

Empirical adaptive landscapes in  
variable environments

Marjon G. J. de Vos



# Empirical adaptive landscapes in variable environments

## PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Technische Universiteit Delft,  
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,  
voorzitter van het College voor Promoties,  
in het openbaar te verdedigen op dinsdag 31 januari 2012 om 12.30 uur  
door

**Marjon Gerarda Johanna DE VOS**

Ingenieur in de Biologie  
geboren te Heerlen

Dit proefschrift is goedgekeurd door de promotor:  
Prof. dr. S. J. Tans

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. S. J. Tans,	Technische Universiteit Delft, promotor
Prof. dr. J. T. Pronk,	Technische Universiteit Delft
Prof. dr. J. Krug,	Universität zu Köln
Prof. dr. S. Bonhoeffer,	Eidgenössische Technische Hochschule Zürich
Prof. dr. B. Teusink,	Vrije universiteit Amsterdam
Prof. dr. N. M. van Straalen,	Vrije universiteit Amsterdam
Dr. C. J. A. Danelon,	Technische Universiteit Delft
Prof. dr. N. Dekker,	Technische Universiteit Delft, reservelid



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 program of the 'Stichting voor Fundamenteel onderzoek der Materie (FOM)', which is financially supported by the 'Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)'.

Nederlandse titel: Empirische adaptieve landschappen in variabele omgevingen.

ISBN 978-90-77209-59-2

A digital version of this thesis can be obtained from <http://www.amolf.nl> and from <http://repository.tudelft.nl>. Printed copies can be obtained by request via email to [library@amolf.nl](mailto:library@amolf.nl).

Cover by Joost van Ingen

© 2011 Marjon G. J. de Vos

Printed by Ponsen & Looijen, the Netherlands

*To my three grandmothers*



## Contents

<b>Chapter 1- Introduction</b>	<b>9</b>
1.1 Before the beginning until the beginning	10
1.2 The evolution of evolutionary thought	10
1.3 Testing evolutionary theories- the Modern Synthesis	13
1.4 The gene as the center of evolutionary thinking	15
1.5 The functional synthesis	16
1.6 Experiments on fitness landscapes	17
1.7 The <i>lac</i> repressor – a model system	19
1.8 This thesis	20
<b>Chapter 2- Revealing evolutionary pathways by fitness landscape reconstruction</b>	<b>23</b>
2.1 Introduction	24
2.2 Intra-genic epistasis	24
2.3 Inter-genic epistasis	27
2.4 Conclusions and perspective	30
<b>Chapter 3- Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape</b>	<b>31</b>
3.1 Introduction	32
3.2 Description of the system	33
3.3 Algorithm	36
3.4 Results	37
3.5 Discussion	43
<b>Chapter 4- Tradeoffs and optimality in the evolution of gene regulation</b>	<b>47</b>
4.1 Introduction	48
4.2 Experimental system and fitness in constant environments	49
4.3 Tradeoffs for fixed expression phenotypes in variable environments	49
4.4 Fitness of regulated phenotypes in variable environments	52
4.5 Competition in variable environments	52
4.6 Evolution under directional selection in variable environments	54
4.7 Genetic basis of local constraint	57
4.8 Molecular mechanism of evolved phenotypes	59
4.9 Discussion	60
4.10 Materials and methods	62

<b>Chapter 5- Environmental dependence of genetic epistasis in the <i>E. coli</i> <i>lac</i> repressor</b>	<b>75</b>
5.1 Introduction	76
5.2 Environmental dependence of sign epistasis	77
5.3 Visualizing higher-order genotype-environment interactions	80
5.4 Higher-order genotype-environment interactions in LacI	81
5.5 Relation between higher-order interactions and system architecture	84
5.6 Conclusions and discussion	87
5.7 Materials and methods	89
<b>Chapter 6- Crossing multipeaked landscapes in variable environments</b>	<b>91</b>
6.1 Introduction	92
6.2 Genotype space for lock-key molecular recognition	93
6.3 Multiple peaks in the <i>lac</i> repressor-operator recognition landscape	96
6.4 Genetic interactions underlying landscape ruggedness	97
6.5 Tradeoffs between expression and repression	98
6.6 Valley crossing in heterogeneous environments	102
6.7 Discussion	103
6.8 Materials and methods	105
<b>Bibliography</b>	<b>107</b>
<b>Summary</b>	<b>121</b>
<b>Samenvatting</b>	<b>123</b>
<b>Dankwoord</b>	<b>125</b>
<b>List of publications</b>	<b>129</b>
<b>Curriculum Vitae</b>	<b>131</b>

# Chapter 1

## Introduction

### 1.1 Before the beginning until the beginning

The first fossil record of unicellular prokaryotic life is laid down approximately 3.5 billion years ago [1], followed by the last eukaryotic common ancestor around 1-1.4 billion years ago [2]. Around 540 million years ago, just before the start of the Cambrian, the earthly environment gradually changed from oxygen poor to oxygen rich [3], during this time the first large, architecturally complex, soft bodied life forms appeared [4,5,6].

During the Cambrian, most major phyla of modern animal life emerged, some of which already resembled today's life forms, including the early vertebrates [7,8,9,10,11]. Whether this 'explosion' into more complex life forms was driven by an environmental change, or whether it was merely a developmental saltation is still under debate [12,13,14].

Early life on land started around 460 million years ago with terrestrial plants that evolved from green algae, in the form of mosses around shallow waters, possibly in symbiosis with the fungi [11]. The first arthropods followed nearly 450 million years ago [15] and flying insects, for instance the first beetles, appeared 270 million years ago [16]. Early primates and birds did not develop until 65 million years ago [17,18], followed much later by the first hominid species of which the fossil record was dated to about 7 million years ago [19,20] (Fig. 1.1).

The genus that includes modern humans, *Homo*, is thought to be about 2.3-2.4 million years old [21]. *Homo sapiens* came into existence 200,000 years ago, whereas civilization has thought to have started around 10,000 years ago. It was shortly after, on geological timescales that is, when man started to investigate its own descent and the origin of other species.

### 1.2 The evolution of evolutionary thought

Until the 18<sup>th</sup> century the common conception in the western biological thinking was that species are unalterable, that they are ordered in a hierarchical ladder of life and need a soul to live. This notion came forth of Aristotles' (384 BC – 322 BC) idea that organisms were arranged in a hierarchical *scala naturae* in which he classified organisms on basis of their perfection, with man being most perfect.

During the Age of Enlightenment, around 1650-1700, the dualistic philosophy advocated by Descartes (1596 –1650) became more popular. He basically separated the notion of body and soul in an organism; the body became a mechanic entity that consists of many parts that each have their own function, much alike a machine.

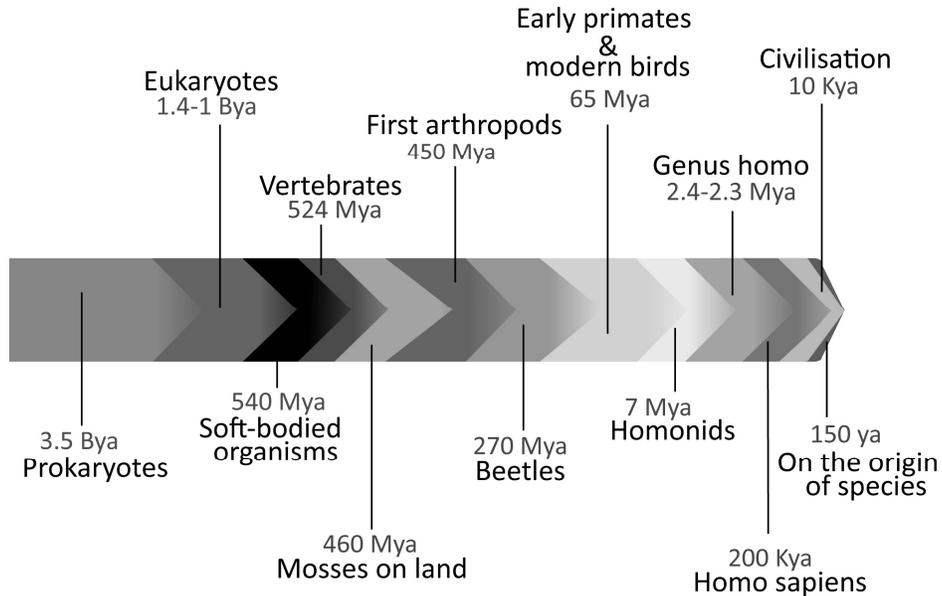


Figure 1. 1 Timeline of evolutionary development. From the first prokaryotes 3.5 billion years ago to On the origin of species, written by Charles Darwin in 1859.

