system induction and catabolite are separated, we included IPTG induction and anti-induction for high concentrations of Pgal (section 6.4) in the model, based on independent measurements of the expression levels of LacZ. Although we obtained a clear qualitative agreement, a quantitative agreement was reached only for higher concentrations of IPTG. Data and model for 220 μM IPTG are shown in figure 6.4, and model predictions other IPTG concentrations are given in section 6.4. The observed discrepancies point at an interesting issue: whereas expression levels only rise marginally for IPTG concentrations up to 5 μM, cost and benefit already diversified the obtained growth rates (see figure 6.1). Adjustment of the model is required (section 6.4).

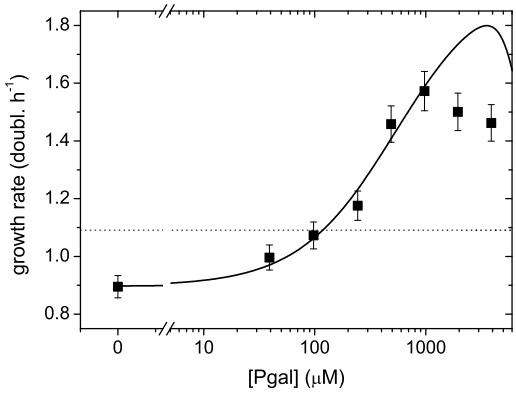


Figure 6.4: Fit of growth data at high induction (220 μM IPTG) using a reaction kinetics model (section 6.4). The dotted horizontal line indicates the growth rate in the absence of Pgal and IPTG. The growth difference between this line and the data point at [Pgal] = 0 represents the cost of protein expression. For higher Pgal concentrations this cost is compensated and eventually dominated by the benefit of Pgal metabolism. Cost and benefit are balanced for a Pgal concentration of $1.2{\cdot}10^2$ μM.

6.2.1 Evolution in constant environments

We performed a serial dilution experiment in a number of constant environments with different concentrations of IPTG and Pgal, as indicated schematically in figure 6.5. For each condition, a 10 ml culture was grown and diluted twice daily 300-500 fold for a total of ~800 generations. Each week a sample of each culture was stored at -80^{o}C to preserve snapshots of its evolutionary history. Afterwards, the LacZ activity[1] of the adapting populations was determined for different time points during the experiment.

The history traces in figure 6.6 indeed show that the expression level of LacZ changes during the adaptation experiment. As expected on the basis of the optimality curve in figure 6.3, cultures grown in the presence of 350 μM Pgal, but at IPTG concentrations that do not fully induce expression (fig. 6.6a and e), increase their uninduced expression levels. In these cases we observed a loss of repression. For the population grown without IPTG it takes ~200 generations before the uninduced levels resemble the induced levels. Notably, two replicate experiments performed at this condition (squares and triangles in fig. 6.6a), are indistinguishable.

[1]The measured LacZ activity in principle is determined by both the expression level and the kinetic parameters of LacZ. Therefore the LacZ activity is indicative for the expression level if we assume that the kinetic parameters of LacZ remain unchanged, as we do in this chapter. We argue that this assumption reasonable since we do not see an increase in maximum LacZ activity for adaptation lines that select for higher expression (fig. 6.6).

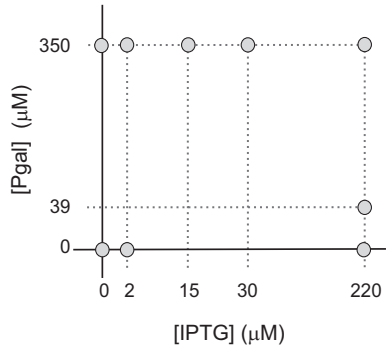


Figure 6.5: Overview of Pgal and IPTG concentrations of the constant environments in which adaptation experiments were performed.

Comparing the history traces of the cultures without IPTG (fig. 6.6a) to the trace of the culture grown at 2 μM IPTG (fig. 6.6e), we observe that the rates of adaptation are markedly different. If both traces are fitted with a simple competition model (assuming a single mutant fixation and a sufficiently high mutation rate to be able to neglect stochasticity due to bottlenecking the population, see section 6.4), we find that the selection coefficient of the population growing without IPTG is more than 4 times larger than that of the population at 2 μM IPTG (s = 0.055 versus 0.013)[2]. Although we would indeed expect the selection coefficient to decrease for increasing concentrations of IPTG, the observed large difference between 0 and 2 μM IPTG is remarkable in the face of the small expression differences between these IPTG concentrations in wild-type cells (see fig. 6.1b). On the other hand, from fig. 6.1a, we can see that wild-type cells for a Pgal concentration of 0.5 mM already realize more than half of their expression benefit at 5 μM IPTG. Consequently, at this IPTG concentration the additional selective advantage of abolishing regulation is decreased considerably compared to cells growing in the absence of IPTG. We further note that the absolute values of the selection coefficients found from the history traces are lower than expected on the basis of the wild-type growth rates. On the basis of the landscape in figure 6.2, we would expect a selection coefficient on the order of 0.2 for adaptation in the absence of IPTG.

In figure 6.6b we show the evolutionary trace of a culture grown under conditions of high amounts of carbon source ([Pgal] = 350 μM) and high induction ([IPTG] = 220 μM). No significant adaptation is observed, and as these conditions provide near optimal growth rates, this is as expected. Unchanged expression levels were similarly found for the culture grown at 350 μM Pgal and 30 μM IPTG (not shown). When fully induced, the regulatory system is in principle free to lose regulation by neutral drift: mutations that deactivate the repressor do not affect the growth rate. Since mutations that restore

[2]The selection coefficients determine the rate at which a mutant is fixed in the population, which in a history trace corresponds to the steepness of the curve (see section 6.4).

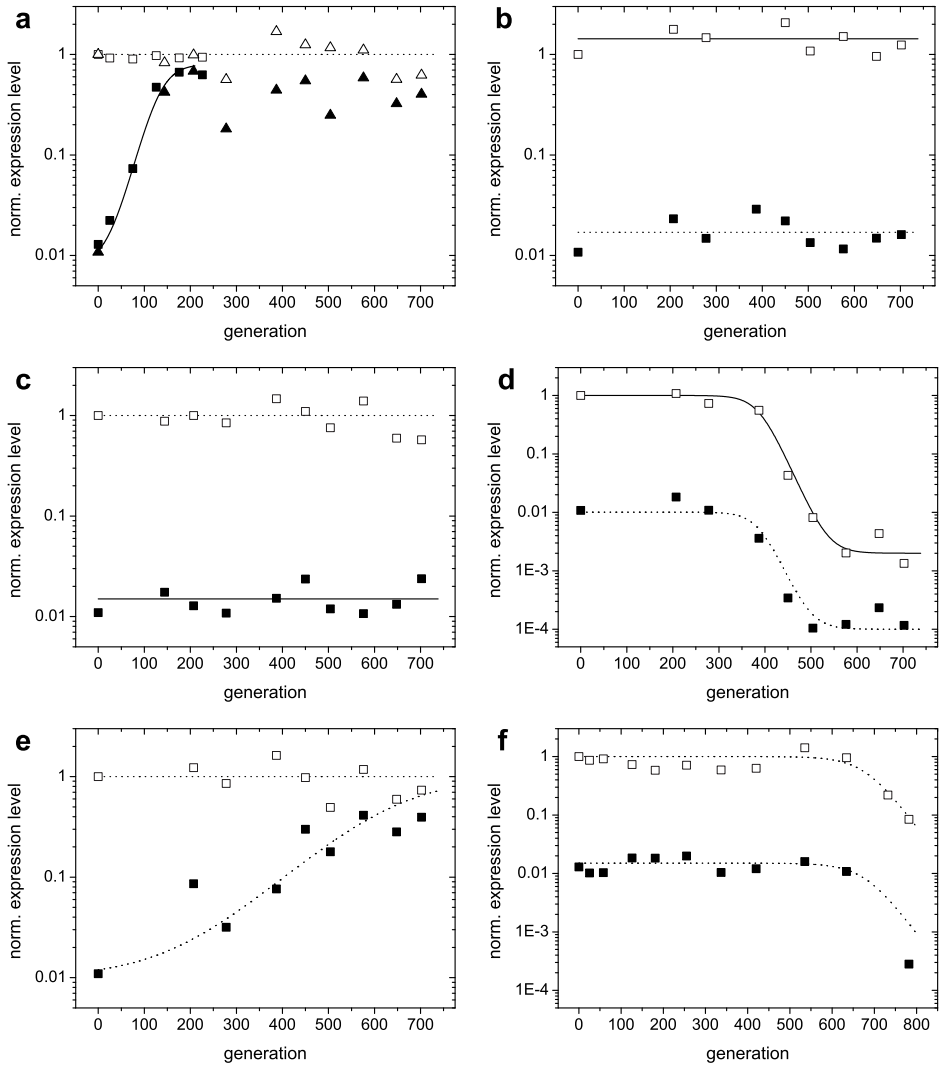


Figure 6.6: History traces of expression levels (population averages) for a subset of populations adapting in constant environments. Open symbols represent induced expression levels (at 220 μM IPTG), solid symbols are uninduced expression levels. Curves are fits based on growth rate differences under exponential growth (section 6.4). Where induction levels are the same as in the environment to which the populations adapted, the curves are solid. a) 0 μM IPTG, 350 μM Pgal. Two populations evolved in parallel are shown to yield the same adaptation dynamics (triangles and squares). b) 220 μM IPTG, 350 μM Pgal. c) 0 μM IPTG, 0 μM Pgal. d) 220 μM IPTG, 0 μM Pgal. e) 2 μM IPTG, 350 μM Pgal. f) 2 μM IPTG, 0 μM Pgal.

repressor function are in general much less likely to occur, in the long run repressor null mutants will fix in the population. However, the expected rate at which this would occur is on the order of $1/\mu$ generations [203], where μ is the mutation rate towards *lacI*$^-$ mutants, being ~$1{\cdot}10^{-6}$ [204]. If repressor deactivation is neutral, fixation would only be expected after $1{\cdot}10^6$ generations. Interestingly, a null mutation in the promoter controlling the transcription of the repressor may actually be selectively favored, since it should reduce the cost associated with the production of repressor protein. On the basis of the low amount of repressor protein compared to the other *lac* gene products, we expect that the selection coefficient associated with the loss of repressor production are too low to be observable within the time course of the experiments performed here.

Figure 6.6c also shows an unchanged regulation. Here the medium contains no IPTG and no Pgal, and expression of *lac* operon products would only incur a cost. A similar argument as above for the neutral loss, now of downstream operon products (LacZ, LacY, and LacA) could be developed. Indeed no change is expected within the 800 generations followed here.

Expression is drastically reduced during growth on 200 μM IPTG and 0 μM Pgal (fig. 6.6d). The rate at which the expression decreases in the population suggests a selection coefficient of around 0.067, comparable with the loss of repression observed in fig. 6.6a. However, the fact that fixation occurs at later generations indicates that this type of mutants occur less frequently than the repressor null mutants.

Figure 6.6f shows the evolutionary history of a population growing without Pgal, but with 2 μM IPTG. Here again expression of downstream genes is decreased. We observe a selection coefficient that is roughly half of that in fig. 6.6d. Apparently, 2 μM IPTG increases expression of downstream genes enough to be selected against in the absence of carbon source.

Two conditions indicated in fig. 6.5 remain. For medium containing 39 μM Pgal and 220 μM IPTG we observed no significant change in expression levels, while we found (fig. 6.3) that the optimal expression level at this Pgal concentration would be zero. However, from our landscape we would predict the selection coefficient for this condition to be around 2.5 times lower than for medium containing no Pgal and 220 μM IPTG. Assuming a similar mutation rate towards low expression for the medium condition discussed here, we do not expect observable changes in expression levels before generation ~900.

For the constant environment at 15 μM IPTG and 350 μM Pgal we found an altered induction profile (figure 6.7), such that the expression level at 15 μM IPTG remained similar to wild-type expression (taking into account Pgal anti-induction). From the data shown in our landscape in fig. 6.2 we can infer only a marginal difference in fitness between the expression level as induced with 15 μM IPTG compared to that at 220 μM IPTG. Interestingly, in this population a mutant was fixed that did not abolish repression altogether. Once fixed, the fitness advantage of a constitutive mutant would be minimal, and hence its chance to invade the population would be small.

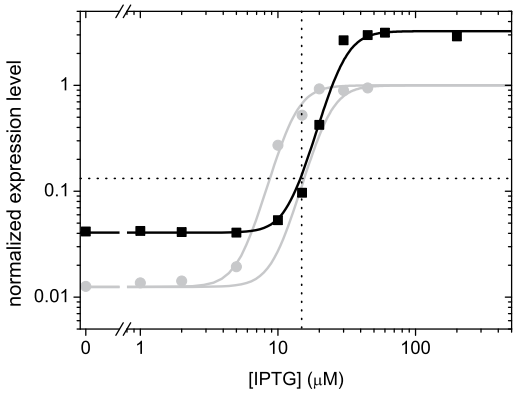


Figure 6.7: Induction profile (black squares and curve) of a population evolved for approximately 700 generations in the presence of 15 μM IPTG and 350 μM Pgal. Also shown is the wild-type induction profile (grey data points and fit), as well as the profile for wild-type incorporating Pgal anti-induction (grey curve shifted to the right). The expression levels at 15 μM IPTG for wild-type (with anti-induction) and evolved strain are very similar.

From 8 clonal isolates after the serial dilution experiment we sequenced the chromosomal region consisting of the *lac* repressor, the *lac* promoter (upstream of *lacZ*), until 420 base pairs into the *lacZ* coding sequence (see fig. 1.4 in chapter 1). Compared to the reference GenBank nucleotide sequence of the *lac* operon (accession number J01636.1), all isolates contain a often occurring *lacI* polymorphism (C857T) that does not affect LacI function, and a silent mutation in the coding sequence of *lacZ*. From earlier work we know that C857T pre-existed in the MG1655 strain, and we assume that the *lacZ* mutation did also. Apart from these mutations, three clones isolated from the population adapted to 350 μM Pgal, 0 μM IPTG all showed a known hotspot frameshift deletion of four base pairs from a triply repeated `TGGC` (nucleotides 593-604 of the *lacI* coding sequence) [204]. This frameshift leads to complete inactivation of the repressor [204], which corresponds to our observation. One clone sequenced from adaptation on 350 μM Pgal, 220 μM IPTG and another from 0 μM Pgal, 0 μM IPTG, which retained wild-type induction characteristics, did not reveal any mutations. Remarkably, three clones sequenced from the population that adapted to 220 μM IPTG, 0 μM Pgal, also showed the hotspot frameshift. These isolates do not show a constitutive expression, but instead a greatly reduced expression, which means that they must carry another mutation. However, since these isolates did not contain mutations in the promoter controlling *lacZ* expression, no cause for the observed loss of LacZ activity (which originated from selection against expression cost, not against activity) can be identified at present.

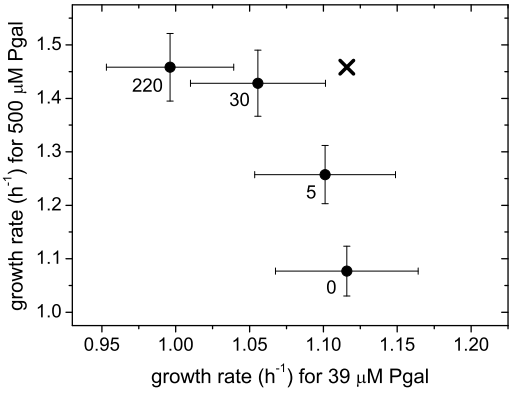


Figure 6.8: Example of trade-offs experienced when expression is not regulated in an environment that alternates between a low Pgal concentration (39 μM) and a high Pgal concentration (350 μM). Data points denote the growth rate in each environment for a certain constant expression level (at the indicated concentrations of IPTG in μM). This trade-off data is directly obtained from figure 6.1. Low expression levels (0 μM IPTG) yield optimal growth in medium with low Pgal concentrations, but non-optimal growth in medium containing high Pgal concentrations, and *vice versa* for high expression levels. Only when expression is regulated (low in low Pgal conditions and high in high Pgal conditions), overall growth over both environments can be optimal (indicated by the black cross).

6.2.2 Evolution in alternating environments

Regulation is favorable when an organism is confronted with a fitness trade-off, which occurs when optimizing the expression level in one state decreases the fitness in the other state. Using the growth data in figure 6.1 we can visualize such trade-offs. In figure 6.8, we plotted the growth rate in an environment with a high Pgal concentration (500 μM) versus the growth rate in an environment with a low Pgal concentration (39 μM). The IPTG concentrations to which these data points belong are given in the figure (in μM). The point of maximum growth rate under both conditions is marked with a cross. The figure suggests that there is no intermediate inducer concentration (hence expression level) that provides optimal growth under both conditions. From the concave shape of the curve we would expect that in an environment that alternates between these states in equal periods, an unregulated expression level would be evolutionary adjusted towards an intermediate level between low and high expression (see chapter 4).

We performed a number of serial dilution experiments in which the environment was alternating between two states (fig. 6.9). A change of environment was accomplished once or twice daily (see section 6.3). For 4 out of 6 experiments (marked with

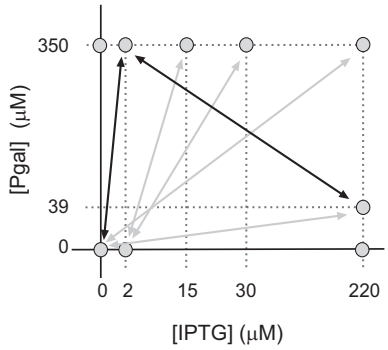


Figure 6.9: Overview of Pgal and IPTG concentrations of the alternating environments in which adaptation experiments were performed. Grey arrows indicate conditions which resulted in unaltered induction profiles. Black arrows did result in adapted profiles (see text).

grey arrows in the figure) we found no significant change of the induction profile. Interestingly, from the experiments in a constant environment we know that there is a selective pressure to decrease expression levels for a population grown at 2 μM IPTG and 0 μM Pgal. However, for the cultures alternating between this condition and high Pgal (350 μM) plus moderately high IPTG (15 and 30 μM), we found no response to decrease the expression level at low IPTG concentrations. Growth in the high Pgal condition prevents loss of expression, so that it would be advantageous here to increase the ratio between low and high expression levels. Since this was not observed, this may indicate a functional constraint in the system. For the environment alternating between no IPTG, no Pgal and 220 μM IPTG, 39 μM Pgal, we would expect an overall decrease of expression on the basis of the optimality curve in fig. 6.3, but evolution in a constant environment of 220 μM IPTG, 39 μM Pgal already showed that the selection coefficients are probably too small to see adaptation here within 800 generations.

In two alternating environments the induction profile did change. First, alternating between 2 μM IPTG, 350 μM Pgal and 220 μM IPTG, 39 μM Pgal almost fully abolished regulation and acquired a high constitutive expression. These conditions were intended to elicit an inverted regulatory response to IPTG (as was accomplished in chapter 4), but the selective pressure to decrease expression in the presence of 39 μM Pgal was not strong enough to lower the expression level at high concentration of IPTG (at least not within the time course of this experiment). The recovery of a constitutive expression suggests that the inverted response is genetically less accessible. Whether continuation of the adaptation experiment would result in the optimal inverted response, or whether the fixation of an inactivated repressor constitutes an evolutionary dead-end cannot be decided at the moment.

Second, when the environment alternates between no IPTG, no Pgal and 2 μM IPTG, 350 μM Pgal, a constitutive expression results. An optimal regulatory strategy would have been here to change the inflexion point of the induction curve to lower

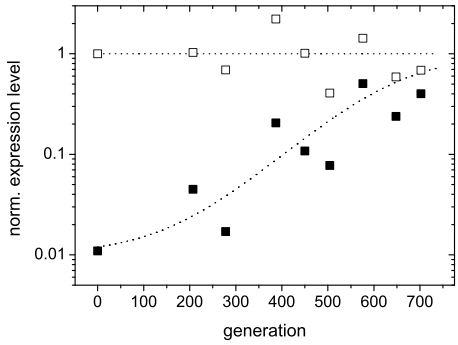


Figure 6.10: History trace of expression levels for the population evolved in an environment alternating between no IPTG, no Pgal and 2 μM IPTG, 350 μM Pgal. The population evolves towards a constitutive expression.

IPTG concentrations, which could result from a higher affinity of the repressor for IPTG. It is likely, however, that genetic changes that accomplish this are not easily accessible. The adaptation that occurred here maximizes growth in the environmental state with Pgal. However, the predicted fitness loss due to spurious expression in the state without Pgal is comparable. The fact that the mutation fixes rapidly in the population (fig. 6.10) on the other hand implies a considerable selection coefficient.

6.2.3 Conclusions and outlook

The decoupling of inducer and carbon source presented in this work provides a framework for studying the evolutionary plasticity of gene regulation. In most cases within the course of a few hundreds of generations a predicted optimal expression level was reached. In some cases optimal expression levels under the same conditions were not reached by populations that evolved in an alternating environment, when the other environmental state imposes a different optimal expression level. This implies that the lack of evolutionary change in these cases is not due to a lack of selective pressure, but points at functional constraints of the gene regulatory system.

We observed that mutations inactivating the repressor are more easily accessible than mutations that abolish expression, in cases where the selective advantages for such mutations are comparable. The high accessibility of the hotspot frameshift mutation [204], leading to *lac*$^-$ phenotypes at a frequency that is an order of magnitude higher than would be expected from the genomic mutation rate[3], is interesting in the

[3]The genomic mutation rate being $5.4{\cdot}10^{-10}$ per base pair per replication for E. coli [205], we would expect mutations to occur in the *lac* coding sequence (1080 base pairs) at a rate of $\sim 6{\cdot}10^{-7}$. It is known that roughly a quarter of substitutions is synonymous, and that around half of the amino acid substitutions in the *lac* repressor are deleterious [177]. This yields an expected rate of *lacI*$^-$ mutations of $\sim 2{\cdot}10^{-7}$. The observed rate is $2{\cdot}10^{-6}$ [204]

light of regulatory evolution. In fact, both deletions and additions of the 4 base pair repeat are observed at high frequency [204], which implies that reversals of hotspot mutations will also be more likely than reversals of e.g. base pair substitution that deactivate *lacI*. Together with the observation that the *lacI* coding sequence surrounding the hotspot is highly structured (palindromic), which elevates the mutation frequency due to slippage of a replicating DNA polymerase [206], it seems as if the *lac* repressor also has an in-built mutational regulation.

We obtained one instance (evolved at 350 μM Pgal, 15 μM IPTG) where the induction profile was not changed towards either constitutive expression or loss off expression. In some cases evolving towards constitutive expression was not predicted to be optimal on the basis of the measured growth rates from fig. 6.1. For example the culture alternating between 350 μM Pgal, 2 μM IPTG and 39 μM Pgal, 220 μM IPTG, would be optimal when the response to inducer would be inverted. However, it acquired a mutation which abolished repression first, after which the additional fitness gain of lowering expression for the high IPTG condition was diminished, and probably not enough to adapt within the course of 800 generations (even apart from whether such phenotype is still genetically accessible after the first mutation). The combination of fitness effect and mutational accessibility will determine which mutation will fix in the population first. This initial fixation is likely to lead to a 'law of diminishing returns' for subsequent mutations [27], and hence longer fixation times.

As a final remark, we note that the superior mutant would cheat our decoupling between inducer and carbon source and become responsive to Pgal, so that

$$Z_{\text{opt}}(P) = Z(P) \tag{6.7}$$

None of the evolved populations here was found to be induced by Pgal. Although this is an interesting issue, we expect that for this type of mutation to occur, a range of environments and an amount of generations need to be surveyed that is not easily accessible in this type of laboratory evolution experiments. Adaptation of inducer specificity is a subject probably more effectively studied at a higher level of control over the system, as was achieved in chapter 4.

In this work we presented an experimental approach to explore the adaptation of a gene regulatory response. Due to the many ways an environment can be variable, the survey of environmental conditions and fluctuations was necessarily limited. To obtain a high level of understanding of the evolutionary aspects of gene regulation will require a major effort, but which is necessary, as it increasingly appears that evolution is more strongly driven by regulatory changes than by modifications in non-regulatory proteins [165].

6.3 Materials and methods

Strains and media

We used *Escherichia coli* strain MG1655 [188]. All experiments were performed in M9

minimal medium, consisting of M9 salts (Sigma-Aldrich), supplemented with 0.1 mM $CaCl_2$ (Merck Eurolab), 1 mM $MgSO_4$ (Merck Eurolab), and 5 g/l casamino acids (BD Biosciences). When indicated, media contained isopropyl-β-D-thiogalactopyranoside (IPTG) and phenyl-β-D-galactoside (Pgal), both obtained from Sigma-Aldrich. All cultures were grown at 37^{o}C.

Determination of growth rates

Growth rate determinations were performed after overnight growth in the medium described above, without IPTG or Pgal, followed by at least 3 hours growth in medium with the appropriate concentrations of Pgal and IPTG. Subsequently the cultures were diluted to an optical density of $\sim 5 \cdot 10^{-4}$ and transferred to a pre-warmed flat bottom 96 well microtiter plate (VWR 351172), at 200 μl per well. Optical density at 600 nm was recorded in a Perkin & Elmer Victor^3 plate reader every 4 minutes, and every 29 minutes 9 μl sterile water was added to each well to counteract evaporation. When not measuring, the plate reader was shaking the plate at double orbit with a diameter of 2 mm. All presented growth values are averages of 3 independent measurements. From measurements in which all 96 wells were inoculated with wild-type MG1655, we determined the error margins on our averaged growth data to be 4.3%.

Determination of β-galactosidase activity

To determine the activity of β-galactosidase (LacZ) we used essentially the same method as described in section 4.1. Before transfer to a 96 well plate, cultures were grown overnight without IPTG and Pgal, and then diluted to an optical density of $\sim 5 \cdot 10^{-4}$. When expression levels were high so that overnight passage through stationary phase resulted in 'superinduced' LacZ activity levels (see section 6.4), growth times before fluorescence determination were prolonged.

As some of the expression levels in the current work were one or two orders of magnitude lower than the lowest expression levels in chapter 4, determination of the initial slope of these FDG curves was inaccurate due to experimental noise. In these cases we used the maximum slope at long timescales, which is proportional to the slope at t=0 (by a factor ~30, see section 4.1).

As before, during the assay of hydrolysis concentrations of IPTG and Pgal in each sample are made equal, to prevent unfair comparison due to competitive inhibition of LacZ by IPTG or Pgal.

Serial dilution experiments

10 ml cultures were grown in 50 ml flasks in a 37^{o}C water bath under vigorous shaking (200 rpm). Cultures were diluted 300-500x twice daily in fresh medium. As stationary cultures contain $\sim 10^9$ cells ml^{-1}, this implies bottleneck sizes of $\sim 10^7$ cells (for 10 ml total culture volume). The alternating conditions were either switched twice daily (for the cultures alternating between 2 μM IPTG, 0 μM Pgal and 15/30 μM IPTG, 350 μM Pgal, see fig. 6.9), or once daily (for the remaining conditions). When switching from a higher concentration of IPTG or Pgal to a lower one, cultures were washed 3x in minimal medium. Each four days a sample of the cultures was frozen at -80^{o}C. Re-inoculation

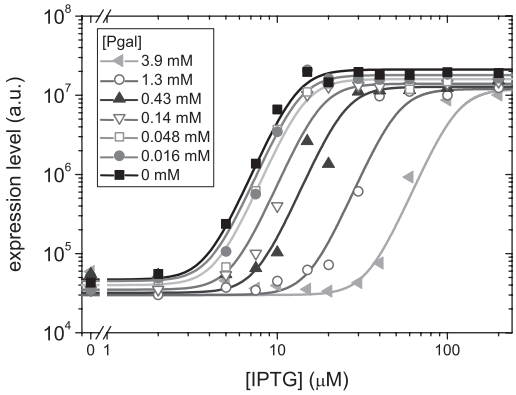


Figure 6.11: Induction profiles for wild-type cultures grown at different Pgal concentrations. The data is fitted with a Hill function that incorporates anti-induction by Pgal.

occurred after thawing and 3x washing in minimal medium.

6.4 Supplementary information

Analysis of induction curves and competitive inhibition by Pgal

The *lac* repressor does not only bind IPTG, it also has a affinity for Pgal. Pgal does not induce the repressor, but does competitively prevent IPTG from binding and thus effectively anti-induces the repressor. Since the equilibrium dissociation constants differ by three orders of magnitude ($K_D = 1{\cdot}10^{-6}$ for IPTG and $1{\cdot}10^{-3}$ for Pgal [201]), this effect is only noticeable when the Pgal concentration is much higher than the IPTG concentration. In other cases IPTG and Pgal can be considered to be decoupled with respect to induction of the *lac* repressor. We measured the effect of Pgal anti-induction by growing cultures under different concentrations of Pgal and IPTG. Immediately before the LacZ assay, all concentrations of IPTG and Pgal were equalized to prevent unequal inhibition at the level of LacZ. The results are given in figure 6.11.

The data for [Pgal] = 0 was fitted with a general Hill function (compare ref. [207])

$$\alpha_{\mathrm{IPTG}} = \alpha_0 \frac{1 + F\,([\mathrm{IPTG}]/C_{\mathrm{IPTG}})^m}{1 + ([\mathrm{IPTG}]/C_{\mathrm{IPTG}})^m} \tag{6.8}$$

where α_0 is a parameter relating the number of LacZ molecules to the measured LacZ activity, F is the ratio between induced and uninduced LacZ activity, C_{IPTG} is a dissociation constant associated with the affinity of IPTG to the repressor, and m is a phenomenological Hill exponent incorporating non-linear behavior of the *lac* induction (due to e.g. cooperative binding of the repressor to multiple operators).

The curves for higher Pgal concentrations in fig. 6.11 were obtained by incorporating Pgal anti-induction into the Hill description of the system. This is done in a similar way as in chapter 4, where we described and measured the competitive binding of FDG and IPTG to LacZ. Pgal anti-induction increases the effective equilibrium dissociation constant of IPTG to the *lac* repressor, (or equivalently lowers the effective IPTG concentration).

$$C_{\mathrm{IPTG,eff}} = C_{\mathrm{IPTG}}(1 + \frac{[\mathrm{Pgal}]}{K_P}) \tag{6.9}$$

where K_P is the apparent equilibrium dissociation constant of Pgal binding to the repressor (which incorporates a potential difference in internal and external Pgal concentrations).

All curves in fig. 6.11 could be fitted using a K_P of 0.45 mM. A minor vertical offset of the curves was observed due to the different growth rates under the different (Pgal, IPTG) conditions, as can be expected on the basis of a slightly different dilution rate (see e.g. equation (7.1) for the relation between growth rate and expression level).

Comparison to reaction kinetics model for transport and degradation

Here we modeled our system according to the cost-benefit analysis for a fixed expression level as reported in [73]. We modify this model by including induction by IPTG and anti-induction by Pgal. A comparison of this model to the obtained growth data in figure 6.1 is made.

The proposed [73] functional form for relative growth due to the cost and benefit of *lac* operon gene expression is

$$\Delta g = -\eta(Z) + B(Z,L) = -\frac{\eta_0 \cdot Z}{1 - Z/M} + \delta \frac{Z \cdot L}{K_Y + L} \tag{6.10}$$

where η is the cost term and B the benefit term, that depend on the concentration of LacZ (Z), lactose (L), and the equilibrium dissociation constant of the LacY permease and lactose (K_Y). This expression was derived under the assumption of low lactose concentrations (< 1 mM). This assumption assures that the rate limiting step in lactose metabolism is the import of lactose into the cell by LacY. In our system this condition is also fulfilled, as the equilibrium dissociation constant of LacY for Pgal is of the same order of magnitude as that for lactose, being 1.3 mM [208]. Pgal has a ~10 times higher affinity for LacZ [209], which justifies the assumption for Pgal.

We thus modified the expression for relative growth as follows

$$\Delta g = -\eta(Z(I,P)) + B(Z(I,P),P) = -\frac{\eta_0 \cdot Z(I,P)}{1 - Z(I,P)/M} + \delta \frac{Z(I,P)) \cdot P}{K_Y + P} \tag{6.11}$$

where the LacZ expression $Z(I,P)$ depends on IPTG as well as on Pgal concentration (the latter being of influence only at high concentrations).

Using this model we fitted the growth at high IPTG concentrations (220 μM), which works well for Pgal concentrations up until 1 mM (fig. 6.4). Indeed for higher concentrations the assumptions of the model may be violated.

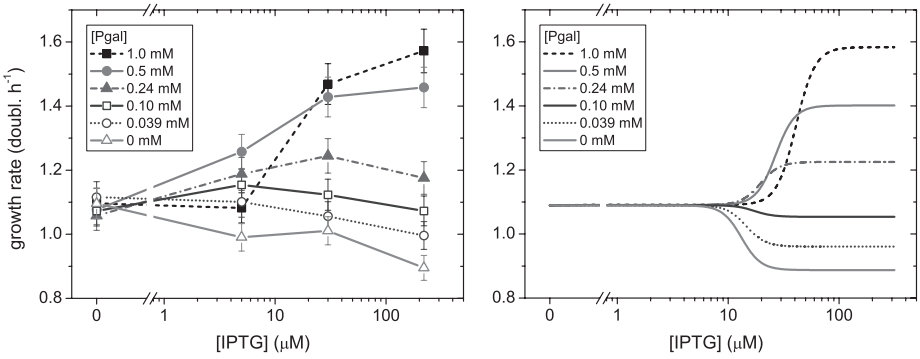


Figure 6.12: Comparison of growth data in the presence of various concentrations of IPTG and Pgal (left), with a model incorporating transport and catalysis of Pgal, as well as induction by IPTG and anti-induction by Pgal (right). The qualitative trend and the predictions at higher IPTG concentrations correspond well. For lower IPTG concentration there is a discrepancy: cost and benefit of expression occur at lower induction levels than is predicted by the model, on the basis of the measured induction profiles (fig. 6.11).

When we compare model predictions using the obtained parameters from the fit for 200 μM IPTG to reproduce the data at lower IPTG concentrations, we observe a qualitative correspondence only (fig. 6.12). The major difference is the occurrence of a cost and a benefit at very low IPTG concentrations (e.g. 5 μM), while LacZ expression levels have only increased marginally (from fig. 6.11 we see that they are still a factor ~100 below fully induced levels). These observations suggest that the cost and benefit terms may exhibit a steeper dependence on operon expression levels than assumed in the model. Alternatively, the model might need to incorporate competition between Pgal and IPTG for LacZ. However this would imply that Pgal import by LacY is not rate limiting, and hence violate the assumptions underlying the present model, which precludes an analytical solution. A numerical description of the system is ongoing work.

Non-stochastic competition model

Evolutionary traces were fitted (dotted and solid curves in expression history graphs) using a non-stochastic model for the change in expression when a mutant fixes in the population. It is known that the fate of mutants in a population that is periodically bottlenecked is influenced by 'sampling noise' when the mutation is initially only present in a few individuals [210]. However, when the mutation rate is such that the expected number of mutants after bottlenecking is significantly larger than 1 ($\mu b \gg 1$, where b is the bottleneck size and μ the mutation rate), these stochastic effects can be ignored. This seems to hold in our case at least for the hotspot mutations that occur at a rate of $\sim 1 \cdot 10^{-6}$, while our bottleneck size is $\sim 10^{7}$. Moreover, the selection coefficients are estimated from the rate of the fixation process, which is independent of the mutation

rate (unless the population is very small or the mutation rate very high).

As both the wild-type and the mutant population grow exponentially in between the bottlenecks we have for their numbers

$$\begin{aligned} \text{wild}-\text{type} \qquad & N(t) = N_0 \mathrm{e}^{\ln 2 g t} \\ \text{mutant} \qquad & N^*(t) = N_0^* \mathrm{e}^{\ln 2 g(1+s) t} \end{aligned} \tag{6.12}$$

where g is the wild-type growth rate, s is the selection coefficient. On the basis of these numbers of individuals, we have for the expression levels of a population average

$$E_{\text{ave}}(t) = \frac{E^* N^*(t) + E N(t)}{N^*(t) + N(t)} = E \frac{E_r R_0 \mathrm{e}^{\ln 2 g s t} + 1}{R_0 \mathrm{e}^{\ln 2 g s t} + 1} \tag{6.13}$$

where $R_0 = N_0^* / N_0$ is the initial ratio of mutants, and $E_r = E^* / E$ the ratio of the expression levels of the mutant and the wild-type. This expression assumes that a mutant arises close to the beginning of the experiment and further does not address consecutive mutations.

Enzyme dilution

Important for both the correct determination of expression levels, as well as important to take into account when setting up an experiment with alternating medium conditions, is the fact that we observed a 'superinduced' LacZ activity for cells after spending a stationary phase at high expression levels. We found that the expression levels of induced cells as determined immediately after they leave stationary phase, can be up to a factor of 10 higher that the expression during exponential growth. If this happens, it can take very long before LacZ molecules are diluted out by cellular division, even when their production is low. To demonstrate this effect, a culture of wild-type cells was grown overnight at full induction (200 μM IPTG). The next morning the culture was washed and grown in fresh medium without IPTG. At specific time points samples were taken and frozen at -80^{o}C. Afterwards the expression levels for these samples was determined (figure 6.13). In the figure induced and uninduced levels of expression for an exponentially growing population are given as dotted lines. We indeed observe that the cells initially have a much higher LacZ expression than exponentially growing induced cells. The expression levels decrease over time, which corresponds to the observed growth rate of the cells. Remarkably, even after 8 hours of growth the expression level of exponentially growing uninduced cells has not yet been reached.

In our determination of expression levels (see above), we have taken into account the long times it may take to be able to determine the expression levels associated with exponential growth. Importantly, for evolution experiments under alternating conditions, it is essential to take into account enzyme dilution effects on the response times of the regulatory system.

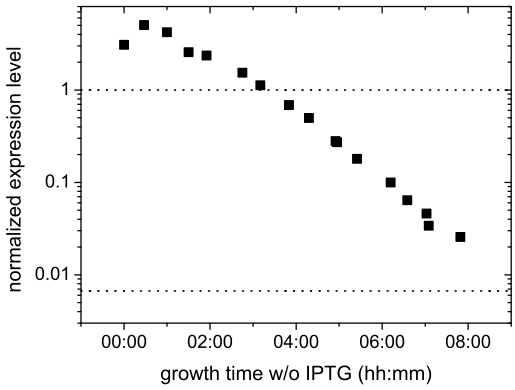


Figure 6.13: Enzyme dilution by cellular division, visualized by decreasing expression levels. At t=0 a stationary overnight culture of induced (200 μM) wild-type cells is inoculated in fresh medium. At the indicated time points, samples are taken from this culture of which the expression is determined. Dotted lines are uninduced (lower) and fully induced (upper) expression levels of an exponentially growing wild-type population. On the logarithmic vertical scale we observe a near linear decay, corresponding to exponential dilution of the enzyme. From the decay rates a slowly increasing decay rate ($1/t_{1/2}$) was determined, starting at 0.92 h^{-1} and ending at 1.3 h^{-1}. The initial rate corresponds well with the growth rate of the population at full induction, whereas the end rate is somewhat higher than that of an uninduced population, which might be caused by intrinsic degradation of LacZ.

CHAPTER

7

Residual affinity of induced repressors alters the shape of the induction curve

Why do we still study E. coli?

Jeffrey H. Miller,
Experiments in Molecular Genetics, 1972

In recent years our understanding of gene regulation and regulatory networks has greatly increased owing to the development of quantitative thermodynamic and stochastic modeling approaches. Among one of the best characterized systems is the lactose utilization operon of Escherichia coli. We argue, however, that to quantitatively account for the behavior of the system, a missing ingredient in recent extended modeling efforts is the residual DNA binding affinity of induced repressor species - although it has in fact been experimentally determined in early studies on lac regulation. We present a basic thermodynamic model that incorporates residual affinity in the lac system, and together with experimental data on overexpressed wild-type and mutant repressors, we investigate its consequences. We conclude that for the interpretation of kinetic as well as equilibrium behavior, residual affinity plays an important role. For overexpressing systems, it significantly alters the expected shape of the induction curve. If it is not taken into account, extraction of quantitative information from such measurements suffers from inconsistencies in interpretation. We expect that residual affinity plays a similar role in systems with regulatory allosteric interactions in general. Accounting for its effect should be particularly important in the field of synthetic biology and rational network design.

Although the *lac* operon in *Escherichia coli* is since its initial description [89] undoubtedly experimentally and theoretically the most well-studied example of bac-

terial regulation [93, 201, 211–213], some important aspects are still under scrutiny. Recently a focus has been on repressor mediated DNA looping [214–218], the observed cooperativity in the *lac* operon induction kinetics [207, 219] and the presence of multistability [220–222]. Quantitative thermodynamic models have been formulated that explain additional sharpness of the inducer response by incorporating a detailed model of inducer bound repressor states [207, 219], as well as by including CAP/CRP mediated DNA bending [207]. However, these models, and more in general models of transcriptional regulation responding to an intra- or extra-cellular signal, do not take into account an important ingredient: the residual DNA binding affinity of the 'inactivated' transcription factor. In the case of the *lac* operon this means that repressors that are fully saturated with inducer retain an affinity for the operator site, where they will still competitively interfere with the binding of RNA polymerase and thus inhibit transcription. This effect, although it was discussed in the earlier literature [201, 223] and binding constants were experimentally determined [224], has been neglected in the recent modeling literature. The reason for it is understandable: within the wild-type *lac* system the residual binding does indeed not seem very prominent (which we will argue has been evolutionary tuned to be so). But when the transcriptional regulators are overexpressed, as is the case for specific *lac* mutants [225] or often for artificial systems, the effects cannot be neglected. In fact, an important fraction of experimental work on the *lac* system has been performed in overexpressing systems (e.g. [216, 226, 227]). Also in studies focusing on the rational design of synthetic gene regulation networks, transcription factors are usually not operating in their 'natural' concentration ranges, which may lead to a discrepancy between conceived and actual behavior of the networks (e.g. [153, 178]). Apart from the consequences for the equilibrium description of the *lac* system, a kinetic description of the system will also suffer from not taking residual binding into account. Effectively the absence of residual binding would imply that the time constant of induction would be equal to the time constant of dissociation of the repressor-operator complex (see section 7.3), which is certainly not true for the *lac* operon.

In this work we will present a somewhat simplified, but transparent thermodynamic equilibrium model of *lac* transcription regulation that does take into account residual binding of induced repressor states. Our model is discussed in close reference to experimental data we obtained from a highly overexpressing system and a mutant system with a reduced DNA binding affinity. The consequences of residual binding are made clear by comparison to the case where residual binding is neglected. We show that overexpression and residual binding can have unexpected effects and can explain a number of apparent experimental inconsistencies and seemingly anomalous behaviors, from literature data, as well as from our own data. We argue that, in order to meaningfully extract thermodynamic information from experiments on transcription regulation, it should be assessed whether residual binding plays a role in that particular system, and if so, its effect should be taken into account. Especially in the field of synthetic biology

it is important to realize the potential effects of residual binding.