During this period the first naturalists, such as George Cuvier (1769-1832), studied paleontology, in which they compared aspects of alive and extinct organisms. In an article that he published around 1800 [22] he compared the skeleton remains of an elephant and an extinct mammoth. He hereby demonstrated the extinction of species, which was under heavy debate until then. Also the comparative anatomy the naturalists administered, contributed to the increased awareness of the differences between species in natural studies.

The advent of romanticism in the second part of the 18<sup>th</sup> century, in which the study of nature and the cherishing of history became highly appreciated, together with the deducting mentality of the Enlightenment, led to the rise of many theories about nature's origin, from which the theory of evolution, as later written down in Darwin's *On the origin of species* would become the most influential.

Erasmus Darwin (1731-1802), a then well known romantic poet, philosopher, physician, naturalist, and Charles Darwin's grandfather, published his first ideas on evolution in *Zoonomia, the laws of organic life* (1794-1796) [23]. He writes: 'Owing to the imperfection of language the offspring is termed a *new* animal, but is in truth a branch or elongation of the parent; since a part of the embryon-animal is, or was, a

part of the parent; and therefore in strict language it cannot be said to be entirely new at the time of its production; and therefore it may retain some of the habits of the parent-system'. Here he thus introduced the concept of heritability and changes in the offspring.

The other concept of the evolutionary theory, known today as natural selection, by which populations change over time due to the variation between populations and the influence of the environment [24], was in that same time introduced by James Hutton (1726 – 1797), a geologist, in his book *Investigation of the principles of knowledge* (1794) [25].

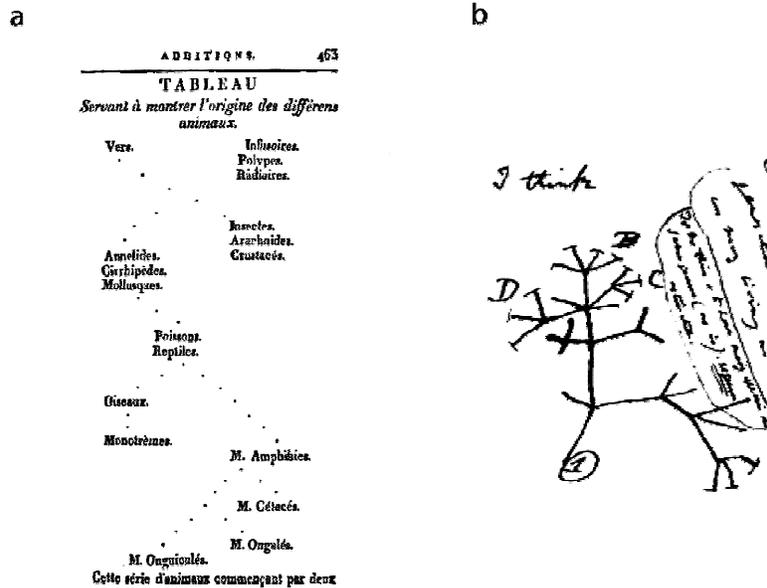
Shortly after, Jean-Baptiste Lamarck (1744-1829) published his *Philosophie zoologique* in 1809 [26], in which he generalized the concept of evolution, as a gradual change in the formation of new forms (Figure 1.2a). Thus at the start of the 19<sup>th</sup> century the first principles of current evolutionary thinking had already been laid down. What was missing was a mechanism by which natural selection, the impact of the environment on a current population, and heredity, the inheritance of traits from parents to their offspring, could be connected.

This connection came in 1859 by Charles Darwin's *On the origin of species* [27]. In this famous book he proposed the theory that populations contain heritable variation that can be selected by natural selection depending on the environmental circumstances (Fig 1.2b). The most important difference between Darwin's and Lamarck's theory is that where Lamarck believed that the *use* of certain traits during the lifetime of an organism determined whether they would be passed on to their offspring, Darwin argued that there was *inherited genetic variation* for traits in the population. Depending on the environment these genetic variants were more or less beneficial or detrimental, which determined the reproductive success of a population. The most fit, the best-adapted to a particular environment, produce the most offspring, and their traits are passed on to future generations.

However, also for Darwin's theory there was a fierce opposition, mostly to the idea that all organisms, including humans, were subject to natural selection. The notion that humans descended from apes aversed a crowd, that had not so much a problem with accepting that other organisms had emerged by natural selection. Thanks to Thomas Huxley's (1825-1895) defense in the famous debate in 1860 in Oxford his ideas found a wider acceptance.

Further, Darwin had not only to compete with different conceptions on the origin of species at that time; he was forced to publish his ideas, due to the peer pressure of Alfred Russel Wallace (1823-1913), who around the same time, also developed his theory of evolution, which was very similar to Darwin's.

They decided to present their work together for the Linnean Society in London, in 1858.



**Figure 1. 2 The First hypothetical evolutionary trees.** a) Drawn by Lamarck, published in his *Philosophie zoologique* [26]. b) Drawn by Darwin, in *On the origin of species* [27].

Thus in the second half of the 19<sup>th</sup> century the time was ripe for evolutionary theories about man's and other species' descent, but there were no experiments to test these hypotheses, yet.

### 1.3 Testing evolutionary theories- the Modern Synthesis

The validation of these evolutionary hypotheses started with the work of Gregor Mendel (1822-1884) which was published in 1866 [28]. His famous study described the independent segregation of traits in garden peas in which 'particular factors' were inherited from parents to their offspring. Although it did not gain much attention then, in 1900 his work was re-discovered and it was married to the concept of natural selection soon after 1930, which started the Modern Synthesis. The Modern Evolutionary Synthesis, as it is also referred to, joins ideas from several biological fields such as genetics, botany, paleontology, ecology and systematic, and describes the then accepted ideas about the theory of evolution. The name was coined by Julian

Huxley (the grandson of Thomas Huxley mentioned above) in his book *Evolution: the modern synthesis* (1942)[29].

In the run-up to the modern synthesis, the advent of population genetics was an important step. Since the heredity 'factors' under selection were not identified yet, it was under debate what these factors could be. Because the discrete 'factors' found in Mendel's experiments seemed to be in contradiction with the gradual evolution proposed by Charles Darwin, the biometrics school under the lead of Karl Pearson (1857- 1936) was in favor of continuous physical variation in populations, in contrast to the discrete 'Mendelian factors'.

In two important publications that Ronald Aylmer Fisher (1890-1962) published in 1918 and 1930 [30,31], he introduced a genetic model that describes the action of discrete genetic characters that can explain the continuous traits observed in nature, on which natural selection could act, favoring the more beneficial alleles over the detrimental ones in populations. John Burdon Sanderson Haldane (1892 –1964) then worked out the effect of natural selection on the changes of these gene frequencies in these populations in a series of ten papers called: *A Mathematical Theory of Natural and Artificial Selection* published from 1924 to 1934. The third of the three big theoretical population geneticists is Sewall Wright (1889 –1988). In addition to various statistical contributions, he introduced a representation of the relationships of genotype, phenotype and fitness as surfaces on a fitness landscape, [32,33], which turned out to be a valuable metaphor for testing evolutionary scenarios.

Central figures in shaping the Modern Synthesis were the experimentalists Theodosius Grigorievich Dobzhansky (1900–1975) and Ernst Walter Mayr (1904-2005). Theodosius Dobzhansky pioneered the genetic work with his experiments with fruitflies, *Drosophila melanogaster*. Besides his famous phrase: 'Nothing in biology makes sense except in the light of evolution' (the name of an essay he wrote in 1973) he is renowned for his *Genetics and the origin of species* (1937)[34]. In this account he validated, by experimental research, the theoretical arguments that genetic changes in populations are an important factor in the evolutionary process. The work boosted the interest in evolutionary studies enormously at that time [35].

Ernst Walter Mayr shed light on how species evolve, in his *Systematics and the Origin of Species* (1942) [36]. He articulated the concept of biological species as groups of organisms that can only breed among themselves, and attributed roles to geography, ecology, and life history for their evolution. This was new then, and in contrast to what Darwin's title *On the origins of species* suggests, untouched terrain. It should, however, be noted that Wallace had already mentioned that species were separated from other species by 'distinctive characters' [37], but he had not formalized it as completely as Mayr did later.