7.1 Model

The most straightforward effect of repressor overexpression is that the operon cannot be fully derepressed (see e.g. [211, 223]), which is the hall mark of residual affinity in the system. Even when the transcription factors are saturated with their inducer they still retain an affinity for their operator, and the expression from the operon is less than the there were no repressor present. However, this fact is often obscured as the experimentally obtained expression levels are generally normalized to the value at saturated induction.

Our model consists of a set of coupled reaction equations, based on relevant thermodynamic equilibria. It describes the binding and unbinding of the molecular species at the level of signal integration on the *lac* promoter and the resulting expression of the downstream genes. Expression levels are generally determined in assays of the concentration of the catabolic enzyme LacZ [154]. Since the focus is to provide an intuitive view on the effects of residual binding, we have made the following simplifications. 1) the two auxiliary operators of the *lac* system O2 and O3 (see fig. 1.4 on page 22 for an overview of the architecture of the *lac* regulatory region) are lumped in one compound operator, here termed O_2. This reduces the amount of looped complexes that have to be taken into account. This will not greatly affect the description as can be seen e.g. from ref [227] where deleting O2 only has a small effect on measured repression values. Mind that throughout this chapter subscripted operator species O_1 and O_2 refer to the model parameters. 2) we assume that a repressor will change to its induced state upon binding of one inducer. Other models have demonstrated the additional cooperativity resulting from partially induced tetrameric repressors ([207, 219] and see section 7.5). This implies that our model will not provide an *ab initio* prediction for the sharpness of the induction curve, which is therefore incorporated via a phenomenological Hill coefficient. Note that the predictions for either the zero or saturating inducer limits do not depend on the cooperativity and are therefore not influenced by assumption 2).

The expression level of a repressible promoter is determined by the competitive binding of the repressor and the RNA polymerase [228, 229]. Since binding processes are generally fast with respect to the time scale of transcription, expression levels are commonly assumed to be proportional the promoter activity P_A, which is proportional to one minus the fractional saturation of the operator by the repressor ($P_A \propto 1 - S_f$) (see [90, 230]), where S_f is defined as the averaged time fraction that a binding site is occupied. However, the *lac* system is also activated by the CAP/CRP protein, the action of which is modulated by catabolite repression [231]. Interestingly in the *lac* system one of the operators partially overlaps with the binding site for the CAP/CRP protein [232, 233], so that the level of activation is also determined by a competitive binding mechanism. As we will see, this removes the proportionality of the expression

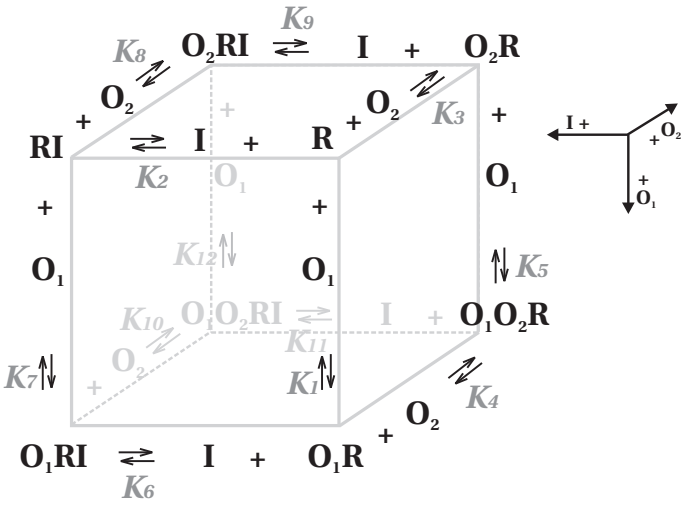


Figure 7.1: Cube of reaction equations used to describe the *lac* system. *K*'s are equilibrium dissociation constants, *R* is repressor, *I* is the intracellular inducer concentration. O_1 and O_2 denote the main and auxiliary operator, where O_2 represents a lumped variable for the *lac* O2 and O3 operators (see fig. 1.4). Downward reactions denote repressor species binding to O_1, leftward represents binding to I, and backward reactions represent binding to O_2.

to the fractional saturation S_f of the main operator.

The concentration of the product of downstream gene *lacZ*, or expression is given by

$$E = \frac{AP_A\alpha_o}{\gamma} \tag{7.1}$$

where A is the factor of activation, α_o is the un-activated, un-repressed enzyme production rate and γ is the dilution rate, resulting from cellular division and protein degradation. These elements are incorporated into the model by coupling the relevant reaction equations (figure 7.1). The species O_1O_2R and O_1O_2RI denote the looped complexes of main operator, auxiliary operator and repressor with or without inducer bound.

The functional form of the promoter activity is given by

$$P_A = O_1 \cdot O_2 + fO_1 \cdot O_2R + fO_1 \cdot O_2RI \tag{7.2}$$

where O_1 and O_2 denote the average time fraction that the main and auxiliary operators are free, and O_2R and O_2RI denote the fraction of time that the auxiliary operator is occupied by respectively free repressor R and induced repressor RI. A dot denotes a product of these fractions. f is a factor (its value being between 0 and 1) to account for the competitive binding of repressor species and CAP/CRP to the auxiliary operator, given by $f = 1/A$. In this way the promoter activity is 1 if both main and auxiliary operator are unbound ($O_1 = O_2 = 1$, other terms 0), the activity is f if the main operator

is free while the auxiliary operator is bound ($O_1 = 1$, and $O_2R + O_2RI = 1$), and 0 if the main operator is bound ($O_1 = 0$).

In order to find an expression for P_A in terms of equilibrium dissociation constants, and repressor and inducer concentrations, we should state the equilibrium relations $\frac{A\,B}{AB} = K_{\mathrm{AB}}$ for the different binding events. However, in order not to overspecify the system, one should properly take into account the constraints of the system: stating three equilibrium relations on a face of the cube in figure 7.1, determines the fourth one because of detailed balance. However, the relations on the faces are in their turn constrained since they are part of a cube, and one constraint would be doubly counted. In total we are left with 6 - 1 = 5 constraints, and 12 equilibrium constants. We thus fully determine the system by stating 7 equilibrium relations.

$$\frac{O_1\ R}{O_1R} = K_1 \qquad \frac{I\ \ R}{IR} = K_2 \qquad \frac{O_2\ R}{O_2R} = K_3$$
$$\frac{O_1R\ \ O_2}{O_1O_2R} = K_4 \qquad \frac{O_1\ \ RI}{O_1RI} = K_7 \qquad \frac{O_2\ \ RI}{O_2RI} = K_8 \qquad \frac{O_1RI\ \ O_2}{O_1O_2RI} = K_{10} \tag{7.3}$$

which are chosen on the basis of either their experimental accessibility, or their usefulness in considering limit cases of the system (see section 7.3).

Next the conservation relations for operator and repressor are specified

$$\begin{aligned} R + RI &= R_{\mathrm{tot}} \\ O_1 + O_1R + O_1RI + O_1O_2R + O_1O_2RI &= O_{1\mathrm{tot}} \\ O_2 + O_2R + O_2RI + O_1O_2R + O_1O_2RI &= O_{2\mathrm{tot}} \\ O_{1\mathrm{tot}} = O_{2\mathrm{tot}} &= 1 \end{aligned} \tag{7.4}$$

Note that repressor bound operator states do not contribute to the conservation relation of repressors, since they are negligible compared to the total repressor concentration. Also there is no conservation relation for inducer, as the inducer (here the gratuitous inducer isopropyl-β-D-thiogalactopyranoside (IPTG)) easily crosses the cellular membrane (e.g. [234]) so that the intracellular inducer concentration is buffered by the extracellular reservoir. Note further that I can be considered equal to the extracellular concentration, only in the case if membrane transport is by diffusion, hence in strains deleted for the permease LacY.

Given these relations, we can analytically solve for P_A (given by equation (7.2)), which yields

$$P_A = \frac{4p\,(2 + y - 2\sqrt{1+y})}{y^2} \tag{7.5}$$

where

$$p = \frac{K_1K_7(I+K_2)(K_2K_8(K_3 + fR_{\mathrm{tot}}) + IK_3(K_8 + fR_{\mathrm{tot}}))}{(K_2K_7(K_1 + R_{\mathrm{tot}}) + IK_1(K_7 + R_{\mathrm{tot}}))(K_2K_8(K_3 + R_{\mathrm{tot}}) + IK_3(K_8 + R_{\mathrm{tot}}))} \tag{7.6}$$

$$y = \frac{4K_3K_8R_{\text{tot}}(I+K_2)(IK_1K_4+K_2K_7K_{10})}{K_4K_{10}(K_2K_7(K_1+R_{\text{tot}})+IK_1(K_7+R_{\text{tot}}))(K_2K_8(K_3+R_{\text{tot}})+IK_3(K_8+R_{\text{tot}}))} \tag{7.7}$$

For small y, equation (7.5) simplifies to p; when y is large, this reads $\frac{4p}{y}$. In fact, p is the full description for the case when no tetramerization occurs, and only dimers can bind (see section 7.3).

In order to compare the case without residual binding, we consider the limit of this model where none of the inducer-bound repressor species is able to bind (K_7, K_8, and K_{10} (and hence K_{12}) to infinity). This changes the functional forms of p and y, which are given in section 7.3. Other limits and their consequences are discussed in relation to experimental data in section 7.3.

7.2 Results and Discussion

We constructed a plasmid that overexpresses wild-type LacI by a factor of $5{\cdot}10^3$ (see section 7.4), thus yielding around $5{\cdot}10^4$ tetrameric *lac* repressors per cell. A reporter plasmid containing the wild-type *lac* promoter controlling *lacZ* expression was used to measure the induction curve as a function of IPTG concentration. The induction curve is given in figure 7.2a, square symbols. The expression values are normalized here with respect to the maximum operon expression level, which we determined experimentally in the absence of a repressor.

Indeed, it can be seen that the overexpressed system cannot be induced to the maximum expression levels: induced expression is a factor of ~100 lower than the maximum. Moreover, the IPTG concentration, $I_{1/2}$, at which the expression is halfway between the repressed and induced values (plotted in log-log), which for wild-type repressor levels is ~20 μM (see [207] for the Δ*lacY* strain TK150), has shifted to higher values: 140 μM. This is a strikingly large shift: whereas at wild-type repressor levels (in a Δ*lacY* strain) full induction is reached at around 200 μM IPTG, in the overexpression strain expression is only ~1/30 of the induced level, and full induction is only reached for IPTG concentrations larger than 2 mM.

This effect is especially important to account for when using transcription factors in synthetic networks. Functioning of a designed network that contains transcription factors that either operate outside their natural concentration ranges, or where their concentration is regulated itself, often behave quite differently than expected (see e.g. [153, 178]). An interesting situation occurs when $I_{1/2}$ shifts upwards, while adding a higher concentration of inducer would be toxic to the cell. This was for example the case for TetR induction by doxycycline in the initially constructed repressor cascade network in chapter 4. Practically, this causes the repressor to be uninducible. When we evolved the cascade network under a selective pressure to be inducible by non-toxic levels of doxycycline ($\lesssim$60 ng/ml), we mainly recovered *tetR* mutations which presumably destabilized the *tet* repressor and decreased its capacity to repress. Indeed, also when we randomly mutate the *lac* repressor to an extent that its repression is decreased

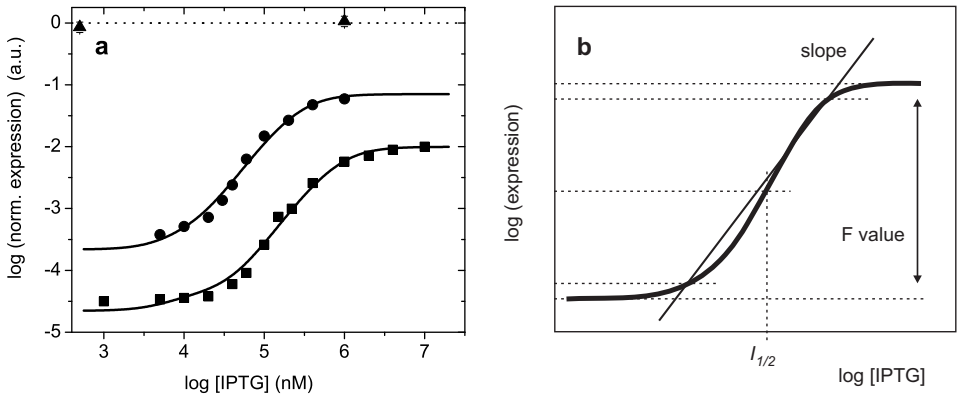


Figure 7.2: (a) Expression data of a LacI overexpressing operon as a function of IPTG. Square symbols are data for overexpressed wild-type sequence LacI, round symbols for a mutant LacI. Curves are generated by the model described in the text. Triangles indicate maximum operon expression levels, measured in the absence of repressor (leftmost triangle indicates level without inducer). This expression level serves as normalization. (b) Definition of the curve shape parameters. The F value is the ratio of fully induced over uninduced expression. The slope is here defined by the line that intersects the induction curve at 5% and 95% of the expression range, plotted logarithmically. $I_{1/2}$ is the inducer concentration where the expression is halfway its dynamic range, again plotted logarithmically.

(figure 7.2, round symbols) the value for $I_{1/2}$ is shifted back from 140 μM to 58 μM.

The shape of the curve could be reproduced by equation (7.5), with the model parameters as given in section 7.6. The mutant data was fitted with the same parameters, where the lower affinity of the mutant repressor-operator interaction was captured by a single 'mutation parameter'. With the obtained parameters we use the model to investigate the effect of repressor overexpression quantitatively. We will focus on changes in parameters describing the shape of the curves, as indicated in figure 7.2b. These are besides $I_{1/2}$: the F value or regulation factor F, being the induced expression divided by the uninduced expression, and the slope, defined here as 0.9 F divided by the inducer concentration interval between 5% and 95% of induction. We used this practical definition for the slope, and not the maximum derivative since the model allows for non-sigmoidal curve shapes where the maximum derivative would not be a good measure for the concentration of half-induction.

The shapes of the induction curves for repressor concentrations between 1 nM and 1 M are given in figure 7.3a, where 1 nM corresponds to about 1 repressor per cell[1]. Figures 7.3b-d demonstrate the changes in the induction curve shape parameters (solid lines) as a function of repressor concentration, in comparison to the situation without

[1]Note that the full concentration range might not be physiological for *E. coli*, as extreme overexpression ($R_{\mathrm{tot}} \gtrsim 0.1$ mM) will interfere with cellular growth.

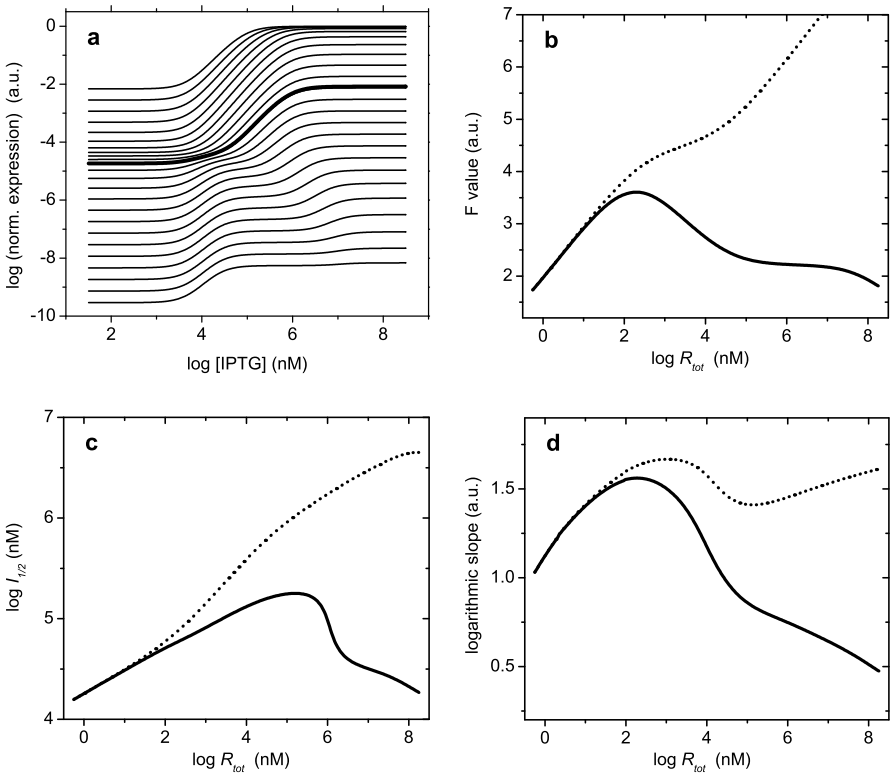


Figure 7.3: Influence of overexpression and residual affinity on the shape of the induction curve. (a) Changes in the curve shape as a function of total repressor concentration, ranging from 1 nM to 1 M for the model including residual affinity. Other model parameters are as used to describe the induction profile of wt-sequence LacI (curve through the square symbols in figure 7.2 and which is here shown in bold). (b, c, d) Influence of LacI overexpression on the curve shape parameters defined in figure (7.2b). Solid curves include residual affinity, dotted curves do not.

residual affinity (dotted lines). From the figures we see that for wild-type expression levels (10-20 repressors per cell [235], or log R_{tot} ~1 - 1.3) the effects of residual affinity are minimal: the solid lines and the dotted lines do not differ much. An intuitive evolutionary argument can be given why this is the case. In fact the optimal operational concentration for a repressor is subject to two opposite selection pressures: to maximize the F value and to minimize the cost of repressor protein expression. The requirement of a large F value sets a minimum to the repressor concentration. At the same time, a higher repressor concentration would tune down the overall operon expression, but this would be a very inefficient way to achieve a lower expression: downtuning the promoter strength would achieve the same, but would impose a lower cost associated

with repressor protein production and therefore be more advantageous in evolutionary terms. A more quantitative derivation can be given (see section 7.3) why the optimum repressor concentration should actually be roughly *equal* to the dissociation constant specifying the residual affinity.

For mildly overexpressing systems (~10x wild-type levels, $\log R_{\mathrm{tot}} \approx 2 - 2.5$), figure 7.3b shows that the regulation factor F increases, which corresponds well with the observation in ref. [226] that F values increased from 1300 to 6700 for 5x overexpression. However for stronger overexpression F decreases again, as is observed in figure 7.2. Note that not taking into account residual binding would imply that the fully induced expression level is the maximum expression level (as if no repressor were present), and the F value would grow without bound (dotted lines). The difference between the solid and the dotted curve in figure 7.2 therefore also directly demonstrates the mismatch between determining the F value either by induction, or by absence (or knock-out) of the repressor (see also section 7.3).

Interestingly, the shift in $I_{1/2}$ combined to the limited increase in F value for overexpressing systems, may influence the quantitative estimation of parameters from an induction curve. If the expression level at saturated induction is interpreted as the maximum expression level for the operon (no repressor present), a model reproducing the experimental data may largely overestimate the dissociation constant for inducer binding, K_2, due to the shift in $I_{1/2}$. Although K_2 has been determined experimentally [224] to be ~1.0 μM, in the modeling literature generally much higher values are used (15-30 μM). This, in turn, may lead to wrong predictions on the behavior of the regulatory system, especially when integrated in a larger network.

Figure 7.3d shows the change in the steepness of the induction curve as a function of repressor concentration. As remarked earlier, based on the model in section 7.1, we cannot make precise statements about the values for the slopes of the curve. From our model we see that for (mild) overexpression it increases, which has also been remarked in [219]. However, comparing the case with and without residual affinity, we observe that a maximum value for the slope is reached at repressor concentrations that are lower by a factor of about 10. Moreover, the slope decreases quickly at higher overexpression, to a lower value than for wild-type repressor levels. This is qualitatively different when there is no residual affinity: the slope of the curve is then predicted to stay high. Again, as the sharpness of response is an important quantity in a signal transducing system (e.g. [236–238]), care should be taken in its description.

In this work we have discussed the effects of transcription factor overexpression using a model that explicitly incorporates the residual affinity of induced states. We argued that a description of transcriptional regulation is not complete and often will not be accurate without giving attention to residual affinity. Although we have focused mainly on the *lac* transcriptional regulation, we expect that similar effects occur for other systems with an allosteric transition between an 'active' and an 'inactive' state [239–241]. Especially when acting in a network where the concentration of allosteric

regulators itself is regulated, operational concentration effects can be expected to play a role. For a realistic quantitative description of network behavior (e.g. [242]) and their stable states, and especially for rational design of networks in synthetic biology [243], explicitly considering these effects is important.

7.3 Useful limits and comparison to data

In this section a number of limits are worked out that demonstrate in an intuitive way how changes in the system affect the shape of the induction curve and the F values. An emphasis is placed on the role and consequences of residual affinity. Outcomes are discussed in the light of measured data, and some interpretation issues in existing literature are clarified.

No residual affinity.
The case of no residual binding is obtained by taking the limits of large K_7, K_8, and K_{10} of equation (7.5). The functional form of P_A remains the same, but the terms p and y now read

$$p = \frac{K_1(I+K_2)(IK_3+K_2(K_3+fR_{\text{tot}}))}{(IK_1+K_2(K_1+R_{\text{tot}}))(IK_3+K_2(K_3+R_{\text{tot}}))} \tag{7.8}$$

and

$$y = \frac{4K_2K_3R_{\text{tot}}(I+K_2)}{K_4(IK_1+K_2(K_1+R_{\text{tot}}))(IK_3+K_2(K_3+R_{\text{tot}}))} \tag{7.9}$$

which are the relations used in figure 7.3 (dotted curves) where the effects of having residual binding are demonstrated.

Only one operator, O1.
If the auxiliary operators are knocked out or removed the F value is greatly reduced (see e.g. [226, 244]). We can find this limit by solving for the front face of the cube, or, equivalently take the limit to infinity for K_3, K_4, K_8, and K_{10}.
This results in

$$P_A = \frac{K_1K_7(I+K_2)}{K_2K_7(K_1+R_{\text{tot}})+IK_1(K_7+R_{\text{tot}})} \tag{7.10}$$

The F value for the single operator system is given by

$$F = \frac{K_7(K_1+R_{\text{tot}})}{K_1(K_7+R_{\text{tot}})} \tag{7.11}$$

from which we can see that the maximum F value is given by $\frac{K_7}{K_1}$ (being the ratio of dissociation constants for induced and uninduced repressor to the main operator), for sufficiently large R_{tot}. Interestingly, for wild-type LacI concentrations R_{tot} is not larger

than K_7, which means that the full dynamic range is only observed in overexpression. This is indeed the case. The group of Müller-Hill measured an F value of 18 in the wt i^+ background, but a factor of 60 in a (5x) LacI overexpressing system [226].

Only one operator, O1 & no residual binding.
As seen above, one of the most notable effects of residual binding of induced repressors is that an operon cannot be fully derepressed by induction. This implies that in experiments that use non-functional repressors to determine the maximum operon expression, the *apparent* F value of an operon grows unlimitedly for increasing repressor concentrations. This is a main source of confusion in the interpretation of measurements and the extraction of thermodynamic quantities from them. That the effects can be large can be seen from e.g. [227], where a repression value of 230 is reported for a 5x overexpressed (dimeric) repressor, determined by comparison to non-functional repressor. In [226] the same quantity is determined by induction, and yields a factor of 60. These fairly large differences are often interpreted as experimental noise, but are perfectly explained by the presence of residual binding. The combined case (only O1 and no residual affinity) discussed here can be obtained by setting K_3, K_4, K_7, K_8, and K_{10} to infinity. We obtain

$$P_A = \frac{K_1(I + K_2)}{IK_1 + K_2(K_1 + R_{\text{tot}})} \tag{7.12}$$

which is the same form as the first mechanistic model of *lac* induction kinetics in the seminal paper by Yagil and Yagil [91]. The equation yields a regulation factor of $F = 1 + \frac{R_{\text{tot}}}{K_1}$. This is also the apparent regulation factor when using a non-functional mutant in the experimental determination of the F value. In contrast, determining the F value using induction yields equation (7.11). Indeed for overexpressing systems a discrepancy between the two methods is expected, where for sufficiently high repressor concentrations the induction method will result in a lower F value, which is independent of the degree of overexpression.

No tetramerization / repressor acts as dimer.
Repressor mutants exist that do not form tetramers, but act as dimer (e.g. the *lacI*^{adi} mutant [226, 245]). Such mutant repressors are not able to bind two operators simultaneously and therefore not able to form loop the *lac* promoter DNA. Their observed repression values are lower than those of tetrameric LacI, when the repressor concentration is comparable to wild-type, but when the repressor is overexpressed the repression values for dimers and tetramers become similar (see e.g. [226, 227]). From our model the expression for dimers is obtained by taking the limit of P_A given by equation (7.5) for large K_4 and K_{10}. P_A then reduces to p, as given by equation (7.6). From this we can see that indeed for high repression values there is no distinction between having dimers or tetramers, since the limit of equation (7.5) for large repressor concentrations is equally described by p. This can be understood by realizing that for large tetrameric

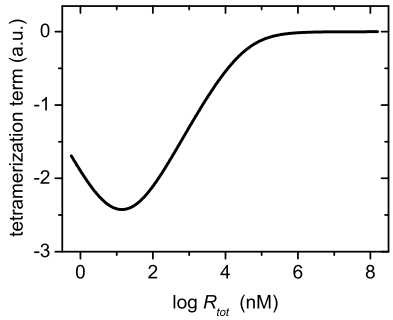


Figure 7.4: Logarithm of the factor given by equation (7.13), specifying the ratio of uninduced expression levels accomplished by tetramers and by dimers as a function of the total repressor concentration. When the ratio is 1 (here its log being 0), there is no difference in repression between tetramers and dimers. Used parameters are given in section 7.6.

repressor concentrations each operator is occupied by a separate *lac* repressor, as is necessarily the case for dimers. In fact, the factor

$$\frac{4\,(2+y-2\sqrt{1+y})}{y^2} \tag{7.13}$$

from equation (7.5) is the change of the functional form of the induction curve due to the ability of tetramers to bind to two operators and loop the DNA.

From figure 7.4 we can see that the factor in equation (7.13) indeed reaches 1 for high overexpression of repressor.

The F value in the no-tetramerization case is given by

$$F = \frac{K_7(K_1+R_{\rm tot})(K_3+R_{\rm tot})(K_8+fR_{\rm tot})}{K_1(K_7+R_{\rm tot})(K_3+fR_{\rm tot})(K_8+R_{\rm tot})} \tag{7.14}$$

which for large concentrations of (either dimeric or tetrameric) repressor, reduces to $F = K_7/K_1$. Moreover, since for large $R_{\rm tot}$ the functional form of the induction curve is a sigmoidal where $I_{1/2}$ coincides with the inflexion point of the curve, we can find the location of $I_{1/2}$ by solving $\frac{d^2}{dI^2}P_A = 0$, which yields $I_{1/2} = K_2\sqrt{K_7/K_1}$.

O1 knockout, only auxiliary operators.
Several effects of the auxiliary operators have been proposed. First, the presence of the auxiliary operators would result in an increase in the local repressor concentration around O1 [246]. Second, the induction kinetics would become more cooperative because of DNA looping, potentially aided by CAP/CRP binding [207,219,227,244]. Thirdly, since the operator O3 and the CAP/CRP activator binding site partially overlap,

binding of repressor to O3 sterically hinders CAP/CRP activation, and thus lowers expression [227,233]. The observation that when only O3 is functional, there is still a large difference in repression values between smaller dimers and bulkier tetramers [227] supports the third explanation. In our model the effect of DNA looping is incorporated by the reaction constants. The effect of CAP/CRP interference is incorporated via the factor f, where CAP/CRP activation is a factor of $1/f$ (see the next sub-section for its value). The case of an inactivated operator O1 can be obtained in our model by taking the limit of K_1 and K_7 to infinity, from which we obtain an F value of

$$F = \frac{(K_3 + R_{\text{tot}})(K_8 + fR_{\text{tot}})}{(K_3 + fR_{\text{tot}})(K_8 + R_{\text{tot}})} \tag{7.15}$$

As $K_8 \gg K_3$ and for wild-type $K_8 \gg R_{\text{tot}}$, the regulation factor becomes:

$$F = \frac{K_3 + R_{\text{tot}}}{K_3 + fR_{\text{tot}}} \tag{7.16}$$

which is 1 for $f = 1$ (no interference with CAP/CRP), and $1 + \frac{R_{\text{tot}}}{K_3}$ for $f = 0$ (realistic values are closer to 0, see below). Assuming that K_3 is of the order of 10 nM (roughly a factor 5 less than K_1, see above), for wild-type repressor concentration the repression value will be of the order of 2. The measured value in [226] is 1.9.

Induction and CAP/CRP activation.
The most obvious effect in a CAP/CRP deletion strain is the lowered expression of *lac* operon expression due to a lack of activation. A 50 fold difference has been found between the expression of a wild-type and CAP/CRP deletion strains [247], and a factor of 28 by maximum repression of a mutant containing only operator O3 [227]. This effect is incorporated in our model, since the *lac* operon expression is proportional to P_A/f (section 7.1). By changing f, we therefore do not only change the functional form of P_A, but also the overall expression level. Our measurements are performed by growth in 0.2% glucose medium, which due to catabolite repression will reduce CAP/CRP activation ~3 fold [207,248]. As the best-fit value for f in our model is around 0.1, we would interpolate the full CAP/CRP activation factor to be on the order of $1/(0.1/3) = 30$, close to the experimental value of 28 [227], or 50 in [247]. Interestingly, the mere factor of 3 that is due to catabolite repression may not be the main function of CAP/CRP signal integration in the *lac* operon, nor the overall activation factor of ~30 (which could equally be accomplished by increasing the promotor strength). An alternative suggestion would be that the main role of CAP/CRP activation is steepening and increasing the dynamical range of the response to lactose, as a result of the competitive binding with the *lac* repressor. Such a general activator (CAP/CRP at the same time interacts with many other operons) could well be an efficient general mechanism for simultaneously obtaining sharp response curves for a large number of operons.

Residual affinity and the kinetics of induction.
Not only an equilibrium but also a kinetic description of the system will suffer from neglecting residual affinity, and even at wild-type repressor levels. Its absence would imply that the time constant of induction would be equal to the time constant of dissociation of the *OR* complex. This can be made intuitive by looking at front face of figure 7.1. When the operon is in the repressed state, upon addition of inducer, I will bind to the free repressors R, which effectively lowers the concentration of repressor. However, at saturating concentrations of inducer, the induction *kinetics* are determined by the rate of dissociation of the *OR* or *ORI* complex. The assumption that the repressor-inducer complex has no affinity for the operator would imply that K_7 is very large (infinite). However, by detailed balance, this would mean that the dissociation constant K_6 is also very large, implying that inducer does not bind *OR*. This would mean that after addition of inducer, the time in which protein synthesis rate reaches its maximum is at least equal to the dissociation time of the *OR* complex (20-30 minutes [249, 250]), which is clearly incorrect when compared to experimental values. For example the operator clearance rate (however for dimeric *lac* repressor) was determined at less than 1 s^{-1} in [234]. If we assume that inducer binding does not affect the association but only the dissociation rate of repressor to operator, our estimate for operator clearance rate would be the reciprocal of ~25 min divided by the F value, which would indeed be on the order of 1300 / 1500s = 0.9 s^{-1}.

Evolutionary argument on wild-type repressor levels.
The argument why the optimal concentration of repressors should lie around the equilibrium dissociation constant expressing the residual affinity is most clearly made by reference to a simple system containing one operator (see case above 'Only one operator, O1'. We suggest that the optimum repressor concentration is a result of two counteracting selective pressures. First, the repressor concentration should be high enough to be able to accommodate the maximum F value in the system (if a lower F value would suffice for operation, there would be no reason to have an in-built, but never used, capacity for higher F values). Earlier it was already shown that equation (7.11) for one operator leads to a maximum F value for repressor concentrations larger than K_7, the equilibrium dissociation constant that determines the residual affinity of induced repressors. However, as the fully induced promoter activity for one operator is given by

$$P_A = \frac{K_7}{K_7 + R_{\text{tot}}} \tag{7.17}$$

we see that for repressor concentrations larger than K_7 the operon cannot be fully derepressed. We argue that this is an evolutionary unfavorable situation, since if a lowered expression would be more favorable, it could be accomplished by down tuning the promoter strength. This would achieve the same regulation, but with a lower cost, since no extra repressor protein need to be produced. These arguments taken together, we expect that the operating concentration of the repressor will be of the order of K_7. Indeed,

when the *lac* operon is described in a simplified way as a one-operator system (as was the case before the discovery of the auxiliary operators), the measured residual affinity is on the order of 10^{-8} M (see [201] and section 7.6 below). This number amounts to around 10 repressors per cell which also is the consensus value [93, 235]. It would be interesting to see whether the same holds for other repressible operons.

7.4 Material and Methods

Induction experiments are performed using *Escherichia coli* K12 strain MC1061 [184], which carries a deletion of the complete *lac* operon. This strain was obtained from Avidity LLC, Denver CO, USA, as electrocompetent strain EVB100 (containing an additional chromosomal *birA* gene). We constructed a LacI overexpressing plasmid, based on the pZ vector system [186], in which LacI is expressed from the $P_L$tetO1 promoter in a medium copy plasmid (p15A ori). This plasmid yields an estimated constitutive expression of $\sim 5 \cdot 10^4$ *lac* tetramers per cell (by comparison of the PN25 promoter expression as stated in [186], and the comparable promoter strength of P_L [251]). No influence of this extent of overexpression on the bacterial growth could be established within an error of roughly 5% (for methods of growth measurements see chapter 4). Repression strengths were determined by either ONPG or FDG expression assays (see chapter 4), using a pTrc99A [187] based plasmid where *lacI* and Ptrc were replaced by the Plac-*lacZ* fragment from strain MG1655 [188]. All growth and expression measurements are performed in Defined Rich medium (Teknova, Hollister, CA, USA, cat. nr. M2105), with 0.2% glucose as carbon source, and supplemented with 1 mM thiamine HCl. Mutants were created in a mutagenic polymerase chain reaction using the Stratagene Genemorph II Random Mutagenesis kit.

7.5 Appendix A: not modeled induced repressor states

For simplicity, and given the scope of the present work, induction was modeled as if the binding of one inducer molecule would bring the repressor in the induced state. Although this might be the case for some repressors, LacI is a dimer of dimers [252], which causes the induced repressor states to be more complex (see figure 7.5).

Both recent studies on the origin of the cooperativity in the *lac* system [207, 219] include the possibility of partial induction. Without inducer, both dimers can bind and the repressor can form loops (repressor in the black dotted box). At intermediate inducer levels one dimer can be induced, while the other is still active (grey boxes). However, both studies assume that the 'unboxed' states abolish all operator infinity, i.e. do not take into account residual affinity of the fully induced states.

The reaction equation model described in the present work does not take into account multiple binding events of the inducer and had therefore to be modified with a phenomenological Hill coefficient to account for increased steepness of the response

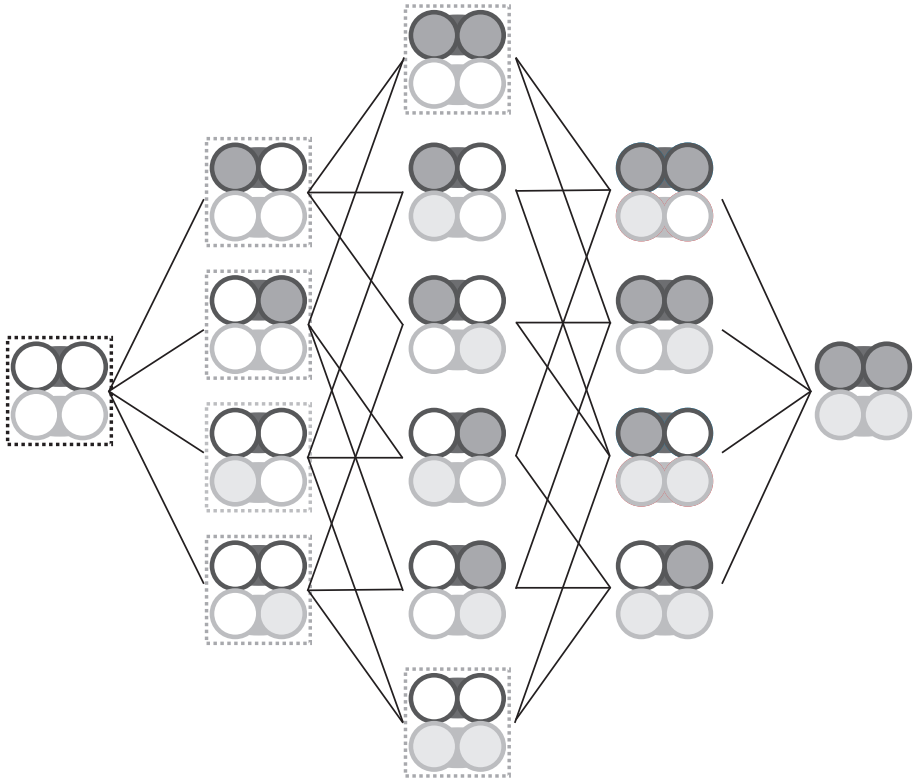


Figure 7.5: Partially induced repressor states. As the *lac* repressor is a dimer of dimers, in which each monomer has a binding site for an inducer molecule, there are various (partially) induced repressor states. Different dimers are depicted as dark grey and light grey connected circles. Filled circles represent monomers with an inducer molecule in their binding pocket. When an inducer molecule is bound to (at least) one monomer, the corresponding dimer is induced. Possible transitions between repressor states are indicated by a black line. Two recent *lac* cooperativity studies [207, 219] incorporate this scheme, but assume only the boxed states to be able to bind to DNA. Figure after ref. [207].

due to the inducer cooperativity. Importantly however, while the model might not provide an *ab initio* prediction for the steepness of the response, the limits for low and high inducer levels are independent of cooperative behavior and therefore the same as when partial induction would have been included. A full model including both partial induction states and residual binding is best approached with a partition sum description and is left for future work.

7.6 Appendix B: reaction constants

Here we give the values of equilibrium dissociation constants used in the model description of the data given in figure 7.2a. For comparison specific experimentally obtained constants are stated.

To model the mutant data in figure 7.2a, we used the same data, together with a 'mutation parameter' that raises all K's describing protein-DNA interactions (all but K_2) with the same factor. The obtained factor was 5.9.

constant	model value[a]	lit. value[b]	refs.
K_1	6.3 nM	1 nM	[257]
K_2	4.9 μM	1 μM	[224]
K_3	31 nM	10 nM	[226]
K_4	2 pM	N/A	–
K_7	120 nM	60 nM	[226]
K_8	82 mM[c]	N/A	–
K_{10}	3.5 μM	N/A	–
R_{tot}	14 μM	N/A	–
f	0.116	~0.1	[207, 227]
Hill c.	1.9	N/A	–

[a] All literature values specifying DNA-protein interactions are corrected for aspecific interactions [253–255], which cause that most of the repressors are aspecifically bound. Only a fraction of around 0.01 is free in the cytosol [223, 256]. This could be incorporated by either scaling the total repressor concentration or the equilibrium constants by a factor of roughly 100. We opted for the last. Note that this introduces some uncertainty with respect to what is listed as the literature values. This would introduce a (small) scaling factor between the model parameters and the listed literature values, which indeed seems the case.

[b] Values for compounded operators are estimated on the basis of the induction characteristics in the cited references.

[c] This value implies that the affinity is of the order of the affinity for aspecific DNA, being 100 μM- 10 mM [201] in the scaled units used here.

CHAPTER

8 Reciprocal sign epistasis and multiple peaks in the fitness landscape

The whole organism is so tied together [...] that when slight variations in one part occur, and are accumulated through natural selection, other parts become modified. This is a very important subject, most imperfectly understood.

Charles Darwin,
The Origin of Species

Epistasis refers to the situation when the effect of a mutation depends on the genetic background in which it occurs. In the context of fitness landscapes it has been shown that the presence of epistatic interactions affects the selective accessibility of evolutionary trajectories. Here we focus on 'reciprocal sign epistasis' that occurs when mutating one locus is either deleterious or advantageous depending on the state of another locus, but also vice versa for the second locus with respect to the first. We show that it is a necessary condition for the occurrence of multiple peaks in the adaptive landscape. This implies that for any landscape harboring multiple peaks, there is at least one reciprocal sign epistasis motif at the level of the most elementary mutations. We also briefly indicate why a reciprocal sign epistasis motif is not a sufficient condition for the existence of multiple peaks.

Organisms are highly integrated functional systems, where change in one component often affects the functioning of other components or the system as a whole. That this interdependence of parts presents challenges for evolutionary processes was already remarked by Darwin [101]. Manifestations of these dependencies at the genetic and physiological level are pleiotropy, where a single gene or mutation has an effect on

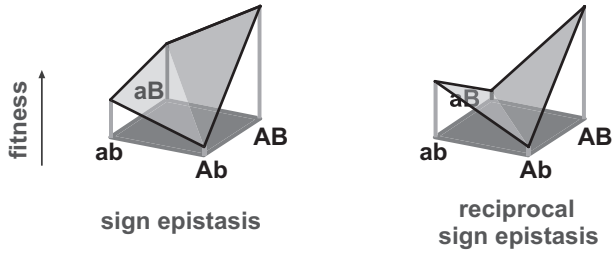


Figure 8.1: Different manifestations of epistasis along a path from a suboptimal allele *ab* towards the optimal *AB*. Left: Sign epistasis: the fitness effect of a mutation from *a* to *A* differs in sign (is either beneficial or deleterious) depending on whether the other locus is *b* or *B*. Right: Reciprocal sign epistasis: sign epistasis occurs for both loci A/a and B/b reciprocally.

multiple phenotypic traits, and epistasis, defined as the interaction between genes or mutations in their effect on the phenotype. In the latter case the effect of a mutation depends on the genetic background in which it occurs. The basis for the conditionality of epistatic effects often lies in the physical interactions between the gene products and other biomolecules, but it can for example also result from the interplay between protein stability and catalytic activity [104], from Watson-Crick base pairing within an RNA molecule [258], or from interactions between modular metabolic networks [259]. In general such dependencies between parts may constrain the evolutionary optimization of a biological system. An intuitive picture here is that of a key and a lock: changes in multiple parts are needed to obtain a new functional combination, while intermediates are non-functional. In an evolutionary process this implies long 'waiting times' before the relevant new genetic combinations appear in a population (see e.g. [260]).