The paleontologists were the first to recognize the concept of extinction of species (see above- George Cuvier), but they were not too eager to accept the concept of natural selection as the main mechanism of evolution. George Gaylord Simpson's (1902-1984) work contributed to their acceptance of this mechanism by publishing his work *Tempo and Mode in Evolution* in 1944 [38] in which he combined the micro-evolutionary thinking at population level to the macro-evolutionary change of species, as observed in fossil records. Note that all this progress is remarkable, given that the nature of the genetic material was not yet elucidated!

#### **1.4 The gene as the center of evolutionary thinking**

The discovery of DNA as the inherited genetic material by Alfred Hershey and Martha Chase in 1952 [39] and the resolution of the DNA crystal structure by James Watson, Francis Crick [40] and Rosalind Franklin in 1953 evoked the development of a gene-centric view of evolution, in which the gene, in contrast to the organism, is the objective for natural selection to act on.

Interestingly already before these discoveries, the terms genotype, genetic content, and phenotype, a type of trait or character, like for instance wing color in moths, were coined by Johannsen in 1909 [41,42]. Also the relationship between genes and proteins, in the one-gene-one-protein hypothesis was already proposed by Beadle and Tatum in 1941 on mutants of the bread mold *Neurospora crassa* [43] and although it later was found to be an oversimplification, it was a major step forward for evolutionary thinking.

That the genotype could interact with the environment, such that phenotypically plastic organisms could arise was already known for some time [44]. However whether this genotype by environment interaction was the direct consequence of a character that was independent of trait means, and thus a separate plasticity factor, or whether the organism's response to the environment was the consequence of a direct interaction with the distinct character, in the absence of separate plasticity factors was unclear [44,45]. Scheiner and Lyman tested the hypothesis "that plasticity is due to genes that determine the magnitude of response to environmental effects which interact with genes that determine the average expression of the character" [46,47,48], which they called the "epistasis model" [47].

Their results supported the model that the plasticity of a character is determined by special plasticity factors. They described the effect of these factors as overdominance, pleiotropy and epistasis. Epistasis appears when the effects of one trait are modified by one or several other *loci*, which is in contrast to overdominance, where several *alleles* act on the same trait. Pleiotropy occurs when one allele or locus affects multiple phenotypic traits. Note that in this thesis we use the definition of

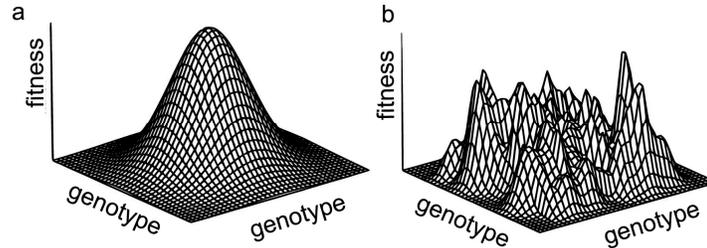
epistasis for the genetic background dependent effect of both *alleles* and *loci* on phenotype or fitness, we thus do not discriminate between overdominance and epistasis as described above.

The British Richard Dawkins (1942), now mainly a writer of popular-scientific books, is a strong advocate of the gene-centric view of evolution, and popularized this conception to the common public by his 1976 book *The Selfish gene* (1976) [49]. His American opponent on this view was Stephan Jay Gould (1941-2002). He believed that natural selection acts on combinations of genes, on their whole genome, thus on organism level. In his book *The wonderful life: The Burgess Shale and the Nature of History* (named after a geological formation in which many Cambrian species are found), he furthermore claimed that the evolution of species is highly unpredictable [7]. In contrast, Simon Conway Morris (1951), who actually worked as a paleontologist on the fossils from this Burgess Shale, claims that evolution is convergent and thus that the evolution of species is predictable, and subject to chance processes [50]. These contradictions are not fully resolved until today.

### **1.5 The functional synthesis**

The field of evolution and molecular biology became more and more specialized in the 20<sup>th</sup> century, which led to a divergence of the fields. In the last decade of the 20<sup>th</sup> century, it was time for re-marrying the fields investigating evolution and the field of what had become molecular biology. Such that 'today, the reductionist culture of molecular biology is being fused with the historical realism of evolutionary biology, creating a new functional synthesis' as Anthony Dean and Joseph Thornton put it in their review describing the current status of the field [51]. In this field the techniques of evolutionary research, such as statistical measures of functions and phylogenetic analysis are combined with the tools of molecular biology or biochemistry, for the (re-)construction of mutants, and the study of their functional performance *in vitro* or *in vivo*. The key innovation of the field is that it can assess the functional relationship between mutant genotypes and their corresponding phenotype or fitness in an organism, without altering other aspects that affect function, such as the environment or other changes of the genetic background, which is virtually impossible 'in the wild'. This can elucidate many questions that were previously difficult to tackle such as how do mutations affect function, does the functionality of mutants depend on their genetic background, and what is the effect of both of these questions on the evolutionary progress?

An interesting example that Dean & Thornton mention in their review and which addresses the former two questions, uses a combination of the techniques of



**Figure 1.3 Fitness landscapes from [56].** a) In a single peaked fitness landscape, the fitness peak is selectively accessible. b) In a multi-peaked fitness landscape evolution is constrained by epistasis, such that evolving populations can undergo stasis at suboptimal peaks.

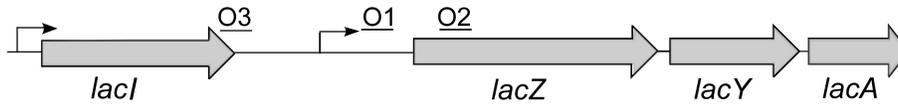
evolutionary biology, biochemistry and molecular biology is the reconstruction of the ancestral opsin, a visual pigment found in the eye that is necessary for sight. Changes in the amino acid residues in this protein alter the sensitivity to different wavelengths, such that different colors of light can be perceived by different proteins. Phylogenetic analysis has revealed that in rhodopsins, a special class of opsins used for dimlight vision, an evolutionary change from green light sensitivity to blue light sensitivity was brought about by only a few amino acid substitutions in the protein. By replacing the 'new' residues with the 'old', a change back in time and function was established by Yokoyama and co-workers, such that the more recent version of the protein, with the 'new' key amino acids residues replaced by the 'old' was sensitive to green light again [52,53,54].

This example beautifully demonstrates the elucidated relationship between history and function by the combined usage of evolutionary, biochemical and molecular biological techniques. However, it does not address the effect of the genetic substitutions on adaptive trajectories, and thus the evolutionary progress. The fitness landscape metaphor originally introduced by Wright in 1932 and since then heavily discussed on a theoretical level [33,55,56,57,58,59] was found useful for the attribution of this question and was therefore adopted by the functional synthesis.

### 1.6 Experiments on adaptive landscapes

There are a few ways in which the fitness landscape metaphor is used. One is a description of the gene frequencies present in a population, which is mainly used in theoretical studies. The other is the representation of genetic combinations and their effect on phenotype or fitness, or phenotypic traits and their effect on fitness [59].