When genotype, phenotype, and organismal fitness are considered from a fitness landscape perspective, epistatic interactions have a marked impact on the topology of the surface. As was shown by Kauffman in the mathematical framework of his NK landscape models [52]: the more epistasis, the more rugged the fitness landscape, and the more local adaptive peaks arise, which increases the chance of entrapment on a suboptimal adaptive solution. The occurrence of so-called 'sign epistasis' in the mapping from genotype to fitness (fig. 8.1, left), which means that the fitness effect of a certain mutation can be both advantageous and deleterious depending on the genetic background, was shown to be a necessary and sufficient condition for the selective inaccessibility of a subset of the evolutionary trajectories towards a fitness optimum [54]. If *all* paths towards a landscape optimum are inaccessible, this implies that the current population is located on a local adaptive optimum. In that case there are at least two adaptive peaks in the landscape. We will show here that in this case the landscape necessarily contains an epistatic motif, that is referred to as 'reciprocal sign epistasis' (fig. 8.1, right) [55]. This motif occurs when two loci (which can be intragenic) exhibit sign epistatic interactions with respect to each other: mutation A/a is sign epistatic with re-

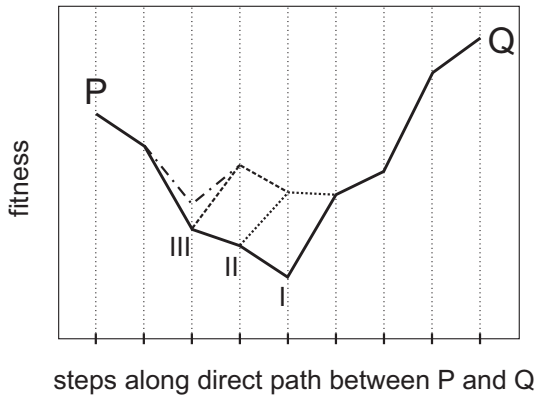


Figure 8.2: Procedure for finding the Reciprocal Sign Epistasis motif. A random direct path between peaks P and Q has a minimum at location I. By reversing the order of the mutations leading to and from this minimum, a new minimum occurs at location II. Again the order of the two mutations around this minimum are reversed. A new minimum occurs at III. When the relevant mutations are reversed here, however, the minimum does not change its location. At this location a RSE motif is found.

spect to 'background' B/b, and mutation B/b is also sign epistatic with respect to background A/a. This motif applies e.g., when two mutations are separately deleterious but jointly enhance fitness, or in the case of compensatory mutations [261].

To demonstrate that reciprocal sign epistasis (RSE) is a necessary condition for the existence of multiple peaks, we have to show that multiple peaks imply the existence of a RSE motif somewhere in the landscape. Here we give a short version of the argument, a more extended version can be found in section 8. We can look at direct paths (no detours or 'backmutations') between two maxima. If we take one of such direct paths (see solid line in fig. 8.2), we will find a combination of mutations that confers a minimum fitness (indicated with *I* in fig. 8.2), at some (Hamming) distance[1] from the initial peak P and the new peak Q. We can now try to optimize the minimum fitness of this path by reversal of the mutations leading to and from this minimum. In many cases the reversal will cause the minimum to shift to a different position between P and Q (e.g. to position *II* in fig. 8.2). We then again reverse the order of mutations to and from this minimum, and so on until the minimum remains located at the same mutational step. At that moment we will have found a RSE motif. This optimization procedure has to come to an end - because the path's minimum cannot be higher that the lowest fitness peak - and this means that we will necessarily find an RSE motif[2].

Next, we ask whether a sufficient condition could be formulated for the existence

[1] The Hamming distance d_H between two sequences is defined as the number of positions at which they differ.

[2] Finding the overall highest minimum from P to Q is a strict criterion required for the proof. In fact, we already will have found a RSE motif when for some position the minimum does not shift to another position

of multiple peaks that would *not* involve assaying the full sequence space. Although for the two locus-two allele case of fig. 8.1 RSE is a sufficient condition for multiple peaks (being *ab* and *AB*), in higher dimensions it is not. Noting that we can describe the discrete sequence domain fitness landscape as a continuous, although triangulated, multidimensional surface, we can apply Morse theory [262], linking the dimension and connectedness of the surface to the number of stationary points, like minima, maxima, simple and higher order saddles. To obtain a sufficient condition for multiple peaks, we need a lower limit on the number of maxima (M > 1), and within the context of Morse theory such a lower limit consists of a sum of the number of higher order saddles, with alternating signs. However, any sufficient condition for multiple peaks based on these topological considerations would thus involve the exclusion of saddles of a certain order, which can only be done by inspection of the complete surface or landscape. Hence using Morse theory, no practical sufficient condition can be formulated.

Recently studies have emerged that seek to assess the extent to which natural fitness landscapes are rugged and multiply peaked [25, 55–58, 61, 263–266] (and see also chapter 5). It should be noted that due to the sheer amount of combinatorial possibilities associated with the genetic sequence, it will never be possible to conclusively demonstrate that two adaptive peaks are really separated and that they are not connected by a ridge in the landscape that avoids the adaptive valley between them. However, mutational paths that accomplish such a detour may involve such long waiting times that from a practical point of view the fitness maxima are separated [263]. Here we have shown that multiple fitness peaks imply a reciprocal sign epistasis motif at the level of elementary mutations. This implies that in order to move from a lower to a higher adaptive peak, selection has to overcome typical key-lock type interactions, or at least one.

Extended proof

Theorem: In a N-allelic L locus system, *reciprocal sign epistasis* is a necessary condition for the existence of multiple peaks in the fitness landscape.

Assuming that all fitness values are nondegenerate.
First of all, we will define the terms:

Definition 1: Epistasis: means that the fitness effect of a mutation is conditional on the presence of other mutations (the 'genetic background').

Example: $\Delta w_{ab\to Ab} \neq \Delta w_{aB\to AB}$, where Δw is the fitness difference between two mutational states, and B and b can be considered to be the genetic background for states A and a.

For one mutation in different backgrounds, two classes of epistasis can be discerned:

Definition 2: Magnitude epistasis: means that the *magnitude* of the fitness effect of a mutation is conditional on the presence of other mutations.

Example:

$\Delta w_{ab\to Ab} \neq \Delta w_{aB\to AB}$ AND $|\Delta w_{ab\to Ab} + \Delta w_{aB\to AB}| = |\Delta w_{ab\to Ab}| + |\Delta w_{aB\to AB}|$

Definition 3: Sign epistasis: means that the *sign* of the fitness effect of a mutation is conditional on the presence of other mutations.

Example:

$|\Delta w_{ab\to Ab} + \Delta w_{aB\to AB}| < |\Delta w_{ab\to Ab}| + |\Delta w_{aB\to AB}|$

In case the role of mutation and background can be reversed, there is a special case of sign epistasis:

Definition 4: Reciprocal sign epistasis: means that the sign of the fitness effect of mutation $a \to A$ is conditional on whether the state of another locus is b or B, and vice versa.

Example:

$|\Delta w_{ab\to Ab} + \Delta w_{aB\to AB}| < |\Delta w_{ab\to Ab}| + |\Delta w_{aB\to AB}|$ AND

$|\Delta w_{ab\to aB} + \Delta w_{Ab\to AB}| < |\Delta w_{ab\to aB}| + |\Delta w_{Ab\to AB}|$

Now, in order to prove the Theorem, we can restrict ourself to *direct* paths between *two* peaks only (follows below).
If we can prove that for the subset consisting of direct paths reciprocal sign epistasis is a necessary condition for the existence of multiple peaks, we also have proved it for the set of all paths. In other words: if reciprocal sign epistasis is necessarily present in the subset, then also in the superset. Further, it will be shown that it is enough to prove the necessity for two peaks.

Starting with a number of 'sub-theorems' that are used for narrowing down the extent of the proof needed:

T 1: The set of direct paths ($d - d_H = 0$) of length d between two points in a N-allelic L locus system is equal to the set of direct paths of length d in the bi-allelic L locus system.

Which is a complicated way of stating something trivial. *Direct* paths between between two points in sequence space ($d - d_H = 0$) only require one substitution at each locus that is different between begin and end points.
Hence we only need to produce the proof for the bi-allelic case:

T 2: In a bi-allelic L locus system, *reciprocal sign epistasis* is a necessary condition for the existence of multiple peaks in the fitness landscape.

But we can further narrow down what we have to prove, since:

T 3: If two peaks in a bi-allelic L locus system are at a distance L, then adding any other peak will decrease the minimum distance between pairs of peaks in the system.

If the two initial peaks differed at L loci, a possible additional peak would find itself at a distance d from the first one and $L-d$ from the second one. Hence the minimum distance has decreased from L to $L-d$ or to d, whichever is smaller, but in any case $d < L$ and $L-d < L$.

As we can choose to only look at differences at the smallest amount of loci (either $L-d$ or d) between two peaks (e.g. only mutating these positions), this subspace should also contain reciprocal sign epistasis. Then, one only needs to prove the necessity of reciprocal sign epistasis for two peaks at a distance $L-d$ or d in a biallelic $L-d$ or d locus system. So finally, we are left to prove:

T 4: In a bi-allelic L locus system, *reciprocal sign epistasis* is a necessary condition for the existence of *two* peaks at a distance L in the fitness landscape.

We will now prove that two mutations along a path containing the highest fitness minimum in the landscape ($\max(\min(F_i))$, where i specifies the points in sequence space), exhibit reciprocal sign epistasis. Since all paths between two maxima necessarily go through a minimum, the landscape must contain reciprocal sign epistasis.

1) There are two maxima at distance L in a biallelic L locus system.
2) We consider a random direct path (length L) from one maximum to the other one.
3) This path necessarily contains a minimum.
4) We change the path, by reversing the order of the two mutations leading to and away from the minimum.
5) Two cases are possible:
   - 5a) If the fitness minimum of the path is still located between the same mutations, then these mutations exhibit reciprocal sign epistasis.
   - 5b) If the fitness minimum is now located at another step along the path, go back to point 4), now for the two mutations around this new minimum.
6) The repeating of steps 4) and 5) necessarily leads to breaking the loop via case 5a), because the maximization of the minimum along the path is necessarily bounded (by the value of the lowest peak).

This completes the proof of **T 4**, hence the proof of the **Theorem**. ■

Appendices

But visions are still important, as long as we also remember the details.

Stig W. Omholt,
Science 295 2220, 2002

APPENDIX

A Historical lineage of strains DH5α, DH10B, and MC1061

Here we provide a brief overview of the historical derivation of a number the *Escherichia coli* strains used throughout this thesis (parental strains are indicated below derivatives). As a large part of the work in this thesis is performed with a plasmid borne *lac* repressor, we were in need of strains containing a chromosomal deletion of the lactose operon. Contrary to common indications, DH5α and DH10B were found to contain a *lac* repressor. We traced the present confusion to a sudden omission in description of the genotype in the history of strains carrying the genotypic marker ϕ80d*lacZ*Δ(M15), which should correctly read ϕ80d*lacI*^{q}*Z*Δ(M15). The I^q designation is still present in the description of strain JM83 [267], but omitted in the papers describing the construction of the DH lineage of cloning strains [185].

DH5α
F$^-$ ϕ80d*lacI*^{q}*Z*Δ(M15) Δ(*argF-lac*)U169 *deoR recA1 endA1 hsdR17*(r_k^-, m_k^+) *supE44 phoA thi-1* λ^- *gyrA96 relA1*

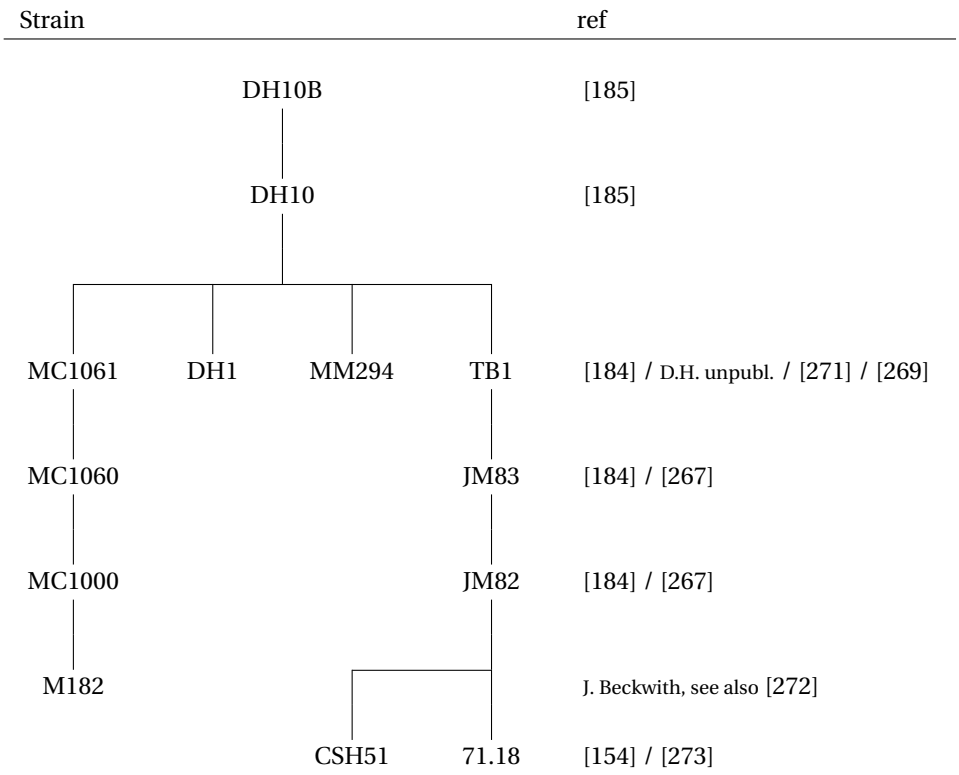
Strain	ref
DH5α	[185]
DH5	D. Hanahan unpubl. / [268] / [269]
SH210	
TB1	
DH1	D. Hanahan unpubl. / [270] / [267]
MC4100	
JM83	

DH10B

F$^-$ ϕ80d*lacI*^{q}*Z*Δ(M15) Δ*lacX74 deoR recA1 endA1 mcrA* Δ(*mrr hsdRMS mcrBC*) *nupG rpsL*(Str^R) *galU galK* Δ(*ara, leu*)7697 *ara*Δ139 λ$^-$

MC1061

F$^-$ Δ*lacX74 mcrB1 e14*$^-$(*mcrA0*) *rpsL150*(Str^R) *galE15 galK16* Δ(*ara, leu*)7697 *ara*Δ139 λ$^-$ *hsdR2*(r_k^-, m_k^+) *spoT1*



Here 'D.H. unpubl.' refers to unpublished data by D. Hanahan, as indicated in ref. [185].

The *lac* deletion ΔX74 is also referred to as DE(codB-lacI)3. The extent of the deletion is reported in [274]. The extent of deletion Δ(*argF-lac*)U169 is described in [275].

Ref. [267] states that JM83 carries ϕ80d*lacI*^{q}*Z*ΔM15. The parent strain that contains the ϕ80, CSH51, is stated to carry ϕ80d*lac*$^+$ by J. Messing [267], as well as by J.H. Miller [154]. The other parent, 71.18, carries *lacI*^{q} and ZΔM15.

APPENDIX

B

Incompatibility of the *lacZα* marker with pBR322 plasmids

We found that transfer of the commonly used α-complementation marker *lacZα* originating from the pUC line of plasmids to plasmids bearing a pBR322 origin of replication did not yield viable transformants. Upon inspection, we found that the pUC *lacZα* marker contains a 90 base pair stretch in its coding sequence that is identical to a region near the origin of replication of pBR322. This suggests that the reason for the decreased viability is interference with replication, as a result of which transformed bacterial cells fail to express the antibiotic resistance marker that was used for plasmid maintenance. We used a pBR322 derivative plasmid (pTrc99A [187]) to create a construct where the *lacZα* marker (from pUC8 [276]) is under tight control of an overexpressed *lacI* repressor, while additionally using a *lacI* overexpressing strain DH5αZ1 [186]. In this strain we indeed observe a strong decrease of viability on plates when we induce expression of the *lacZα* marker (fig. B.1). We constructed a new alter-

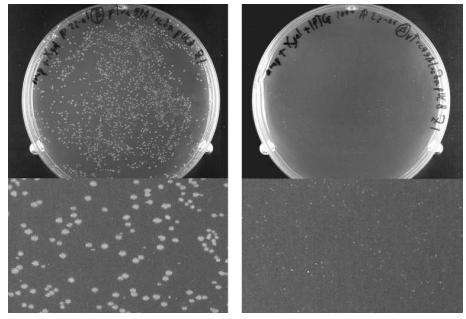


Figure B.1: *E. coli* DH5αZ1 cells harboring a pTrc99A derivative plasmid containing *lacZα* from pUC8 under control of the *lac* repressor. Overnight growth of DH5αZ1 on plates containing ampicillin (100 μg/ml) as a selection marker was normal in the absence of IPTG (left), but strongly reduced at 1 mM IPTG (right) where colonies were hardly visible by eye after 14 hours at 37^{o}C.

native α-complementation marker, by using the first 364 nucleotides from the chromosomal *lacZ* from strain MG1655, to avoid the observed incompatibility. In this way, we obtain the maximum length *lacZα* fragment that does not carry *lac* operator O2 (see fig. 1.4), which may limit the usability of the marker. Decrease of viability upon induction was not observed and α complementation was functional and similar to the pUC *lacZα* marker, as was verified by LacZ activity assays (see chapter 4).

α-complementation is the remarkable phenomenon that the β-galactosidase ac-

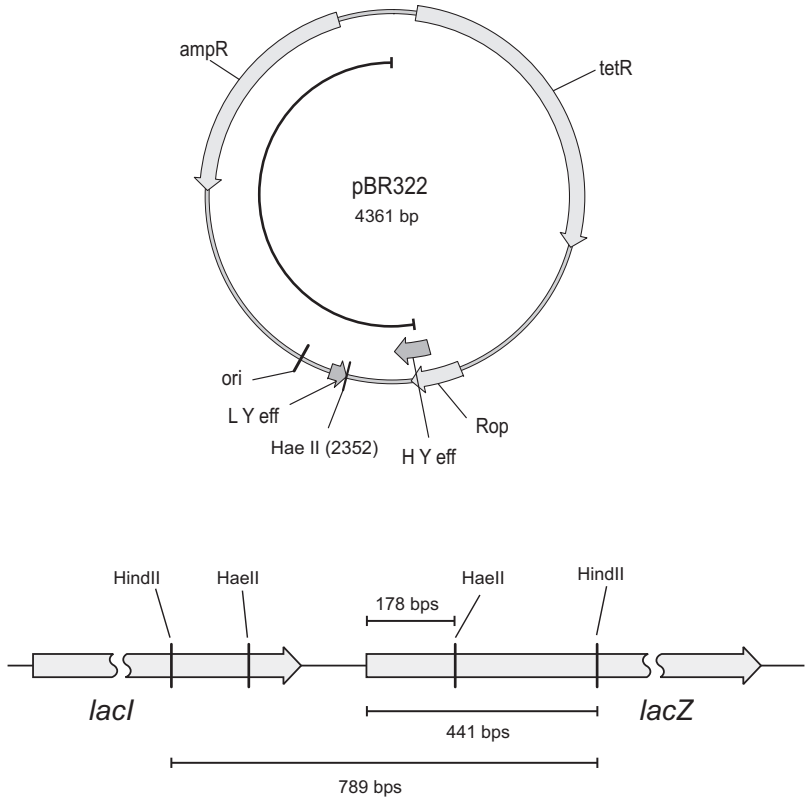


Figure B.2: Overview of cloning fragments involved in the creation of the *lacZα* gene of the pUC plasmid series [276]. Shown above is the pBR322 cloning plasmid [277]. *ampR* indicates an ampicillin resistance gene, *tetR* is a tetracycline resistance gene. 'ori' denotes the replication origin. Rop (also known as Rom) is a small protein modulating replication control [278]. Two Y effector sites [279] are indicated, which can function as origins of DNA replication [280]. The half-circle indicates the pBR322 derivative which received the *lacZα* fragment, and 'HaeII' denotes the location of the restriction site that was used to clone *lacZα* into it. The *lacZα* fragment originates from a phage M13 cloning vector [273], where it is present as a HindII restriction fragment from the *lac* operon, of which the relevant part is shown below (compare fig. 1.4). After HaeII digestion of the phage, the fragment containing the *lac* promoter and the first 178 base pairs from *lacZ* was ligated into the pBR322 derivative.

tivity of an inactive N-terminal deleted *lacZ* gene (*lacZω*) is recovered in the presence of a separate peptide containing the N-terminal part (*lacZα*), both *in vivo* and *in vitro* [281, 282]. Due to the small size of the α-peptide, it is has become one of the most commonly used plasmid or phage borne markers to identify bacterial colonies with a successful insertion of recombinant DNA. The larger ω-fragment is usually stably integrated in the chromosome of bacterial strains designed for gene cloning (see also appendix A). Although different size fragments display complementation [283], the most often used combination has been a *lacZ* mutant lacking amino acids 11-41 as ω donor (*lacZ*Δ(M15) [284]), with an α donor carried by coliphage M13 [273], or its derivative in the pUC series of plasmids [276].

We will briefly trace the construction history of the pUC *lacZα* gene that reveals the cause of the incompatibility observed above. One of the first cloning vehicles of recombinant DNA into *Escherichia coli* was the filamentous coliphage M13 [273, 285]. Recombinant DNA could be inserted in vitro into the DNA of M13 phage, which in turn could infect the bacteria and be stably maintained. To facilitate cloning and screening, the phage was modified by insertion of a *lacZα* marker [273]. This was done by incompletely digesting the phage DNA and making a blunt end ligation with a HindII restriction fragment of the *lac* operon (see fig. B.2). Due to its less infective nature, cloning using plasmid DNA became more popular. Initially natural isolates were used, but they were soon modified and stripped from parts not essential or inconvenient to the molecular biologist (e.g. [277, 286]). As plasmid cloning vectors would also profit from a screenable marker, *lacZα* from a phage M13 derivative was transferred [276] to a derivative from the versatile cloning vector pBR322 [277].

This was done by restriction of phage M13mp7 (containing *lacZα*) with HaeII, which yielded the fragment indicated in fig. B.2 containing a smaller α-marker than was present in the phage. A pBR322 derivative (consisting of base pairs 2067-4361 as indicated with the black half-circle in fig. B.2) was partially digested with HaeII, and a blunt end ligation of the digest with the M13 α-fragment was performed. This yielded some of the early members of the pUC series of plasmids [276]. Since the α-fragment was cloned into the pBR322 derivative without a stop codon, the translated protein contains an additional 30 amino acids, before an accidental stop codon is encountered. When this gene is further subcloned as its full open reading frame, it contains 90 base pairs that belong to the pBR322 origin of replication. As we stated above, this results in an incompatibility with plasmids containing the full pBR322 origin.

Bibliography

[1] Dallinger WH (1887) The President's Address. *J R Microsc Soc* **7**:185–199.

[2] Woltereck R (1909) Weitere experimentelle Untersuchungen über Artveränderung, speziell über das Wesen quantitativer Artunterschiede bei Daphniden. *Verh Deutschen Zool Ges* **1909**:110–172.

[3] Morgan TH (1913) Factors and unit characters in Mendelian heredity. *Am Nat* **47**:5–16.

[4] Falconer DS (1992) Early selection experiments. *Annu Rev Genet* **26**:1–14.

[5] Goodale HD (1937) Can artificial selection produce unlimited change? *Am Nat* **71**:433–459.

[6] Moose SP, Dudley JW, Rocheford TR (2004) Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends Plant Sci* **9**:358–364.

[7] Woese CR (1987) Bacterial evolution. *Microbiol Rev* **51**:221–271.

[8] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**:5463–5467.

[9] Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proc Natl Acad Sci USA* **74**:560–564.

[10] Dykhuizen D, Davies M (1980) An experimental model: Bacterial specialists and generalists competing in chemostats. *Ecology* **61**:1213–1227.

[11] Dykhuizen DE (1990) Experimental studies of natural selection in bacteria. *Annu Rev Ecol Syst* **21**:373–398.

[12] Chao L, Cox EC (1983) Competition between high and low mutating strains of Escherichia coli. *Evolution* **37**:125–134.

[13] Chao L, Levin BR, Stewart FM (1977) A complex community in a simple habitat: an experimental study with bacteria and phage. *Ecology* **58**:369–378.

[14] Jessup CM, Kassen R, Forde SE, Kerr B, Buckling A, et al. (2004) Big questions, small worlds: microbial model systems in ecology. *Trends Ecol Evol* **19**:189–197.

[15] Cooper VS, Lenski RE (2000) The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature* **407**:736–739.

[16] Rainey PB, Travisano M (1998) Adaptive radiation in a heterogeneous environment. *Nature* **394**:69–72.

[17] Meyer JR, Kassen R (2007) The effects of competition and predation on diversification in a model adaptive radiation. *Nature* **446**:432–435.

[18] Rainey P, Buckling A, Kassen R, Travisano M (2000) The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol Evol* **15**:243–247.

[19] Golding GB, Dean AM (1998) The structural basis of molecular adaptation. *Mol Biol Evol* **15**:355–369.

[20] Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* **8**:675–688.

[21] Thornton JW (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* **5**:366–375.

[22] Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**:86–89.

[23] Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA (1990) The ribonuclease from an extinct bovid ruminant. *FEBS Lett* **262**:104–106.

[24] Pauling L, Zuckerkandl E (1963) Chemical paleogenetics; molecular "restoration studies" of extinct forms of life. *Acta Chem Scand* **17**:S9–S16.

[25] Burch CL, Chao L (1999) Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics* **151**:921–927.

[26] Perfeito L, Fernandes L, Mota C, Gordo I (2007) Adaptive mutations in bacteria: high rate and small effects. *Science* **317**:813–815.

[27] Orr HA (2005) The genetic theory of adaptation: a brief history. *Nat Rev Genet* **6**:119–127.

[28] Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**:285–288.

[29] Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* **6**:533–543.

[30] Peisajovich SG, Tawfik DS (2007) Protein engineers turned evolutionists. *Nat Methods* **4**:991–994.

[31] Ricardo A, Carrigan MA, Olcott AN, Benner SA (2004) Borate minerals stabilize ribose. *Science* **303**:196.

[32] Eschenmoser A (1999) Chemical etiology of nucleic acid structure. *Science* **284**:2118–2124.

[33] Geyer CR, Battersby TR, Benner SA (2003) Nucleobase pairing in expanded Watson-Crick-like genetic information systems. *Structure* **11**:1485–1498.

[34] Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**:624–626.

[35] King JL, Jukes TH (1969) Non-Darwinian evolution. *Science* **164**:788–798.

[36] Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* **205**:581–598.

[37] Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, et al. (1985) Developmental constraints and evolution: A perspective from the Mountain Lake conference on development and evolution. *Quart Rev Biol* **60**:265–287.

[38] Antonovics J, van Tienderen PH (1991) Ontoecogenophyloconstraints? The chaos of constraint terminology. *Trends Ecol Evol* **6**:166–168.

[39] Pigliucci I, Kaplan I (2000) The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends Ecol Evol* **15**:66–70.

[40] Frankino WA, Zwaan BJ, Stern DL, Brakefield PM (2005) Natural selection and developmental constraints in the evolution of allometries. *Science* **307**:718–720.

[41] Crill WD, Wichman HA, Bull JJ (2000) Evolutionary reversals during viral adaptation to alternating hosts. *Genetics* **154**:27–37.

[42] Suiter AM, Bänziger O, Dean AM (2003) Fitness consequences of a regulatory polymorphism in a seasonal environment. *Proc Natl Acad Sci USA* **100**:12782–12786.

[43] Maynard Smith J (1978) Optimization theory in evolution. *Annu Rev Ecol Syst* **9**:31–56.

[44] Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Int Cong Genet* **1**:356–366.

[45] Gavrilets S (2004) Fitness Landscapes and the Origin of Species. Princeton: Princeton Univ. Press.

[46] Smith JM (1970) Natural selection and the concept of a protein space. *Nature* **225**:563–564.

[47] Bull JJ, Meyers LA, Lachmann M (2005) Quasispecies made simple. *PLoS Comput Biol* **1**:e61.

[48] Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* **102**:9541–9546.

[49] Gillespie JH (1984) Molecular evolution over the mutational landscape. *Evolution* **38**:1116–1129.

[50] Gillespie JH (1991) The Causes of Molecular Evolution. Oxford: Oxford Univ. Press.

[51] Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**:11–45.

[52] Kauffman SA (1993) The Origins of Order: Self-organization and Selection in Evolution. Oxford: Oxford Univ. Press.

[53] van Nimwegen E, Crutchfield JP (2000) Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull Math Biol* **62**:799–848.

[54] Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**:1165–1174.

[55] Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**:383–386.

[56] Weinreich DM, Delaney NF, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**:111–114.

[57] Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**:97–101.

[58] Lunzer M, Miller SP, Felsheim R, Dean AM (2005) The biochemical architecture of an ancient adaptive landscape. *Science* **310**:499–501.

[59] Poelwijk FJ, Kiviet DJ, Tans SJ (2006) Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data. *PLoS Comput Biol* **2**:e58.

[60] Burch CL, Chao L (2000) Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* **406**:625–628.

[61] Miller SP, Lunzer M, Dean AM (2006) Direct demonstration of an adaptive constraint. *Science* **314**:458–461.

[62] Rokyta DR, Joyce P, Caudle SB, Wichman HA (2005) An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat Genet* **37**:441–444.

[63] Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**:610–618.

[64] Martin G, Elena SF, Lenormand T (2007) Distributions of epistasis in microbes fit predictions from a fitness landscape model. *Nat Genet* **39**:555–560.

[65] van Nimwegen E (2006) Influenza escapes immunity along neutral networks. *Science* **314**:1884–1886.

[66] Koelle K, Cobey S, Grenfell B, Pascual M (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* **314**:1898–1903.

[67] Levene H (1953) Genetic equilibrium when more than one ecological niche is available. *Am Nat* **87**:331–333.

[68] Haldane JBS, Jayakar SD (1963) Polymorphism due to selection of varying direction. *J Genet* **58**:237–242.

[69] Levins R (1968) Evolution in Changing Environments: Some Theoretical Explorations. Princeton: Princeton University Press.

[70] Levins R (1962) Theory of fitness in a heterogeneous environment. I. The fitness set and adaptive function. *Am Nat* **96**:361–373.

[71] Scheiner SM (1993) Genetics and evolution of phenotypic plasticity. *Annu Rev Ecol Syst* **24**:35–68.

[72] Pigliucci M (2005) Evolution of phenotypic plasticity: where are we going now? *Trends Ecol Evol* **20**:481–486.

[73] Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**:588–592.

[74] Savageau MA (1974) Genetic regulatory mechanisms and the ecological niche of Escherichia coli. *Proc Natl Acad Sci USA* **71**:2453–2455.

[75] Savageau MA (1977) Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci USA* **74**:5647–5651.

[76] Shinar G, Dekel E, Tlusty T, Alon U (2006) Rules for biological regulation based on error minimization. *Proc Natl Acad Sci USA* **103**:3999–4004.

[77] Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309**:2075–2078.

[78] Troein C, Ahrén D, Krogh M, Peterson C (2007) Is transcriptional regulation of metabolic pathways an optimal strategy for fitness? *PLoS ONE* **2**:e855.

[79] Jasmin JN, Kassen R (2007) On the experimental evolution of specialization and diversity in heterogeneous environments. *Ecol Lett* **10**:272–281.

[80] Reboud X, Bell G (1997) Experimental evolution in Chlamydomonas. III. Evolution of specialist and generalist types in environments that vary in space and time. *Heredity* **78**:507–514.

[81] MacLean RC, Bell G, Rainey PB (2004) The evolution of a pleiotropic fitness tradeoff in Pseudomonas fluorescens. *Proc Natl Acad Sci USA* **101**:8072–8077.

[82] Carroll SB (2005) Evolution at two levels: on genes and form. *PLoS Biol* **3**:e245.

[83] Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci USA* **104**:8605–8612.

[84] Cases I, de Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* **11**:248–253.

[85] Barrier M, Robichaux RH, Purugganan MD (2001) Accelerated regulatory gene evolution in an adaptive radiation. *Proc Natl Acad Sci USA* **98**:10208–10213.

[86] Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430**:85–88.

[87] McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, et al. (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* **448**:587–590.

[88] Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol* **6**:e38.

[89] Pardee A, Jacob F, Monod J (1959) The genetic control and cytoplasmic expression of inducibility in the synthesis of β-galactosidase by E. coli. *J Mol Biol* **1**:165–178.

[90] Sadler JR, Novick A (1965) The properties of repressor and the kinetics of its action. *J Mol Biol* **12**:305–327.

[91] Yagil G, Yagil E (1971) On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophys J* **11**:11–27.

[92] Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* **33**:D334–D337.

[93] Müller-Hill B (1996) The *lac* Operon. Berlin: Walter de Gruyter.

[94] Weickert MJ, Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem* **267**:15869–15874.

[95] Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3**:e130.

[96] Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. *Nat Genet* **36**:492–496.

[97] Kalisky T, Dekel E, Alon U (2007) Cost-benefit theory and optimal design of gene regulation functions. *Phys Biol* **4**:229–245.

[98] Tănase-Nicola S, ten Wolde PR, *Biophys J, in press.*

[99] Purugganan MD (2000) The molecular population genetics of regulatory genes. *Mol Ecol* **9**:1451–1461.

[100] Simpson GG (1944) Tempo and Mode in Evolution. New York: Columbia Univ. Press.

[101] Darwin C (1859) On the Origin of Species by Means of Natural Selection. London: Murray.

[102] Ugalde JA, Chang BSW, Matz MV (2004) Evolution of coral pigments recreated. *Science* **305**:1433.

[103] Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* **47**:713–719.

[104] DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**:678–687.

[105] Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* **103**:5869–5874.

[106] Zhu G, Golding GB, Dean AM (2005) The selective cause of an ancient adaptation. *Science* **307**:1279–1282.

[107] Hurley JH, Dean AM, Koshland DE, Stroud RM (1991) Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and NADP+ complexes. *Biochemistry* **30**:8671–8678.

[108] Hurley JH, Chen R, Dean AM (1996) Determinants of cofactor specificity in isocitrate dehydrogenase: structure of an engineered NADP+ –> NAD+ specificity-reversal mutant. *Biochemistry* **35**:5670–5678.

[109] Kalodimos CG, Bonvin AMJJ, Salinas RK, Wechselberger R, Boelens R, et al. (2002) Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA-binding domain. *EMBO J* **21**:2866–2876.

[110] Kopke Salinas R, Folkers GE, Bonvin AMJJ, Das D, Boelens R, et al. (2005) Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. *Chembiochem* **6**:1628–1637.

[111] Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**:51–55.

[112] Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, Müller-Hill B (1990) Mutant lac repressors with new specificities hint at rules for protein–DNA recognition. *EMBO J* **9**:615–621.

[113] Barkai N, Leibler S (1997) Robustness in simple biochemical networks. *Nature* **387**:913–917.

[114] Kirschner M, Gerhart J (1998) Evolvability. *Proc Natl Acad Sci USA* **95**:8420–8427.

[115] Kitano H (2004) Biological robustness. *Nat Rev Genet* **5**:826–837.

[116] Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J (2004) Robustness of cellular functions. *Cell* **118**:675–685.

[117] Thattai M, van Oudenaarden A (2004) Stochastic gene expression in fluctuating environments. *Genetics* **167**:523–530.

[118] Arnold FH, Wintrode PL, Miyazaki K, Gershenson A (2001) How enzymes adapt: lessons from directed evolution. *Trends Biochem Sci* **26**:100–106.

[119] Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* **4**:457–469.

[120] Couñago R, Chen S, Shamoo Y (2006) In vivo molecular evolution reveals biophysical origins of organismal fitness. *Mol Cell* **22**:441–449.

[121] Lenski RE, Travisano M (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci USA* **91**:6808–6814.

[122] Stephens SG (1951) Possible significance of duplication in evolution. *Adv Genet* **4**:247–265.

[123] Ohno S (1970) Evolution by Gene Duplication. New York: Springer-Verlag.

[124] Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151–1155.

[125] Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**:1234–1244.

[126] Nguyen CC, Saier MH (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett* **377**:98–102.

[127] Bray D, Lay S (1994) Computer simulated evolution of a network of cell-signaling molecules. *Biophys J* **66**:972–977.

[128] François P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA* **101**:580–585.

[129] Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**:101–113.

[130] Sengupta AM, Djordjevic M, Shraiman BI (2002) Specificity and robustness in transcription control networks. *Proc Natl Acad Sci USA* **99**:2072–2077.

[131] Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* **240**:421–433.

[132] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. *Science* **298**:824–827.

[133] Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**:119–124.

[134] Koch AL (1983) The protein burden of lac operon products. *J Mol Evol* **19**:455–462.

[135] Fukami-Kobayashi K, Tateno Y, Nishikawa K (2003) Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol Biol Evol* **20**:267–277.

[136] Lehming N (1990) Regeln für Protein/DNA-Erkennung (PhD Thesis). Universität zu Köln.

[137] Hollis M, Valenzuela D, Pioli D, Wharton R, Ptashne M (1988) A repressor heterodimer binds to a chimeric operator. *Proc Natl Acad Sci USA* **85**:5834–5838.

[138] MacArthur S, Brookfield JFY (2004) Expected rates and modes of evolution of enhancer sequences. *Mol Biol Evol* **21**:1064–1073.

[139] Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**:2052–2058.

[140] Lynch M (2005) Simple evolutionary pathways to complex proteins. *Protein Sci* **14**:2217–2225.

[141] Francino MP (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**:573–577.

[142] Jürgens C, Strom A, Wegener D, Hettwer S, Wilmanns M, et al. (2000) Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc Natl Acad Sci USA* **97**:9925–9930.

[143] Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biol* **3**:research0008.

[144] Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**:186–189.

[145] Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* **4**:42.

[146] Inagaki Y, Doolittle WF, Baldauf SL, Roger AJ (2002) Lateral transfer of an EF-1α gene: origin and evolution of the large subunit of ATP sulfurylase in eubacteria. *Curr Biol* **12**:772–776.

[147] Stoebel DM (2005) Lack of evidence for horizontal transfer of the lac operon into Escherichia coli. *Mol Biol Evol* **22**:683–690.

[148] Bhan A, Galas DJ, Dewey TG (2002) A duplication growth model of gene expression networks. *Bioinformatics* **18**:1486–1493.

[149] Wagner A (2003) How the global structure of protein interaction networks evolves. *Proc Biol Sci* **270**:457–466.

[150] Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, et al. (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci USA* **101**:8414–8419.

[151] Weber W, Fussenegger M (2002) Artificial mammalian gene regulation networks-novel approaches for gene therapy and bioengineering. *J Biotechnol* **98**:161–187.

[152] Farmer WR, Liao JC (2000) Improving lycopene production in Escherichia coli by engineering metabolic control. *Nat Biotechnol* **18**:533–537.

[153] Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. *Proc Natl Acad Sci USA* **99**:16587–16591.

[154] Miller JH (1972) Experiments in Molecular Genetics. New York: Cold Spring Harbor Laboratory Press.

[155] Sadler JR, Sasmor H, Betz JL (1983) A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc Natl Acad Sci USA* **80**:6785–6789.

[156] Betz JL, Sasmor HM, Buck F, Insley MY, Caruthers MH (1986) Base substitution mutants of the lac operator: in vivo and in vitro affinities for lac repressor. *Gene* **50**:123–132.

[157] Dubertret B, Liu S, Ouyang Q, Libchaber A (2001) Dynamics of DNA-protein interaction deduced from in vitro DNA evolution. *Phys Rev Lett* **86**:6022–6025.

[158] Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**:64–68.

[159] Portalier R, Robert-Baudouy J, Stoeber F (1980) Regulation of Escherichia coli K-12 hexuronate system genes: exu regulon. *J Bacteriol* **143**:1095–1107.

[160] Ritzenthaler P, Mata-Gilsinger M, Stoeber F (1980) Construction and expression of hybrid plasmids containing Escherichia coli K-12 uxu genes. *J Bacteriol* **143**:1116–1126.

[161] Bates Utz C, Nguyen AB, Smalley DJ, Anderson AB, Conway T (2004) GntP is the Escherichia coli Fructuronic acid transporter and belongs to the UxuR regulon. *J Bacteriol* **186**:7690–7696.

[162] Ritzenthaler P, Blanco C, Mata-Gilsinger M (1983) Interchangeability of repressors for the control of the uxu and uid operons in E. coli K12. *Mol Gen Genet* **191**:263–270.

[163] Rodionov DA, Mironov AA, Rakhmaninova AB, Gelfand MS (2000) Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. *Mol Microbiol* **38**:673–683.

[164] Ritzenthaler P, Blanco C, Mata-Gilsinger M (1985) Genetic analysis of uxuR and exuR genes: evidence for ExuR and UxuR monomer repressors interactions. *Mol Gen Genet* **199**:507–511.

[165] McAdams HH, Srinivasan B, Arkin AP (2004) The evolution of genetic regulatory systems in bacteria. *Nat Rev Genet* **5**:169–178.

[166] Carroll SB (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**:577–580.

[167] Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL (2007) Genetic properties influencing the evolvability of gene expression. *Science* **317**:118–121.

[168] Via S, Gomulkiewicz R, de Jong G, Scheiner SM, Sclichting CD, et al. (1995) Adaptive phenotypic plasticity: consensus and controversy. *Trends Ecol Evol* **10**:212–216.

[169] Scheiner SM (2002) Selection experiments and the study of phenotypic plasticity. *J Evol Biol* **15**:889–898.

[170] Kassen R, Bell G (1998) Experimental evolution in Chlamydomonas. IV. Selection in environments that vary through time at different scales. *Heredity* **80**:732–741.

[171] van Tienderen PH (1997) Generalists, Specialists, and the Evolution of Phenotypic Plasticity in Sympatric Populations of Distinct Species. *Evolution* **51**:1372–1380.

[172] de Mazancourt C, Dieckmann U (2004) Trade-off geometries and frequency-dependent selection. *Am Nat* **164**:765–778.

[173] Agrawal AA (2001) Phenotypic plasticity in the interactions and evolution of species. *Science* **294**:321–326.

[174] Choi KY, Zalkin H (1992) Structural characterization and corepressor binding of the Escherichia coli purine repressor. *J Bacteriol* **174**:6207–6214.

[175] Pfahl M (1976) lac Repressor-operator interaction. Analysis of the X86 repressor mutant. *J Mol Biol* **106**:857–869.

[176] Scholz O, Henssler EM, Bail J, Schubert P, Bogdanska-Urbaniak J, et al. (2004) Activity reversal of Tet repressor caused by single amino acid exchanges. *Mol Microbiol* **53**:777–789.

[177] Kleina LG, Miller JH (1990) Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J Mol Biol* **212**:295–318.

[178] Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. *Science* **296**:1466–1470.

[179] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**:1377–419.

[180] Chesson P (2000) General theory of competitive coexistence in spatially-varying environments. *Theor Popul Biol* **58**:211–237.

[181] Marvin JS, Hellinga HW (2001) Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc Natl Acad Sci USA* **98**:4955–4960.

[182] Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet* **5**:34–42.

[183] Deans TL, Cantor CR, Collins JJ (2007) A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells. *Cell* **130**:363–372.

[184] Casadaban MJ, Cohen SN (1980) Analysis of gene control signals by DNA fusion and cloning in Escherichia coli. *J Mol Biol* **138**:179–207.