It is this last use of the fitness landscape metaphor that is used in the functional synthesis. In this representation the combination of genetic substitutions

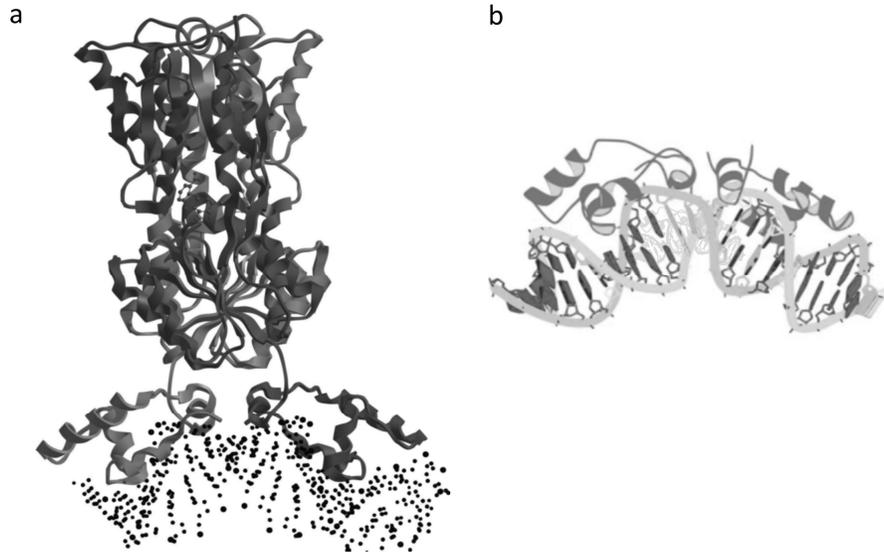


**Figure 1. 4 The wild type *lac* operon, regulated by the *lac* repressor LacI.** LacI regulates the *lac* operon by binding to the O1 and O2 or O3 operators, in the absence of inducer. In the presence of inducer LacI is released and the operon is expressed. The downstream genes can be used for the metabolism of lactose.

visualizes the accessibility of adaptive trajectories. When the fitness landscape is single peaked, fitness increases monotonously with each mutational step until the adaptive peak is reached (Figure 1.3a). However, when the adaptive landscape is multi-peaked, and therefore rugged, the global optimum might be constrained and the adapting population might get stuck on a local sub-optimum (Figure 1.3b).

So far, the systematic mapping of mutational trajectories and their relationship to phenotype or fitness has only yielded single peaked fitness landscapes [60,61,62,63]. However, there have been indications for multi-peaked landscapes, which are underlain by multiple genetic interactions [64,65,66,67,68,69,70]. Interestingly, in a study that involved the mapping of 120 trajectories comprising five mutations on a single peaked landscape towards a TEM  $\beta$ -lactamase with increased cefotaxime resistance, not all genetic substitutions increased resistance [71]. The presence of so-called sign-epistasis [63] was responsible for the fact that in certain genetic backgrounds specific mutations were detrimental, whereas in other genetic backgrounds, the same mutations conferred a resistance increase. Since in all genetic backgrounds at least one mutation increased resistance, this landscape was single peaked for this resistance function. On multi-peaked landscapes, evolutionary trajectories are underlain by genetic interactions that show reciprocal sign epistasis [72] (Figure 1.3b), such that there are no direct adaptive paths from a low fitness genotype to the genetic combination with the highest fitness. Recently Kvittek and Sherlock [64] found in an experimental evolution study that the presence of specific alleles of MTH1 and HXT6/HXT7 in yeast are mutually exclusive, which leads to a valley in a reciprocal sign epistatic motif. Another study by de Visser *et al.* [70] found such a mutual exclusivity for, for example, mutations that led to arginine and pyrimidine deficiency. However, until now, there are no studies that show the systematic mapping of a rugged fitness landscape.

These studies discussed before were all undertaken under constant environments, but evolution mostly occurs in heterogeneous environments. To understand the adaptive potential of populations in heterogeneous environments, it is necessary to know the shape of fitness landscapes in variable environments.



**Figure 1. 5 The *E. coli lac* repressor binds to the operator DNA.** a) The *lac* repressor is a dimer of dimers. One dimer is the unit of binding to one operator DNA (pdb 1EFA). b) Two key amino acid residues in the *lac* repressor protein and four base pairs in the operator DNA are responsible for specific and tight operator-DNA binding (part of pdb 1LBG).

A recent experimental study that addressed the mapping of mutations of the same TEM  $\beta$ -lactamase in two environments showed that alleles have a different mutational effect in different environments [73], which gives the valuable insight that evolution in variable environments cannot be extrapolated from a evolutionary studies investigating a single environment. That adaptive landscapes can be used for the detection of constraining motifs in evolution (this thesis), demonstrates that it is more than a simple metaphor. It can contribute to our understanding of evolution, which could ultimately lead to the predictions of the evolutionary potential of organic systems (see chapter 4).

### 1.7 The *lac* repressor- a model system

In this thesis we use the *lac* repressor gene in the bacterium *Escherichia coli* as a model system to study evolutionary questions. The *lac* repressor, LacI, has been the model of gene regulation in bacteria since its discovery 50 years ago [74]. It regulates the expression of three genes in one operon: *lacZ*, *lacY* and *lacA* (Figure 1.4), which are involved in the metabolism of lactose.

Its regulatory function can be separated in two functions in two environments: binding to the operator DNA in the absence of lactose, thereby preventing the cost of spurious protein expression [75,76]. In the presence of the inducer allo-lactose or IPTG, the *lac* repressor releases the operator DNA, which leads to the expression of the downstream genes, thereby allowing *E. coli* to use lactose for growth. This binding and unbinding of the *lac* repressor to the DNA is mediated by an allosteric conformational change of the protein, caused by the binding of the inducer to the protein.

LacI is a homo-tetrameric protein, of which the dimer is the functional unit for DNA binding (Figure 1.5a). The *lac* operon contains two operators, in addition to the main operator O1 [77]. These two auxiliary operators, O2 and O3, together with the main operator O1 are involved in looping of the DNA by the tetrameric LacI, which results in a highly stable protein-DNA complex that prevents transcription from the downstream genes in the operon [78,79].

LacI recognizes the operator DNA by the binding of amino acids of the second helix in the N-terminal HTH-motif of the headpiece of LacI to base pairs in the major groove of the operator DNA (Figure 1.5b) [80,81,82,83,84]. The combination of amino acids in LacI, together with base pairs in the O1 operator are important for specificity and affinity of binding.

Since the *lac* system responds to the environment and gene regulation by LacI is well-studied, it is a perfect model system for experimental evolutionary studies in multiple environments. In our studies we have only made changes in LacI, or LacI and the O1 operator while keeping the other parts of the microbe in its original state (as far as one can say this for a laboratory strain), by which we could observe the effect of the isolated genetic changes in multiple environments on function (phenotype) and fitness.

## **1.8 This thesis**

This thesis contains six chapters, in which each chapter deals with a particular part of the question revolving around the genetic background dependent effect of mutations, epistasis, and the effect of mutations in variable environments on evolution. Chapter 2 discusses the effect of genetic interactions on the evolutionary accessibility of adaptive trajectories. The genetic background dependent effect of mutations, epistasis, is recognized for its role in adaptive evolution. Due to novel experimental methodologies it has been possible to experimentally determine epistatic interactions, and their effect of phenotype and fitness. With these tools, long standing questions in evolutionary biology can be discussed, such as the shape of fitness landscapes, and the origins of robustness and modularity. In this chapter we

have reviewed and discussed the literature that answers some of these long standing issues, and we have posed some new questions.

Chapter 3 focuses on the constraints induced by epistatic interactions in an empirical adaptive landscape. We used data from a previously reconstructed genotype-phenotype landscape based on *in vivo* measurements of *lac* repressor and operator mutants in *E. coli*. Interactions between two key amino acid residues in the repressor, and four base pairs in the operator determined the genotype-phenotype relation. We found this landscape to be multi-peaked, all direct evolutionary trajectories between peaks contained significant drops in phenotype, that form valleys in the adaptive landscape. Consistent with earlier predictions, we found reciprocal sign epistatic interactions in trajectories between the peaks to constrain the adaptive progress.

Chapter 4 describes the evolution of the *E. coli lac* repressor to the predicted optimum under two adverse environments. Based on measured tradeoffs in fitness of the *wild type lac* repressor in two environments, we have reconstituted the phenotype-fitness landscape. From this landscape we predicted that the optimal adaptive phenotype in these environments would be functionally inverted. We show that genetic constraints impact the trajectory to this functional inversion, but that selection eventually determined the outcome.

Chapter 5 deals with the effect of variable environments on the accessibility of adaptive genetic solutions. Previous studies have described constraints on adaptive trajectories in terms of epistasis. Here, we describe the higher-order interactions between genotype and environment, and their effect on the evolutionary accessibility of adaptive trajectories. We use functionally inverse *lac* repressor mutants from chapter 4, and their reconstructed adaptive intermediates as a case study.

Chapter 6 shows that adaptive solutions that are constrained in sequence space can be reached with the aid of environmental perturbations. We have reconstructed the genotype-phenotype landscapes spanning the sequence space between two *lac* repressor variants. These variants differ by six mutations, two key amino acid residues in the repressor and four base pairs in the operator. We have assayed the genotype-phenotype landscape in two environments, and found the landscape to be multi-peaked in both environments. In addition, we found that all 720 trajectories, from the suboptimal variant to the optimum, contain tradeoffs. By alternating the fixation of one mutation in one environment, by another mutation in the other environment, temporal environmental perturbations allow a large number of adaptive trajectories to walk over the multi-peaked landscapes, until the global optimum was reached.