[185] Grant SG, Jessee J, Bloom FR, Hanahan D (1990) Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proc Natl Acad Sci USA* **87**:4645–4649.

[186] Lutz R, Bujard H (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res* **25**:1203–1210.

[187] Amann E, Ochs B, Abel KJ (1988) Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. *Gene* **69**:301–315.

[188] Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**:1453–1474.

[189] Huang ZJ (1991) Kinetic fluorescence measurement of fluorescein di-β-D-galactoside hydrolysis by β-galactosidase: intermediate channeling in stepwise catalysis by a free single enzyme. *Biochemistry* **30**:8535–8540.

[190] Monod J (1949) The growth of bacterial cultures. *Annu Rev Microbiol* **3**:371–394.

[191] Gay P, Le Coq D, Steinmetz M, Berkelman T, Kado CI (1985) Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. *J Bacteriol* **164**:918–921.

[192] Chambert R, Gonzy-Tréboul G, Dedonder R (1974) Kinetic studies of levansucrase of Bacillus subtilis. *Eur J Biochem* **41**:285–300.

[193] Chambert R, Gonzy-Tréboul G (1976) Levansucrase of Bacillus subtilis: kinetic and thermodynamic aspects of transfructosylation processes. *Eur J Biochem* **62**:55–64.

[194] DeLano WL The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA. (`http://www.pymol.org`).

[195] Stemmer WP (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**:389–391.

[196] Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat Biotechnol* **16**:258–261.

[197] Patel SS, Wong I, Johnson KA (1991) Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry* **30**:511–525.

[198] Stano NM, Jeong YJ, Donmez I, Tummalapalli P, Levin MK, et al. (2005) DNA synthesis provides the driving force to accelerate DNA unwinding by a helicase. *Nature* **435**:370–373.

[199] Zar JH (1999) Biostatistical Analysis, 4th ed. New Jersey: Prentice Hall.

[200] N-way ANOVA (anovan) MatLab 7.0 R14 sp2, Statistics Toolbox 3.0.

[201] Miller JH, Reznikoff WS (1978) The Operon. New York: Cold Spring Harbor Laboratory Press.

[202] Dean AM (1989) Selection and neutrality in lactose operons of Escherichia coli. *Genetics* **123**:441–454.

[203] Hartl DL, Clark AG (2007) Principles of Population Genetics, 4th ed. Sunderland: Sinauer Associates.

[204] Schaaper RM, Danforth BN, Glickman BW (1986) Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the Escherichia coli lacI gene. *J Mol Biol* **189**:273–284.

[205] Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* **148**:1667–1686.

[206] Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* **52**:1243–1253.

[207] Kuhlman T, Zhang Z, Saier MH, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Natl Acad Sci USA* **104**:6043–6048.

[208] Sahin-Tóth M, Gunawan P, Lawrence MC, Toyokuni T, Kaback HR (2002) Binding of hydrophobic D-galactopyranosides to the lactose permease of Escherichia coli. *Biochemistry* **41**:13039–13045.

[209] Tenu JP, Viratelle OM, Garnier J, Yon J (1971) pH dependence of the activity of β-galactosidase from Escherichia coli. *Eur J Biochem* **20**:363–370.

[210] Wahl LM, Gerrish PJ, Saika-Voivod I (2002) Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* **162**:961–971.

[211] Beckwith JR, Zipser D (1970) The Lactose Operon. New York: Cold Spring Harbor Laboratory Press.

[212] Lewis M (2005) The lac repressor. *C R Biol* **328**:521–548.

[213] Wilson CJ, Zhan H, Swint-Kruse L, Matthews KS (2007) The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cell Mol Life Sci* **64**:3–16.

[214] Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, et al. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**:1247–1254.

[215] Vilar JMG, Leibler S (2003) DNA looping and physical constraints on transcription regulation. *J Mol Biol* **331**:981–989.

[216] Oehler S, Alberti S, Müller-Hill B (2006) Induction of the lac promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Res* **34**:606–612.

[217] Cloutier TE, Widom J (2005) DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci USA* **102**:3645–3650.

[218] Saiz L, Vilar JMG (2008) Ab initio thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Res* **36**:726–731.

[219] Narang A (2007) Effect of DNA looping on the induction kinetics of the lac operon. *J Theor Biol* **247**:695–712.

[220] Novick A, Weiner M (1957) Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci USA* **43**:553–566.

[221] Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. *Nature* **427**:737–740.

[222] van Hoek MJA, Hogeweg P (2006) In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. *Biophys J* **91**:2833–2843.

[223] Goldberger RF (1979) Biological Regulation and Development, Vol. I Gene Expression. New York: Plenum Press.

[224] Barkley MD, Riggs AD, Jobe A, Bourgeois S (1975) Interaction of effecting ligands with lac repressor and repressor-operator complex. *Biochemistry* **14**:1700–1712.

[225] Müller-Hill B, Crapo L, Gilbert W (1968) Mutants that make more lac repressor. *Proc Natl Acad Sci USA* **59**:1259–1264.

[226] Oehler S, Eismann ER, Krämer H, Müller-Hill B (1990) The three operators of the lac operon cooperate in repression. *EMBO J* **9**:973–979.

[227] Oehler S, Amouyal M, Kolkhof P, von Wilcken-Bergmann B, Müller-Hill B (1994) Quality and position of the three lac operators of E. coli define efficiency of repression. *EMBO J* **13**:3348–3355.

[228] Schmitz A, Galas DJ (1979) The interaction of RNA polymerase and lac repressor with the lac control region. *Nucleic Acids Res* **6**:111–137.

[229] Lanzer M, Bujard H (1988) Promoters largely determine the efficiency of repressor action. *Proc Natl Acad Sci USA* **85**:8973–8977.

[230] Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181**:211–230.

[231] Zubay G, Schwartz D, Beckwith J (1970) Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proc Natl Acad Sci USA* **66**:104–110.

[232] Schmitz A (1981) Cyclic AMP receptor proteins interacts with lactose operator DNA. *Nucleic Acids Res* **9**:277–292.

[233] Fried MG, Hudson JM (1996) DNA looping and lac repressor-CAP interaction. *Science* **274**:1930–1931.

[234] Elf J, Li GW, Xie XS (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**:1191–1194.

[235] Gilbert W, Müller-Hill B (1966) Isolation of the lac repressor. *Proc Natl Acad Sci USA* **56**:1891–1898.

[236] Goldbeter A, Koshland DE (1981) An amplified sensitivity arising from covalent modification in biological systems. *Proc Natl Acad Sci USA* **78**:6840–6844.

[237] Sourjik V, Berg HC (2002) Receptor sensitivity in bacterial chemotaxis. *Proc Natl Acad Sci USA* **99**:123–127.

[238] Shilo BZ, Barkai N (2007) EGF receptor signaling - a quantitative view. *Curr Biol* **17**:R1038–R1041.

[239] Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: A plausible model. *J Mol Biol* **12**:88–118.

[240] Perutz MF (1989) Mechanisms of cooperativity and allosteric regulation in proteins. *Q Rev Biophys* **22**:139–237.

[241] Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. *Science* **308**:1424–1428.

[242] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**:116–124.

[243] Hasty J, McMillen D, Collins JJ (2002) Engineered gene circuits. *Nature* **420**:224–230.

[244] Mossing MC, Record MT (1986) Upstream operators enhance repression of the lac promoter. *Science* **233**:889–892.

[245] Lehming N, Sartorius J, Oehler S, von Wilcken-Bergmann B, Müller-Hill B (1988) Recognition helices of lac and lambda repressor are oriented in opposite directions and recognize similar DNA sequences. *Proc Natl Acad Sci USA* **85**:7947–7951.

[246] Müller-Hill B (1998) The function of auxiliary operators. *Mol Microbiol* **29**:13–18.

[247] Hopkins JD (1974) A new class of promoter mutations in the lactose operon of Escherichia coli. *J Mol Biol* **87**:715–724.

[248] Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci USA* **100**:7702–7707.

[249] Riggs AD, Newby RF, Bourgeois S (1970) lac repressor–operator interaction. II. Effect of galactosides and other ligands. *J Mol Biol* **51**:303–314.

[250] Riggs AD, Bourgeois S, Cohn M (1970) The lac repressor-operator interaction. III. Kinetic studies. *J Mol Biol* **53**:401–417.

[251] Deuschle U, Kammerer W, Gentz R, Bujard H (1986) Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures. *EMBO J* **5**:2987–2994.

[252] Kania J, Brown DT (1976) The functional repressor parts of a tetrameric lac repressor-β-galactosidase chimaera are organized as dimers. *Proc Natl Acad Sci USA* **73**:3529–3533.

[253] Wang AC, Revzin A, Butler AP, von Hippel PH (1977) Binding of E. coli lac repressor to non-operator DNA. *Nucleic Acids Res* **4**:1579–1593.

[254] Lin SY, Riggs AD (1972) Lac repressor binding to non-operator DNA: detailed studies and a comparison of equilibrium and rate competition methods. *J Mol Biol* **72**:671–690.

[255] von Hippel PH, Revzin A, Gross CA, Wang AC (1974) Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. *Proc Natl Acad Sci USA* **71**:4808–4812.

[256] Kao-Huang Y, Revzin A, Butler AP, O'Conner P, Noble DW, et al. (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**:4228–4232.

[257] Falcon CM, Matthews KS (2000) Operator DNA sequence variation enhances high affinity binding by hinge helix mutants of lactose repressor protein. *Biochemistry* **39**:11074–11083.

[258] Kirby DA, Muse SV, Stephan W (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci USA* **92**:9047–9051.

[259] Segrè D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* **37**:77–83.

[260] Phillips PC (1996) Waiting for a compensatory mutation: phase zero of the shifting-balance process. *Genet Res* **67**:271–283.

[261] Kimura M (1985) The role of compensatory neutral mutations in molecular evolution. *J Genet* **64**:7–19.

[262] Milnor J (1963) Morse Theory. Princeton: Princeton University Press.

[263] Whitlock MC, Phillips PC, Moore FB, Tonsor SJ (1995) Multiple fitness peaks and epistasis. *Annu Rev Ecol Syst* **26**:601–629.

[264] Buckling A, Wills MA, Colegrave N (2003) Adaptation limits diversification of experimental bacterial populations. *Science* **302**:2107–2109.

[265] Korona R, Nakatsu CH, Forney LJ, Lenski RE (1994) Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proc Natl Acad Sci USA* **91**:9037–9041.

[266] Fernàndez G, Clotet B, Martínez MA (2007) Fitness landscape of human immunodeficiency virus type 1 protease quasispecies. *J Virol* **81**:2485–2496.

[267] Messing J (1979) A multi-purpose cloning system based on a single-stranded DNA bacteriophage M13. *Recomb DNA Tech Bull, NIH Publ* **79-99**:43–48.

[268] Schweizer H, Boos W (1983) Transfer of the Δ(argF-lac)U169 mutation between Escherichia coli strains by selection for a closely linked Tn10 insertion. *Mol Gen Genet* **192**:293–294.

[269] Johnston TC, Thompson RB, Baldwin TO (1986) Nucleotide sequence of the luxB gene of Vibrio harveyi and the complete amino acid sequence of the β subunit of bacterial luciferase. *J Biol Chem* **261**:4805–4811.

[270] Emr SD, Hanley-Way S, Silhavy TJ (1981) Suppressor mutations that restore export of a protein with a defective signal sequence. *Cell* **23**:79–88.

[271] Meselson M, Yuan R (1968) DNA restriction enzyme from E. coli. *Nature* **217**:1110–1114.

[272] Beckwith JR, Signer ER (1966) Transposition of the lac region of Escherichia coli. I. Inversion of the lac operon and transduction of lac by phi80. *J Mol Biol* **19**:254–265.

[273] Messing J, Gronenborn B, Müller-Hill B, Hans Hopschneider P (1977) Filamentous coliphage M13 as a cloning vehicle: insertion of a HindII fragment of the lac regulatory region in M13 replicative form in vitro. *Proc Natl Acad Sci USA* **74**:3642–3646.

[274] Gerdes SY, Scholle MD, D'Souza M, Bernal A, Baev MV, et al. (2002) From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J Bacteriol* **184**:4555–4572.

[275] Peters JE, Thate TE, Craig NL (2003) Definition of the Escherichia coli MC4100 genome by use of a DNA array. *J Bacteriol* **185**:2017–2021.

[276] Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**:259–268.

[277] Bolivar F, Rodriguez RL, Greene PJ, Betlach MC, Heyneker HL, et al. (1977) Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene* **2**:95–113.

[278] Cesareni G, Muesing MA, Polisky B (1982) Control of ColE1 DNA replication: the rop gene product negatively affects transcription from the replication primer promoter. *Proc Natl Acad Sci USA* **79**:6313–6317.

[279] Zipursky SL, Marians KJ (1980) Identification of two Escherichia coli factor Y effector sites near the origins of replication of the plasmids ColE1 and pBR322. *Proc Natl Acad Sci USA* **77**:6521–6525.

[280] Zipursky SL, Marians KJ (1981) Escherichia coli factor Y sites of plasmid pBR322 can function as origins of DNA replication. *Proc Natl Acad Sci USA* **78**:6111–6115.

[281] Ullmann A, Perrin D, Jacob F, Monod J (1965) Identification, by in vitro complementation and purification, of a peptide fraction of Escherichia coli β-galactosidase. *J Mol Biol* **12**:918–923.

[282] Ullmann A, Jacob F, Monod J (1967) Characterization by in vitro complementation of a peptide corresponding to an operator-proximal segment of the β-galactosidase structural gene of Escherichia coli. *J Mol Biol* **24**:339–343.

[283] Ullmann A (1992) Complementation in β-galactosidase: from protein structure to genetic engineering. *Bioessays* **14**:201–205.

[284] Langley KE, Fowler AV, Zabin I (1975) Amino acid sequence of β-galactosidase. IV. Sequence of an α-complementing cyanogen bromide peptide, residues 3 to 92. *J Biol Chem* **250**:2587–2592.

[285] van Wezenbeek PM, Hulsebos TJ, Schoenmakers JG (1980) Nucleotide sequence of the filamentous bacteriophage M13 DNA genome: comparison with phage fd. *Gene* **11**:129–148.

[286] Bolivar F, Rodriguez RL, Betlach MC, Boyer HW (1977) Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. *Gene* **2**:75–93.

Summary

Environmental variability has been hotly debated in evolutionary biology, as it is considered the evolutionary cause of cellular regulation, and ultimately responsible for much of an organism's complexity. However, quantitative experimental data has been largely lacking, because of the limited phenotypic understanding of the organisms studied so far, and the technical challenges associated with variable environments. In this thesis we have explored several aspects of the evolution of bacterial gene regulation. In the described work our aim has been to observe how variable selective pressures experienced by an organism drive evolutionary change at the molecular level, and how properties of molecules and their interactions in networks determine evolutionary potential and constraints. Our access to the molecular level is enabled by two approaches. The first is the use of model systems, a model organism, a model regulatory system, or a synthetic system, in order to maximize our functional understanding. The second is the use of fitness landscapes, which reveal how molecular variation affects the fitness of an organism.

In chapter 2 we have described the emerging use of empirical fitness landscapes in evolutionary studies. A major question in this field is to what extent natural fitness landscapes are rugged, since the ruggedness of the fitness surface determines whether a mutation-by-mutation evolutionary process can reach an optimal solution under a certain selective pressure. Most studies sofar have shown that evolution is constrained, but that a subset of possible evolutionary trajectories is accessible, which has interesting implications for the repeatability and predictability of evolution. A potential source of frustration in evolutionary processes are the 'key-lock' issues that arise when two components of a system are co-adapted and exhibit a specific mutual interaction. When such a system is under a selective pressure to change, a new interaction may be reachable only via a deleterious intermediate situation where the interaction is reduced due to a first mutation in one of the interaction partners.

In chapter 3 we focused on a typical key-lock problem that must have been overcome many times during evolutionary history: that of a transcriptional regulator with its DNA binding site. Many of the transcription factors in *Escherichia coli* are part of families whose members display a high level of sequence and structural homology, and must have arisen from ancient duplication events. The present-day members of these families are highly specific for their own binding sites and cross-interactions are generally weak. In chapter 3 we computationally investigated the evolutionary divergence of a duplicated pair of transcription factors and their binding sites, using a large fit-

ness landscape based on experimentally determined binding affinities of *lac* repressor mutants. We showed that the initially redundant network topology can alleviate the key-lock dilemma, so that rapid evolutionary trajectories towards high fitness exist that do not contain deleterious intermediate steps.

Central to understanding the selective pressure on regulation are the performance trade-offs experienced by an organism living in an environment that alternates between different states. Trade-offs arise when optimizing fitness in one environmental state implicates a fitness decrease in another state. Optimality in the context of trade-offs has been analyzed on an abstract level, but lacked experimental verification so far. Although correlations between performances in different environmental states are ubiquitously observed (e.g. adaptation to the dark resulting in diminished performance in the light), they may merely reflect an accumulation of mutations. In chapter 4 we present the first determination of a trade-off relation for the expression level of a gene. Using a synthetic operon we obtained full control over both phenotype and fitness consequences. This allowed us to determine the fitness landscape for gene regulation in alternating environments, thus quantifying the selective pressures. We demonstrated how the trade-off curve shape can be altered, which changes the relative performances of regulatory phenotypes. Adaptation experiments were performed where a regulatory protein and more complex regulatory systems adapted to new, imposed, environments. In this way *lac* repressors with an inverse response to inducer were obtained, as well as 'dual-input' regulatory networks with boolean responses as 'OR' and 'NAND'.

In chapter 5 we zoomed in on the molecular details of the newly evolved inverse repressors. We set out to identify functional mutations and epistasis between mutations at the level of the genotype-fitness landscape. We used PCR-based recombination followed by conservative selection in alternating environments to 'filter out' non-functional mutations. In this way we can reduce the complexity of the fitness landscape that has to be assayed. We developed a statistical method to remove correlations between mutations that arise during such combinatorial approaches. A number of potentially epistatically interacting mutations was identified.

In chapter 6 we use serial dilution of a growing population of *E. coli* to investigate the adaptation and optimality of gene regulation for the natural *lac* operon in constant and alternating environments. To observe adaptation however, one must start with a mal-adapted system. In de presented work we were able to create suboptimal starting points by decoupling the relation between the inductive properties and the catabolic properties of lactose, that arguably has been optimized by natural evolution. Hence we were able to measure fitness for independently varying gene expression levels and concentrations of carbon source in the environment, which allowed to assess regulatory optimality for different environmental conditions. In the serial dilution experiments we generally observed a fast adaptation to near optimal regulatory responses. In some cases the genetic accessibility of fitness-improving but suboptimal mutations

prevented fixation of an optimal mutation, at least temporarily.

A part of our experiments was performed using *lac* repressor overexpressing strains (chapters 4 and 5). For these systems we found a discrepancy between the measured induction curves and the available theoretical descriptions of the induction profile (in combination with experimentally determined reaction constants). We found that an important ingredient lacking in recent theoretical descriptions of *lac* induction is the residual affinity of repressors when they are fully saturated with inducer. In chapter 7 we present a basic thermodynamic model that does incorporate the residual affinity and recover a close match to the induction data in our overexpressing strains. We showed that the presence of residual affinity may well be a determining factor in setting the evolutionary optimal repressor copy number. Residual affinity must be a general property of allosteric regulators (we observed similar effects for the *tet* repressor), and especially when the expression level of regulators is regulated itself, it will be an important factor to take into account in theoretical modeling, as well as in the creation of artificial bio-circuitry in synthetic biology.

In chapter 8 we return to the key-lock issue, now in relation to multiple-peaked fitness landscapes. When a fitness landscape contains several adaptive peaks, this is not only interesting because they represent alternative solutions to the same problem, but also because the peaks might not be equally high, which opens the possibility for an evolving population to get entrapped on a sub-optimal peak. We showed that any fitness landscape harboring multiple peaks contains at least one epistatic motif referred to as 'reciprocal sign epistasis'. This motif in fact represents a key-lock situation between two of the elementary mutations in the landscape's sequence space.

Samenvatting

Bacteriën hebben een verbluffend vermogen om hun leefomgeving waar te nemen en erop te reageren. Daardoor kunnen ze efficiënt gebruik maken van de voedingsstoffen en omstandigheden die ze daar aantreffen. Dit doen ze dankzij hun 'regulatiesystemen', die een signaal uit de omgeving vertalen in bijvoorbeeld de productie van een eiwit. Zo maken sommige bacteriën, wanneer er suiker in hun omgeving aanwezig is, eiwitten die dat suiker kunnen afbreken. Dit levert energie en bouwstoffen, waarmee ze kunnen groeien. Wanneer er geen suiker is, is het niet voordelig als de bacterie de eiwitten blijft maken. De productie van het eiwit kost namelijk ook een beetje energie, maar levert niets op zonder suiker. Het regulatie-systeem van de bacterie schakelt de productie van de eiwitten dan uit.

Hoewel het functioneren van veel bacteriële regulatie-systemen redelijk goed begrepen wordt, is er veel minder bekend over hoe ze zijn ontstaan tijdens de evolutie. In dit proefschrift is een aantal onderzoeken beschreven, waarin is gekeken naar de evolutionaire aanpassing van zo'n regulatie-systeem. De essentie van veel van dit soort systemen is een eiwit (de 'repressor') dat op een precieze plaats op het DNA (de 'operator') kan binden. Als de repressor daar gebonden is, voorkomt het dat andere eiwitten worden gemaakt. En als de repressor loslaat (in ons eerdere voorbeeld doordat er een suikermolecuul aan de repressor bindt), worden de andere eiwitten geproduceerd.