## Chapter 2

# Revealing evolutionary pathways by fitness landscape reconstruction

*The concept of epistasis has since long been used to denote non-additive fitness effects of genetic changes and has played a central role in understanding the evolution of biological systems. Owing to an array of novel experimental methodologies, it has become possible to experimentally determine epistatic interactions as well as more elaborate genotype-fitness maps. These data have opened up the investigation of a host of long-standing questions in evolutionary biology, such as the ruggedness of fitness landscapes and the accessibility of mutational trajectories, the evolution of sex, and the origin of robustness and modularity. Here we review this recent and timely marriage between systems biology and evolutionary biology, which holds the promise to understand evolutionary dynamics in a more mechanistic and predictive manner.*

## 2.1 Introduction

Genomic sequencing has generated a wealth of information on the molecular basis of organisms and their evolutionary relationships [85,86]. However, we remain largely ignorant about the interactions between genes that are central to organismal functions and phenotype [87]. Information on how phenotypes depend on these interactions is not only relevant for understanding the architecture of cellular functions [88,89,90] but also has profound implications for their evolutionary origin [51,61,91,92].

The concept of epistasis provides an elementary description of genetic interactions that are involved in function or fitness [33,93,94]. About 100 years ago, William Bateson introduced this term to describe phenotypic deviations from Mendelian segregation ratios due to genes masking the effects of others [95]. Broadly defined, epistasis denotes cases in which the effect of a mutation depends on the genetic background in which it occurs [96]. For instance, a mutation that is beneficial in one genetic background can be neutral or deleterious in another. Epistasis has played a central role in many evolutionary theories, including those that address speciation [97], the evolution of sex [98,99,100,101,102] and adaptation [32,96,103,104].

Owing to novel experimental approaches, epistatic interactions can now be studied in an increasingly systematic manner. These efforts address a diverse array of biological systems and scientific questions. Some have a predominantly functional perspective, while others are motivated by evolutionary questions. In some cases the focus is on epistatic interactions within a single gene (intra-genic epistasis), while other studies consider larger networks of interacting genes (inter-genic epistasis). The aim of this review is to sketch these recent developments, by giving a number of illustrative examples of these diverse directions. In doing so, we hope to provide an overview of the current possibilities and limitations, and to identify new questions within this exciting new field.

## 2.2 Intra-genic epistasis

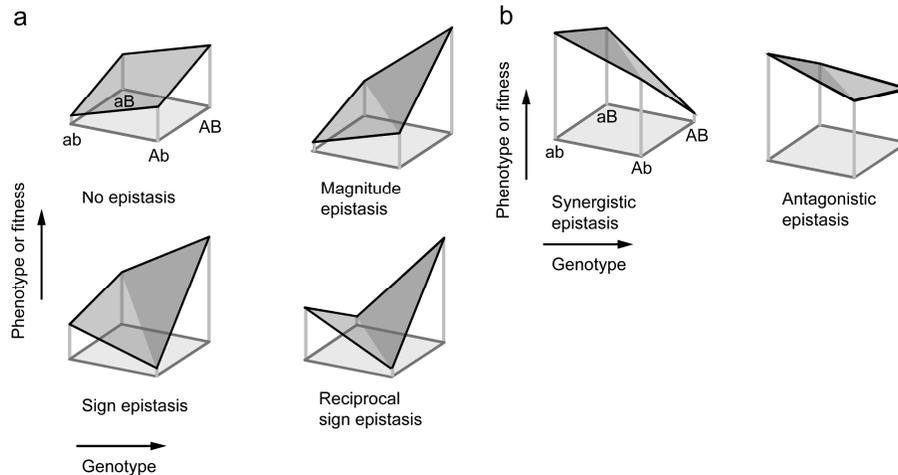
The nature of the epistatic interactions within a biological system is intrinsically linked to its evolutionary origin and potential. For instance in the absence of epistatic interactions between loci, genetic changes at these loci contribute independently to fitness (Fig. 2.1a). Note that for a Malthusian fitness parameter such as the bacterial growth rate, independence implies that fitness effects are additive, while when fitness is defined as the number of offspring, the individual contributions

to fitness multiply. Starting from a sub-optimal phenotype, all adaptive trajectories towards the optimum then rise monotonically in fitness. Consequently, these trajectories are all equally probable to be followed during adaptation. In contrast, when required genetic changes exhibit sign-epistatic interactions (Fig. 2.1a), some trajectories contain fitness decreasing steps, making them much less probable, though other trajectories do exhibit monotonously increasing fitness. We note that any type of epistatic interaction (not only sign epistasis) will result in some difference in the probabilities for different paths. Fitness landscapes are an intuitive concept to consider multiple possible trajectories between two points, and in both previous cases the landscape is smooth and singlepeaked. However, landscapes can also be rugged and have more than one single fitness peak. Adapting from one peak to the other then requires two or more simultaneous genetic changes, which is denoted as reciprocal sign epistasis (Fig. 2.1a). What are the shapes of actual fitness landscapes? This longstanding question is now starting to be addressed. Weinreich and colleagues focused on the protein  $\beta$ -lactamase in *Escherichia coli*. One variant was known to confer resistance to penicillin, while adding five mutations conferred resistance to the newer antibiotic cefotaxime.

To gain insight in all possible mutational trajectories between these two variants, all  $2^5 = 32$  possible intermediates were constructed and assayed on survival ability in cefotaxime, which is taken as a measure of fitness. The data revealed that a majority of the trajectories contained fitness decreasing or neutral steps, resulting in much reduced chances of being followed by natural selection [71].

Sign-epistatic interactions underlie these landscape features. For example, Gly238Ser in wild-type background increases the resistance, even though it increases protein aggregation by lowering the thermodynamic stability. This loss of stability is rescued by Met182Thr which alone modestly reduces resistance. Paths that fix Gly238Ser before Met182Thr are therefore plausible, but the reverse order is not. Such a balancing between functional and structural benefits is a more general evolutionary mechanism [105,106] and provides a mechanistic rationale for sign epistasis [51,96].

One may consider epistasis in fitness or at the functional level. The relation between the two was investigated using isopropylmalate dehydrogenase (IMDH) as a model system. IMDH is involved in biosynthesis of the amino acid leucine, and uses the coenzyme nicotinamide adenine dinucleotide ( $\text{NAD}^+$ ) as a hydride acceptor during an oxidative decarboxylation. Upon six mutations IMDH exchanges this coenzyme for another, nicotinamide adenine dinucleotide phosphate ( $\text{NADP}^+$ ) which is also used by a highly divergent paralog isocitrate dehydrogenase (IDH) [91].



**Figure 2. 1. Two classifications of epistatic interactions.** a) Paths composed of two mutations are considered, from an initial sequence “ab” towards the optimum sequence “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 may differ in magnitude, but not in sign. For sign epistasis, the sign of the fitness effect changes. Consequently, some paths become inaccessible. Finally, such a change in sign of the fitness effect can occur for both mutations; this is denoted as reciprocal sign epistasis, and is required for having multiple peaks in the fitness landscape. b) Two types of epistasis that distinguish possible interactions between two genes. Paths are considered from the initial optimal sequence “ab” towards the double knockouts “AB”. If the fitness effect of the double knockout is larger than expected from the sum of their individual effects is denoted as synergistic epistasis, while a smaller than additive effect is termed as antagonistic epistasis.

The construction of a large number of IMDH intermediates, and analysis of their enzymatic activity *in vitro*, showed a lack of epistasis: all investigated mutations contributed roughly additively to the enzymatic activity. Thus, the genotype–phenotype relation for IMDH coenzyme use is a single featureless peak [61].

Assays of the same mutants *in vivo*, in which the corresponding growth rates were measured, yielded insight into the relation between genotype and fitness. Within this landscape, many – but not all – mutational trajectories from NADP<sup>+</sup> to NAD<sup>+</sup> usage exhibited a fitness dip, which indicates sign epistasis. This introduction of epistasis can be understood from the nonlinear relation between enzymatic activity and growth rate, in combination with the competition between two coenzymes: the fitness decreases when the recognition of NADP<sup>+</sup> is broken down, and rises again when NAD<sup>+</sup> recognition is built up. However, because apparently some mutations

exist that simultaneously decrease NADP<sup>+</sup> interaction and increase NAD<sup>+</sup> interaction, monotonously increasing trajectories are also possible. Thus although the genotype-phenotype map that depends on the functional activity of an enzyme can be free of epistasis, the corresponding phenotype-fitness map can contain (sign) epistatic features.

### 2.3 Inter-genic epistasis

Epistatic interactions can occur between different genetic components when they are functionally related within a network. An elementary example is the recognition between a transcription factor and its binding site within a regulatory region. Which types of epistasis underlie such a molecular recognition and which landscape features they present, was investigated using the *lac* repressor and operator as a model system [67]. Earlier mutational analysis had shown that two repressor residues and four base-pairs in the operator were central to altering the specificity of binding [83]. A large set of mutants with substitutions at these loci had been assayed on their repression value, the ratio between repressed and unrepressed expression of the controlled gene. This genotype-phenotype relation exhibited several distinct peaks: optimal repressor-operator combinations whose paths between them contained significant decreases in repression [67]. The existence of multiple peaks cannot be explained by sign epistasis alone, but requires the more severe reciprocal sign epistasis (Fig. 2.1a). An intuitive rationale for this type of epistasis is the following: mutating one binding partner likely only benefits a new interaction if the other binding partner is mutated first and *vice versa*. The alternative outcome, in which optima are bridged by a “master key” repressor that binds to multiple operators, was not observed in the data.

As the repressor-operator binding landscape contains many sub-optima, they would seem to act as evolutionary traps that hamper adaptation to the global optima. One phenomenon that may affect this outcome is the presence of duplicated genes and their mutational divergence. *E. coli* contains several homologs of the *lac* repressor, which regulate the expression of different operons independently and must therefore have evolved specific repressor-operator recognition. Indeed, the promiscuous binding of functionally unrelated transcription factors on the *lac* operator should provide a selective pressure on binding specificity. This issue was investigated using the *lac* genotype-repression landscape. It was found that for two identical repressor-operator pairs, many mutational trajectories now did not exhibit decreases in fitness. Mutations appeared to exist for which the obligatory decrease in repression for one repressor was offset by a simultaneous decrease in the penalty for promiscuous binding of the other repressor. Such compensations within biochemical networks are

ubiquitously observed and may therefore be a more general adaptive evolution mechanism [107]. These results also substantiate the suggestion that the robustness of networks may promote evolvability [108,109].

Within the larger web of biological interactions, epistasis originates not only by direct physical recognition, but also through their hidden functional relations. Elena and colleagues explored this issue by creating random single and double mutants in *E. coli*, using Tn10 transposition and P1<sub>vir</sub> transduction [110]. Epistasis is absent if the fitness effect of the double mutant equals the sum of the fitness effects of the single mutants. However, a larger than additive effect of the double mutant is denoted as synergistic epistasis (Fig. 2.1b), while a smaller than additive effect indicates antagonistic epistasis (Fig. 2.1b). The data showed that epistasis between these random knockouts is not rare, and that antagonistic and synergistic interactions occur almost at the same frequency [86,110]. This observation is also supported by experimental studies from other organisms [111,112,113,114,115,116] and by computational modeling [117,118].

The conservation of epistatic interactions across different organisms was addressed by Tischler and colleagues. This study focused on synthetic lethal interactions, a phenomenon in which two non-lethal mutations yield a lethal phenotype when combined. A large number of synthetic lethal interactions between genes are known for *Saccharomyces cerevisiae*, some of which have orthologous pairs in *Caenorhabditis elegans*. These 837 pairs were assayed in *C. elegans* on fitness using RNA interference (RNAi). It was found that a maximum of 5% of synthetic lethal interactions are conserved between *S. cerevisiae* and *C. elegans*. This value is low compared to the conservation of protein-protein interactions (31%) [119]. Thus, surprisingly, even though the gene function between worm and yeast are conserved, the epistatic interactions are not.

The aspect common to the above studies is that they assess epistasis through measurement of fitness. As the molecular mechanisms that underlie fitness are often poorly understood, inference of epistasis from fitness alone usually does not provide a mechanistic explanation of epistasis. Epistasis can also be detected at lower levels of organization, such as transcription, for which the underlying mechanisms are better understood. One study focused on two glucose-adapted lines of *E. coli* [120] and the ancestral line, as well as *crp* knockouts for each of these lines. *crp* is a key global regulatory gene that itself was not altered in the glucose adaptation [121]. Deletion of the *crp* gene appeared more detrimental to growth in the evolved strains compared to the ancestor. The cause of this effect was found in the adapted expression profile: in the ancestral strain, *crp* controls the expression of 171 genes whereas in the evolved strains it controls an additional 115 genes that were not initially *crp*-dependent [122].

This study highlights the importance of lower-level phenotypic characterization when aiming to understand the mechanisms that underlie epistasis.

Epistasis is also thought to be related to the occurrence of sexual recombination. In the 1980s, Kondrashov introduced the mutational deterministic hypothesis, which states that the deleterious mutations that are combined during recombination are purged from the population [98]. This scenario requires synergistic epistasis, as it makes combined mutations more harmful than expected from their individual effects. The relation between epistasis and sexual reproduction has been investigated in a computational study of networks of transcriptional regulators as found in *Drosophila melanogaster* [123,124]. Individuals who attained stable expression patterns upon variation of their network interactions and selection were considered viable, whereas individuals that failed to reach equilibrium were considered not viable. The simulations showed that synergistic epistasis, and accordingly mutational robustness, indeed increased within the network in a sexually reproducing population while it did not in asexual populations. This effect is more prominent especially if the population experiences a high mutation rate and there are many genetic interactions. It remains to be seen whether these results in turn also provide a rationale for the maintenance of sex despite its costs.

The relation between epistasis and modularity was addressed in a computational study of *S. cerevisiae* metabolism. By deleting single and pairs of genes *in silico* and computing the resulting growth rate using flux balance analysis, epistatic interactions could be assessed. Between two functional modules, such as those responsible for respiration and glycolysis, significant epistatic interactions were observed. These epistatic interactions were either consistently synergistic or consistently antagonistic. On the other hand, within a functional module both types of epistasis were observed. These results point to a correlation between network architecture and epistasis, and challenge the common notion that epistasis is stronger within a functional module than between the modules [125,126].

In addition to epistatic interactions between genotypes, interactions between genotype and environment are also common. In this case, the effect of a mutation depends on the environment in which fitness is assayed and vice versa. To investigate genotype by environment interactions, random insertion mutants of *E. coli* were assayed in four different environments. It was found that about 40% of the insertions yielded different fitness effects in the different environments, showing that genotype by environment interactions are common [127].

## 2.4 Conclusions and perspective

The studies reviewed here illustrate that epistasis plays a central role in a broad array of systems, scientific questions and experimental methodologies. The questions range from being predominantly functional, where epistasis is a powerful tool to unravel functional relations between genes, to mainly evolutionary questions, where epistasis provides a mechanistic explanation or even a prediction of adaptive dynamics. Some research focuses on interactions between residues within a single protein, while others consider the full regulatory networks underlying organismal fitness. These diverse approaches have provided an intriguing insight into hidden correlations within the design and evolutionary potential of biological systems.

However, the current approaches have considerable limitations and only begin to scratch the surface of all relevant correlations. For instance, intra-genic studies have so far only explored base-pair substitutions within restricted parts of genotype space. One may argue that organisms can exploit genetic changes in other regions to achieve the same functionality, thus making observed constraint and epistasis irrelevant. However, phylogenetic analysis shows that evolutionary transitions similar to several studied examples have in fact occurred in evolutionary history. Larger scale genomic rearrangements such as recombination are also known to play an essential role in the evolution of some functions. It would be of interest to explore how recombination events as well as ploidity affect fitness landscapes and the evolutionary trajectories within it. On the other hand, studies of epistasis within larger networks do explore more distant interactions, but in return typically only consider knockouts as the genetic change, and lethality as the response.

It would be fruitful to investigate more detailed changes, such as altered expression levels or point mutations resulting in novel functionality, as well as more quantitative fitness measurements.