Er is een aantal redenen waarom het nuttig is om juist de evolutie van dit soort systemen te proberen te begrijpen. Ten eerste bevatten alle organismen regulatie-systemen en denkt men tegenwoordig dat de verschillen tussen verwante organismen vooral bepaald worden door verschillen in hun regulatie-systemen. Organismen hebben vaak voor een groot gedeelte hetzelfde erfelijk materiaal en daardoor dezelfde eiwitten, maar hoe die eiwitten gereguleerd worden is verschillend.

Ten tweede zijn er interessante evolutionaire vragen rondom deze systemen. De repressor past heel precies op zijn operator, als een sleutel in zijn slot. Vaak hoeft er maar één verandering in het DNA plaats te vinden (een mutatie) op de plaats van de operator en de repressor kan niet meer binden. Als dat gebeurt kan het systeem niet meer reageren op een signaal uit de omgeving: het staat altijd aan. Hetzelfde geldt voor kleine veranderingen in de repressor. Er zijn in een veelbestudeerde bacterie als *E. coli* veel 'sleutels' en 'sloten' die erg op elkaar lijken, maar toch alleen in de goede combinatie op elkaar passen. Van deze systemen weten we dat ze stap voor stap veranderd zijn tijdens de evolutie. Een vraag is dus: hoe kan zo'n systeem functioneel blijven en toch veranderen?

Verder is het belangrijk dat deze systemen zijn geëvolueerd in een omgeving die voortdurend verandert (bijvoorbeeld een variërende hoeveelheid suiker). Als een organisme nog geen goed regulatie-systeem heeft, zal het een gedeelte van de tijd niet de optimale hoeveelheid eiwitten produceren. Soms is dat niet erg; bijvoorbeeld wanneer er bijna altijd suiker is en de bacterie altijd veel eiwit produceert. In dat geval is het voordeel van een regulatie-systeem niet groot. Het hangt dus af van *hoe* de omgeving precies verandert, of een regulatie-systeem nuttig is en kan evolueren.

Voordat ik noem welke vragen we in de verschillende hoofdstukken van dit proefschrift precies hebben onderzocht, wil ik eerst een begrip introduceren dat wij gebruiken om de evolutie te beschrijven: namelijk het fitness-landschap.

Een fitness-landschap laat zien bij welke combinatie van mutaties een organisme het best aangepast is aan zijn omgeving. Het fitness-landschap is een enigszins abstract begrip, maar kan vergeleken worden met een berglandschap met pieken en dalen. De pieken stellen combinaties van mutaties voor die maken dat een organisme goed is aangepast aan zijn omgeving. De dalen geven de combinaties weer waarmee organismen slecht overleven. Het meten van fitness-landschappen, door precies te kijken wat de effecten zijn van heel veel combinaties van mutaties, is een vrij nieuwe wetenschappelijke trend.

Als je weet hoe het fitness-landschap eruit ziet, kun je een voorspelling doen hoe de erfelijke eigenschappen van een organisme stap voor stap zullen veranderen, wanneer het zich in een omgeving bevindt waaraan het niet goed is aangepast. Deze reeks van stappen door het fitness-landschap wordt het 'evolutionaire pad' genoemd.

Omdat de natuur vaak alleen verbeteringen toelaat, kun je in een fitness-landschap alleen maar omhoog lopen. Totdat je een top bereikt. Het interessante is nu, dat dat niet altijd de hoogste top is. Het kan een lage berg zijn –wat betekent dat het organisme niet optimaal is aangepast aan zijn omgeving– maar omdat stappen naar beneden niet kunnen, evolueert het toch niet verder.

De vragen die in studies naar fitness-landschappen gesteld worden zijn: is het mogelijk in het landschap van een punt A naar een punt B lopen, zonder ergens halverwege op een lage piek vast komen te zitten? Zijn er veel verschillende paden die je kunt volgen, of zijn er maar weinig? En welke mutaties zijn cruciaal? Vaak is voor iedere stap in het landschap bekend wat voor erfelijke veranderingen er plaatsvinden op het niveau van moleculen zoals eiwitten en DNA en hoe dit de chemische reacties tussen bio-moleculen beïnvloedt. Hierdoor krijg je gedetailleerde informatie over welke eigenschappen van die moleculen maken dat iets makkelijk evolueert of moeilijk.

Ik geef een minder abstract voorbeeld. Wanneer er een nieuw antibioticum wordt gebruikt, verschijnen er meestal snel bacteriën die ook resistent zijn tegen dat nieuwe antibioticum. Dit is een van de meest opvallende voorbeelden van evolutie. Patiënten die geïnfecteerd raken met zulke 'ziekenhuisbacteriën' ondervinden daarvan vaak grote complicaties en ziekenhuizen investeren veel tijd en geld om uitbraken te voorkomen. Daardoor zou het goed zijn wanneer we zouden weten hoe de evolutie van resistentie

verloopt.

Niet zo lang geleden hebben Amerikaanse onderzoekers een fitness-landschap bepaald van het resistentie-eiwit van een bacterie en hebben daarbij gevonden dat er van alle mogelijke evolutionaire paden (van de oude resistentie naar een nieuwe) maar een beperkt aantal gevolgd kan worden. Dit soort informatie kan van groot belang zijn om te bepalen wat voor antibioticum gebruikt moet worden of welke combinatie van verschillende antibiotica.

Zoals gezegd, hebben we in dit proefschrift gekeken naar de evolutie van regulatiesystemen. We doen dit met behulp van fitness-landschappen en laten zien dat deze belangrijke informatie kunnen opleveren. Ook hebben we onderzocht hoe je de evolutie ook een bepaalde richting op kan sturen, wanneer je het fitness-landschap begrijpt.

In hoofdstuk 2 hebben wij de opkomende trend beschreven waarin fitness-landschappen gebruikt worden zoals hierboven is uitgelegd. We laten ondermeer zien wat de verschillen en overeenkomsten zijn tussen de evolutie van eiwitten die chemische stoffen afbreken en de evolutie van regulatie-systemen. Het zojuist genoemde voorbeeld van de resistentie tegen antibiotica is een van de onderzoeken die we beschrijven.

In hoofdstuk 3 kijken we gedetailleerd naar de evolutie van een sleutel-slot systeem. We gebruiken hier een computer-programma om in een fitness-landschap mogelijke evolutionaire paden te vinden die het sleutel-slot dilemma oplossen. Wij zien dat in een belangrijk evolutionair proces – mutaties waar genetisch materiaal verdubbeld wordt – het sleutel-slot probleem omzeild kan worden. Er zijn dan aanvankelijk twee dezelfde sleutels en twee dezelfde sloten die ervoor kunnen zorgen dat het regulatie-systeem niet slechter wordt als er veranderingen optreden. Zo kunnen er toch stap voor stap twee unieke sleutel-slot combinaties ontstaan.

In hoofdstuk 4 beschrijven we experimenten waarin voor de eerste keer een fitnesslandschap bepaald wordt van een regulatie-systeem in een variabele omgeving. We laten zien hoe de vorm van het landschap (hoe spits de piek is) afhangt van chemische reacties die plaatsvinden in de bacteriën. We kunnen daarna de omgeving zo veranderen dat de regulatie-systemen die eerst optimaal waren, nu slecht presteren. Met het fitness-landschap kunnen we dan voorspellen wat voor regulatie-systeem in de nieuwe omgeving optimaal is. We maken mutaties in de systemen en kijken of ze zich aanpassen aan de nieuwe omgeving. We zien inderdaad dat de nieuwe piek bereikt wordt (zie bijvoorbeeld figuur 4.4) .

In hoofdstuk 5 kijken we naar de geëvolueerde systemen uit hoofdstuk 4. We proberen hier inzicht te krijgen in welke mutaties ervoor zorgen dat de systemen beter presteren. We gebruiken een methode die nuttig is in combinatie met fitness-landschappen: we evolueren als het ware terug in de richting van het originele systeem en raken daarbij de mutaties die niet belangrijk waren kwijt.

In hoofdstuk 6 kijken we weer naar de evolutie van regulatie in variabele omgevingen, maar nu zonder dat we zelf mutaties in het regulatiesysteem maken. We kijken dus

of bacteriën zich aanpassen aan de omgeving door spontane mutaties die ze oplopen als ze groeien. We groeien culturen met grote aantallen bacteriën (ongeveer een miljard per flesje) in constante en variërende omgevingen. De bacteriën die een gunstige mutatie oplopen, groeien beter dan de andere en zullen langzaam maar zeker de cultuur overnemen ten koste van de andere. Dit zien we inderdaad gebeuren. We zien dat de bacteriën vaak snel evolueren naar de regulatie waarvan we hadden voorspeld dat hij optimaal zou zijn. In sommige gevallen zijn er beperkingen waardoor dat niet gebeurt; bijvoorbeeld omdat een net-niet-optimale mutatie veel vaker optreedt dan een optimale.

Doordat we veel met een bepaald regulatie-systeem werkten, bleek dat er iets miste in de theoretische modellen in de literatuur die dat systeem beschrijven. In hoofdstuk 7 presenteren we een model, waarin dit wel is opgenomen en laten zien dat dit metingen die we gedaan hebben verklaart. Een correcte beschrijving van het regulatie-systeem is namelijk belangrijk om te kunnen begrijpen hoe het evolueert.

In hoofdstuk 8, tenslotte, kijken we op een wat abstractere manier naar het fitness-landschap en laten zien we dat het sleutel-slot dilemma altijd een rol speelt wanneer een fitness-landschap meerdere pieken heeft. En zoals eerder gezegd heeft het bestaan van meerdere pieken (optimale en minder optimale) grote invloed op het verloop van de evolutie.

Publications

Frank J. Poelwijk, Daniel J. Kiviet, and Sander J. Tans
Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data (chapter 2)
PLoS Comput. Biol. 2: e58 (2006)

Frank J. Poelwijk, Daniel J. Kiviet, Daniel M. Weinreich, and Sander J. Tans
Empirical fitness landscapes reveal accessible evolutionary paths (chapter 3)
Nature 445: 383-386 (2007)

Frank J. Poelwijk and Sander J. Tans
Adaptive landscapes of gene regulatory systems in variable environments (chapter 4)
submitted

Frank J. Poelwijk and Sander J. Tans
Identification of functional mutations and epistasis by reverse neutral evolution (chapter 5)
in preparation

Frank J. Poelwijk, Philip Heijning, Marjon G.J. de Vos, Daniel J. Kiviet, and Sander J. Tans
Maintenance and loss of gene regulation in experimental evolution (chapter 6)
in preparation

Frank J. Poelwijk and Sander J. Tans
Residual affinity of induced repressors alters the shape of induction curves (chapter 7)
in preparation

Frank J. Poelwijk, Sorin Tănase-Nicola, Daniel J. Kiviet, and Sander J. Tans
Reciprocal sign epistasis and multiple peaks in the fitness landscape (chapter 8)
in preparation

Frank J. Poelwijk, Eva E.F. Riemslag, and Sander J. Tans
Incompatibility of the common *lacZα* marker with pBR322 plasmids (appendix B)
in preparation

Dankwoord

Lasciate ogne speranza, voi ch'intrate (Laat alle hoop varen, gij die hier binnentreedt), is niet alleen een regel uit La Divina Commedia van Dante Alighieri, maar staat ook te lezen boven de deur van het microbiologisch laboratorium op AMOLF. Het is daar opgehangen door Martijn van Duijn en mij, niet zo lang nadat dit lab in gebruik werd genomen. Inmiddels is het een enigszins vergeeld velletje papier geworden en ik weet niet of er ooit nog acht op wordt geslagen, maar niemand die hier voor het eerst in zijn of haar carrière microbiologisch werk verricht (en dat is in een fysisch instituut toch het merendeel) kan zich erop beroepen niet vooraf gewaarschuwd te zijn.

Ik zelf dus ook niet. Als fysicus aan het werk met biologische systemen heb ik inderdaad geleerd een bepaald soort hoop te laten varen. De natuur doet in veel gevallen geen duidelijke uitspraken en experimenten bedoeld om een 'ja' of 'nee' antwoord te krijgen, leverden vaak een 'misschien' op, of een 'soms', of gewoon iets vreemds. Veelal blijken de biologische systemen waaraan hier gewerkt wordt lastig te temmen, terwijl ze wel getemd *moeten* worden om ze te kunnen beschrijven op een voldoende kwantitatief niveau. Maar uiteindelijk ligt hierin toch, althans voor mij, ook de grootste bron van genoegen: wanneer het lukt aspecten van een biologisch systeem in een exacte beschrijving te vangen.

Net zoals in evolutie, wordt ook in een promotie het traject in hoge mate bepaald door de omgeving. Ik heb op AMOLF deze omgeving, met z'n constante en variabele factoren, als zeer prettig en stimulerend ervaren. Als eerste wil ik hier mijn begeleider Sander Tans bedanken, dé constante factor tijdens mijn promotie. Ik denk dat je aanvankelijke idee om vanuit een kwantitatief perspectief te kijken naar 'de veranderbaarheid van biochemische netwerken' een unieke niche creëerde hier in Nederland en ik ben erg blij dat je mij toentertijd als promovendus hebt aangenomen. Ik heb het enorm gewaardeerd dat we zo exploratief bezig konden zijn in een gebied dat voor ons beiden nieuw was. Ik waardeer je openheid voor discussies, je enthousiasme en je vertrouwen in een goede afloop (of althans het sterk wekken van die indruk) in de periode dat resultaten nog op zich lieten wachten.

Ik wil mijn promotor Daan Frenkel bedanken, omdat veel van de biologische explo raties op AMOLF mede door hem zijn gekatalyseerd. Hij en de andere groepsleiders van de 'Overloop' hebben het voor elkaar gekregen om een zeer prettige sfeer te creëren, zowel wetenschappelijk als sociaal. Natuurlijk hangt sfeer af van alle aanwezigen, maar ik denk dat de rol van de groepsleiders hierin niet moet worden onderschat. Ik wil Marileen Dogterom ook bedanken voor mijn introductie bij de vakgroep Moleculaire

Cytologie van de UvA, waar we samen DNA preps hebben gedaan en gels gerund (die voor mij de eerste waren, maar zeker niet de laatste). Ik dank Bela Mulder voor de keren dat ik het antwoord op een wiskundige vraag, waarvan we de oplossing niet ter plekke doorzagen, de volgende ochtend grondig uitgewerkt op mijn bureau vond. Ook wil ik Pieter Rein ten Wolde bijzonder bedanken voor de vele discussies en zijn universele bereidheid na te denken over vragen die je hem stelt.

Ik wil hier ook mijn speciale dank uitspreken voor dhr. Smit, biologieleraar op mijn middelbare school, die mijn enthousiasme voor biologie erg heeft gestimuleerd tijdens mijn eerste echte evolutionaire experiment waar we keken of de oogkleur van *Drosophila* invloed heeft op hun reproductieve fitness. Het is bizar om te realiseren hoe dicht dat experiment stond bij wat ik de afgelopen jaren heb gedaan.

Als dit proefschrift een boodschap 'mede mogelijk gemaakt door' zou bevatten, dan zouden Kim Renders en Roland Dries de eersten zijn die daar genoemd worden. Jullie inzet en ondersteuning in het lab zijn voor mij onmisbaar geweest. Ik ken weinig promovendi die zo'n voorrecht hebben gehad. Ook onder dit kopje zou Philip Heijning komen, mijn eerste afstudeerstudent, die zich een toegewijd serial-diluter betoonde en zich met groot enthousiasme op de kosten-baten analyse stortte. Enthousiasme is sowieso voor jou een wezenskenmerk en je droeg daardoor erg bij aan de goede sfeer in de groep. Ik ben benieuwd waar je uiteindelijk zal neerstrijken, in de muziek, de economische wereld, of toch de wetenschap, of een combinatie?

Ik bedank de mensen van de werkplaats en de ontwerpafdeling voor een aantal noodzakelijke en zeer handige gereedschappen in het bio-lab, zoals de replica-plater, de kolonieprikker (versies 1 tot, ik geloof, 3) en de petrischaal-imager. En natuurlijk de mensen van de receptie, de kantine, de bibliotheek, E&I en het magazijn...

Van buiten AMOLF wil ik speciaal bedanken Conrad Woldringh van wiens kritische houding ten opzichte van labgebruiken en protocols ik veel geleerd heb, en Tanneke den Blaauwen voor de discussies, haar hulpvaardigheid, dat ik met haar promovendus Gert-Jan Kremers mocht meelopen om moleculair biologische technieken te leren en later een eigen lab-bench kreeg toen het lab op AMOLF nog niet klaar was. En ik dank natuurlijk Gert-Jan zelf voor zijn geduld om mij in te wijden in de bewuste technieken.

Verder wil ik alle mensen bedanken die voor mij AMOLF een heel fijne plek hebben gemaakt om (samen) te werken en vaak ook buiten het werk voor vermaak zorgden. Allereerst zijn daar mijn huidige en voormalige groepsgenoten: mijn mede-Tanser van het eerste uur en goede vriend Ruud, Eva, Thomas, Matt, Philipp, Daan wiens gevoel voor humor een essentieel onderdeel is van onze groep, Ienas with her kindness, Aileen with her cold-bloodedness in case of lab-fires (AMOLF should know what could have happened without her...), ons sociale powerhouse Marjon, just-father Manju, Jerien, Ndika, en onze oud-studenten Genison, Robert en Merlijn.

Ook wil ik hier mijn kamergenoten bedanken (voor zover niet al genoemd), voor wie ik een constante omgevingsfactor ben geweest aangezien ik hier zit sinds de ingebruikname van de overloop en zo ongeveer wegga als het nieuwe AMOLF gebouw klaar

is. Speciaal wil ik hier noemen Martijn van Duijn, wiens lab-stijl ik erg heb gewaardeerd, Tatiana Schmatko with whom I very much enjoyed the pique-niques in the parks around AMOLF whenever the weather allowed, en Gertjan, mijn paranimf met zijn goede humeur en zijn gevatheid. Gertjan, ik heb onze bespiegelingen over het werk in het biolab en überhaupt over van alles en nog wat erg gewaardeerd; ik wens je veel succes met het laatste deel van je promotie!

Among the past AMOLF-ers whose leaving made me deplore the fast fluctuating environment in science are Rosalind Allen, also a steady picnic-er, whom I could assist in her (partial) conversion from theorist into experimentalist, and Sorin Tănase-Nicola, who would never allow anything remotely in that direction. Both of them I want to thank for the scientific discussions and their friendship. Further, Guillaume Romet-Lemonne who managed his younger brother to give up his Paris apartment so that we could stay there during a visit, en ook Gerbrand Koster die ik nu veel geluk wens bij het opbouwen van een nieuw bestaan in Noorwegen.

Verder zijn er nog heel veel mensen (geweest) op AMOLF die voor de goede sfeer zorgen en gezorgd hebben. Dit zijn zeker, maar niet alleen: Paige and her liveliness, Liedewij, Rutger, Julien that decorates the corridor with paintings of completely unknown girls, Christian and his boat, Nienke, Iza, Marco (2x), Frans die altijd goed is voor een hart onder de riem tijdens late uurtjes op AMOLF, Ioana, Maarten, Maria, Rhoda, Simon, Jacob K., Chantal, Behnaz, Niels die er soms 's middags om half vijf al elf uur op heeft zitten, Thorsten, Siebe die tegelijk met mij ploeterde op zijn proefschrift, Eva, Patrick, Koos, Sanne, Kostya, Ana, Andrea, Marina, Svenja, Nefeli.

Buiten AMOLF is er een aantal mensen die ik wil bedanken voor het hooghouden van de moraal en de gezellige tijden (maar van wie ik er een aantal door alle drukte ook veel minder heb kunnen zien dan ik wel gewild zou hebben). Ik noem hier Reinout, Dagmara en Marieke voor onze goede tijden in Utrecht, België en Triëste, Lotje voor haar meelevendheid (jij nu ook succes met de laatste loodjes!), Christiaan, Janne, Ineke, Charlie, Alice for her constant interest in how I was doing, Bastian voor zijn relativerende invloed en zijn overredingskracht weer te gaan zingen, en Ruben en Matthijs, beiden nu ver weg, maar als we bij elkaar zijn is het altijd als vanouds.

De ultieme dank ben ik verschuldigd aan mijn ouders, voor mijn bestaan, maar ook voor de onvoorwaardelijke steun die ze mij altijd hebben gegeven. Jullie vaste zaterdagochtend-telefoontje de afgelopen maanden naar AMOLF, wanneer ik daar zat te typen, was erg goed om me even te 'ontstressen'.

And finally Laura: thank you for all the fun, your love, and your making me see things more optimistically than I sometimes tended to do. As I type this, it is one-thirty in the night and you're sitting at the corridor a couple of offices away typing your own thesis, with one month to go. Finishing together has its good sides and its bad sides... One of the lesser sides were the countless ready-made dinners we microwaved in the AMOLF canteen. But one of the definitely good sides is that we will be ready together to start a new adventure in the States!

Curriculum vitae

Frank Poelwijk was born in Amsterdam, The Netherlands, on January 15th, 1976. From 1988 to 1994 he attended the 'Gymnasium' at the Coornhert Lyceum in Haarlem. In 1994 he started a Master's Program in Physics at the University of Amsterdam, which he combined a year later with a Master's Program in Philosophy. He obtained a Master's Degree in Physics in 2000, after an internship in the group 'Waves in Complex Media' of prof. dr. Ad Lagendijk. In this group he worked on light amplification and interference effects in random lasers. From September 2000 until August 2001 he studied Philosophy at the Université Paris Sorbonne-Paris IV. In November 2001 he started his PhD research in the group of dr. Sander Tans at the FOM Institute for Atomic and Molecular Physics (AMOLF) in Amsterdam. The results of this work are described in this thesis.