## Chapter 3

# Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape

*Insight into the ruggedness of adaptive landscapes is central to understanding the mechanisms and constraints that shape the course of evolution. While empirical data on adaptive landscapes remain scarce, a handful of recent investigations have revealed genotype-phenotype and genotype-fitness landscapes that appeared smooth and single peaked. Here, we used existing in vivo measurements on lac repressor and operator mutants in Escherichia coli to reconstruct the genotype-phenotype map that details the repression value of this regulatory system as a function of two key repressor residues and four key operator base pairs. We found that this landscape is multipeaked, harboring in total nineteen distinct optima. Analysis showed that all direct evolutionary pathways between peaks involve significant dips in the repression value. Consistent with earlier predictions, we found reciprocal sign epistatic interactions at the repression minimum of the most favorable paths between two peaks. These results suggest that the occurrence of multiple peaks and reciprocal epistatic interactions may be a general feature in coevolving systems like the repressor-operator pair studied here.*

It has long been recognized that the evolution of new functions is not only determined by the external forces of natural selection, but also by diverse internal limitations of the evolving biological system itself. Apart from the hard constraints imposed by physical and chemical laws, the Darwinian process of repeated selection of heritable changes can also give rise to adaptive limitations when some of the genetic changes that are required to reach a more adapted genotype are not unconditionally favorable. One of the most striking situations arises when no single genetic change is favorable while combinations of multiple genetic changes are, as it can lead to evolutionary stasis. This scenario can be seen as an entrapment in a local optimum in a multi-peaked adaptive landscape. While in recent years methodologies have been developed to determine such adaptive landscapes empirically, the evidence for the existence of multiple peaks have been rather scarce and indirect. Here we analyze published experimental data on the expression regulation of mutants of the *lac* repressor and operator. We report the presence of multiple peaks in repression, as the key residues and base pairs for the binding specificity are varied in the transcription factor and its target DNA binding site. Together with our finding, the existence of multiple homologous repressor-operator pairs in *Escherichia coli* indicates that evolution has been able to avoid the frustration associated with local suboptima, and exploits the wide range of solutions available in the genetic space despite the presence of genetic barriers.

### 3.1 Introduction

Determining the architecture of adaptive landscapes is central to understanding the course of evolution. The stepwise adaptation of living systems to new environments by natural selection results from the intricate relationships between genotype and phenotype and between phenotype and fitness [65]. Ever since Wright [33] introduced the metaphor of an adaptive landscape, its shape has been hotly debated, but nonetheless remained essentially unknown due to insufficient empirical data [58,61,63,65,93,96,128,129,130,131].

The architecture of adaptive landscapes is tightly related to the notion of epistasis (Fig. 3.1) [33,132]. Epistasis provides a way to classify how elementary genetic changes correlate in terms of their effect on phenotype and fitness. For magnitude epistasis or in absence of epistasis, mutations give rise to either a positive or a negative fitness or phenotypic effect, regardless of the genetic background (Fig. 3.1a, top). This results in adaptive landscapes that are smooth and single peaked (Fig. 3.1b, left). In the case of sign epistasis, the sign of the fitness or phenotypic effect of a

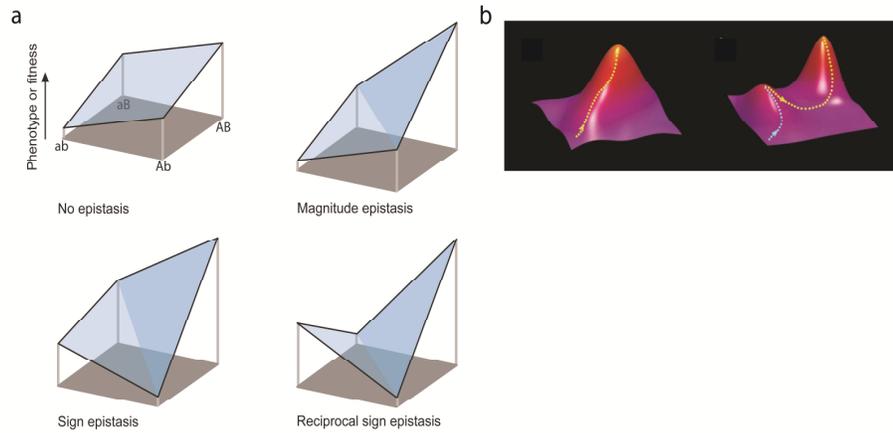
mutation does depend on the genetic background (Fig. 3.1a, bottom left), such that only a fraction of the total paths to the optimum are selectively accessible, i.e., contain only steps that confer a performance increase. A third class of epistatic interactions is that of reciprocal sign epistasis, in which two genetic changes are individually deleterious but jointly advantageous (Fig. 1a, bottom right). It has been suggested that reciprocal sign epistatic interactions play a central role to generate adaptive landscapes with multiple distinct peaks (Fig. 3.1b, right) [96]. The occurrence of multiple peaks can give rise to entrapment on local suboptima, which frustrates adaptation to the global optimum.

Spurred by systematic laboratory reconstructions of the evolutionary intermediates for a handful of well-characterized phenotypes, recent years have seen a renewed interest in the structure of adaptive landscapes [61,71,96,131]. These efforts have revealed the existence of sign epistatic interactions and single peaked landscapes. Here we investigate the structure of the genotype-phenotype landscape for the repression of the *lac* operon by the *lac* repressor and its operator. Using *in vivo* measured data from Müller-Hill and co-workers [83], we have previously reconstructed this landscape to investigate the divergence between two repressor-operator pairs [67]. Here, we aim to determine whether the repression value for a single repressor and operator exhibits more than one distinct peak as a function of its genotype.

In this effort, we developed a recursive algorithm to search for peaks within the genotype space of the repressor-operator system. This analysis showed that the genotype-phenotype landscape is multi-peaked, encompassing in total 19 well-defined optima. Our result contrasts with previous studies that showed single peaked adaptive landscapes [61,71] which we suggest may be understood from the lock-key architecture of the here studied system. This finding, together with the observation that several repressor-operator pairs homologous to the *lac* system exist in *Escherichia coli*, suggests that evolution is able to overcome the frustration of a multi-peaked landscape to exploit a wide diversity of interactions that are available in the genetic space.

### 3.2 Description of the system

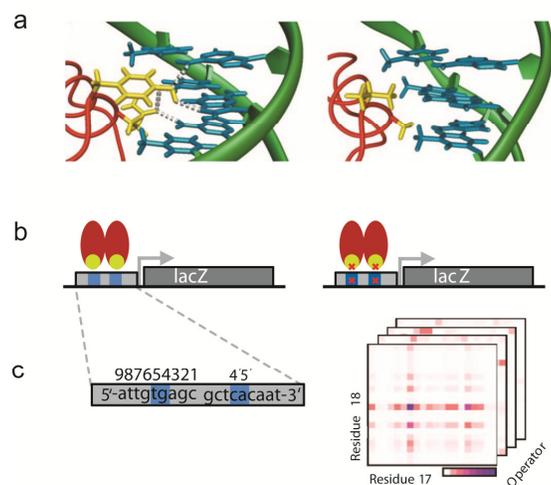
Recognition of DNA by proteins plays a central role in the regulation of transcription in all organisms. The lactose operon of *Escherichia coli* serves as a key biological system to study gene transcription regulation ever since Monod and Jacob [74] discovered it.



**Figure 3.1 Relationship between epistasis and landscape ruggedness.**

a) Schematic representation of different classes of epistatic interactions between mutations at two different genetic loci:  $a \rightarrow A$  and  $b \rightarrow B$ . In the absence of epistasis, mutation  $a \rightarrow A$  yields the same phenotypic or fitness effect in genetic backgrounds  $b$  or  $B$  and vice versa. With magnitude epistasis, the phenotypic or fitness effect differs in magnitude depending on the genetic background. For sign epistasis, the sign of the fitness effect depends on the genetic background; as a result some paths are selectively inaccessible. In the case of reciprocal sign epistasis mutations are individually deleterious but collectively advantageous. b) Continuous surfaces that serve to illustrate ruggedness in fitness landscapes. The left panel shows a single peaked surface where all the paths toward optimum are monotonously increasing in fitness. The right panel shows a multi-peaked surface. All paths from the suboptimal peak to the optimal peak encounter a decrease in fitness.

Transcription regulation of this operon by binding of the *lac* repressor (LacI) to its operator regions in the *lac* promoter (Fig. 3.2) is understood in great detail and continues to be of great value in the study of gene regulation [78,82,84,133,134,135,136,137,138,139,140]. LacI is a prototypic member of the large GalR-LacI family of prokaryotic transcription factors, a group that has more than 1000 members [141,142]. Members of this family possess a conserved N-terminal DNA binding domain, and a central highly versatile domain that, under the same scaffold, functions as a binding pocket for different types of small signaling molecules and promotes oligomerization of the complex by protein-protein interaction between the monomers. Binding of a signaling molecule to the receiving pocket allosterically regulates binding of the transcription factor to the target DNA sequence and thereby modulates mRNA production from the promoter of the operon [142].



**Figure 3.2 Description of the studied system.**

a) Structures illustrating the molecular interactions between the key residues in the *lac* repressor (yellow) and the key base pairs in the operator sequence (blue). The left panel shows a wild-type *E. coli* repressor-operator system, where the side chain of the key residues 17 and 18 from the repressor forms hydrogen bonds (dotted gray lines) with bases 4 and 5 of the symmetrical half operator. The left panel shows another repressor-operator pair. (b) Cartoon representation of the above three-dimensional structure with downstream reporter gene *lacZ* whose expression level is controlled by binding of the *lac* repressor (red) to the *lac* operator. The left panel represents the same but for another pair with mutations in the repressor and the operator (red crosses). (c) Representation of the data set. Genotype-phenotype map showing repression values as a function of residues 17 and 18 and all 16 operator variants based on the *in vivo* measurements. Low and high repression values are indicated by light and dark colors, respectively.

The *lac* system of *E. coli* is well-suited to start addressing the structure of adaptive landscapes for molecular interactions. Residues determining the specific binding between the *lac* repressor and its operator have been identified and circumscribed to only ten base pairs [84], reducing to a large extent the genotypic search space: essentially, two key residues, Tyr-17 and Gln-18, from the recognition helix of the *lac* repressor are responsible for specific recognition of key base pairs 4 and 5 (and symmetrically related base pairs) in the palindromic *lac* operator sequence [143] altogether reducing the determinant factors to six base pairs for the codons of residues 17 and 18 of the repressor, and four base pairs in the palindromic operator (Fig. 3.2a). We note that other residues (e.g., Ser-21, Arg-22 of the recognition helix

and base pair 6 of the operator) do have an effect on affinity, although less on specificity. Müller-Hill and co-workers have measured *in vivo* the repression values of repressor-operator pair variants obtained by extensive base pair substitutions at the aforementioned ten key positions [83]. Repression values for repressor-operator pair variants were determined as the ratio of repressed and unrepressed expression of a downstream  $\beta$ -galactosidase (*lacZ*) reporter gene, as measured via a standard Miller assay [144]. The measured data set covers 1286 out of a total of 6400 possible homodimeric repressor palindromic operator variants (two amino acids and two independent base pairs) [82,83,144]. From the measured data it has been observed that mutations in the key residues of the repressor (residues 17 and 18) contribute additively to the repression value, but the mutations in the key base pairs in the operator (base pairs 4 and 5) do not.

Based on this observed additivity between the repressor residues, the repression values for those mutants for which there were no measured data were determined by interpolation [82,83]. Additionally, to obtain the complete mapping between genotype and phenotype (i.e., repression Value), we have extended the data set to also describe non-palindromic operators, which constitute necessary intermediates for a step-by-step mutation process. To this aim, we have used the observation that each monomer of the dimeric repressor contributes additively to the binding energy with DNA [67,145]. Briefly, extrapolated repression values are calculated according to the general Eq. 2 of Ref. [67],  $F_{0102} = 1 + \sqrt{((F_{01} - 1) * (F_{02} - 1))}$ .

For a palindromic operator,  $F_{0102} = F_{01} = F_{02}$ , where  $F_{0(1/2)}$  is the product of two factors (one for each of the key residues) taken from Table 2 of Ref. [83].  $F_{01}$  and  $F_{02}$  may also be unequal, thus yielding the repression value for a non-palindromic operator. In total, our genotype data set consists of around  $10^6$  sequences (all combinations of ten independent base pairs) of repressor-operator variants, constituted of homodimeric repressors and palindromic or non-palindromic operators (Fig. 3. 2).

### 3.3 Algorithm

To find local repression optima in the genetic space of the repressor-operator pairs, the repression value of each point in the space is compared with the repression values of its nearest (single point mutation distant) neighbors. If at least one neighbor has a better repression value, then the point is not an optimum. If all the neighbors have lower repression values, then the point is an optimum. However, it might be that while none of the neighbors have a higher repression value, not all of them have lower repression values—that is, there might exist neutral neighbors. This does not

necessarily mean that both the assessed point and the neutral neighbors are not optima; on the contrary, they might all together constitute an optimum plateau.

Therefore, during the assessment process of each point of the genetic space, each time a neutral neighbor is found a recursive procedure is started to determine (i) the extent of the associated neutral region and (ii) to test each point of the neutral region for optimality before concluding about the optimality of the entire region—i.e., if one of the points of the neutral region has a neighbor not in the neutral region and with a higher repression value, then the region is not an optimum. At the end of the procedure, each point of the genotype map is defined either as “nonoptimum” or as “optimum-*j*,” where *j* is an integer number that differentiates each distinct and independent local optimum, and that is the same for all neutral points of an optimum plateau.

### 3.4 Results

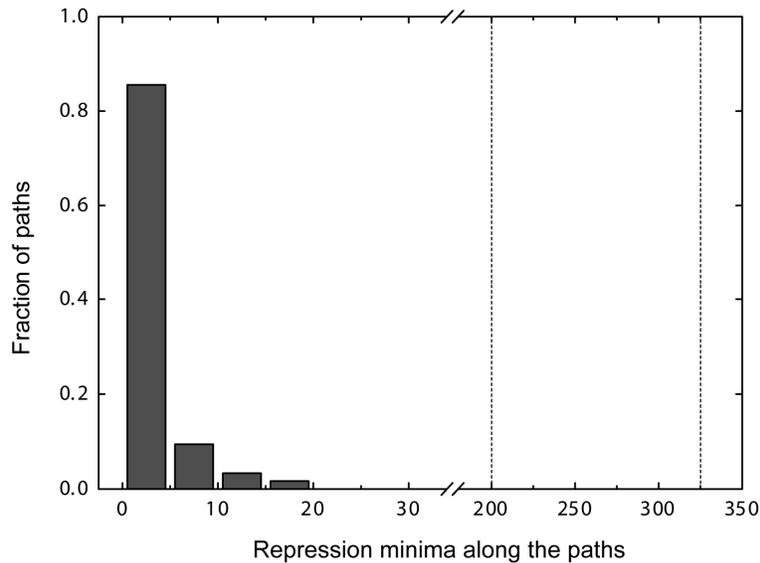
We have reconstructed the genotype-phenotype landscape detailing the repression values (defined as the ratio of unrepressed and repressed expression levels of the downstream *lacZ* gene) for variants of the *lac* repressor-operator system, and analyzed the ruggedness of the landscape. The genotype space contains about  $10^6$  variants, covering all possible combinations of mutations in the repressor and the operator, at the positions known to determine their binding specificity (i.e., base pairs 4, 5, and symmetrically positioned base pairs 4' and 5' in the operator, as well as base pairs coding for residues 17 and 18 in the recognition helix of the *lac* repressor [84] (Fig. 3.2)). A particular operator-repressor variant of the explored genotype space is represented by the sequence at the four key positions in the operator (respectively base pairs 5, 4, 4', and 5' see Fig. 3.2), followed by symbols of the amino acids present respectively at residues 17 and 18 of the LacI protein. Thereby, the wild-type genotype would for instance be designated tgcaYQ. The algorithm described above was used to search for peaks in repression values throughout the entire delineated genotype space. This analysis revealed 19 distinct peaks, i.e., areas of high repression values within the genotype space, isolated from each other by genotypes of strictly lower repression levels. Table 3.1 lists the genotypes and associated repression values of the 19 peaks of the landscape.

To quantify the distinctness of the peaks, we analyzed their relative distance and the decreases in repression values along the paths between them. On average two peaks are separated by a Hamming distance of six mutations, with Hamming distances ranging between two and nine mutations.

Note that a specific situation occurs in the case of serine which is encoded by two independent groups of codons separated by two mutations. This results in the

existence of distinct peaks when this amino acid is present in the repressor. For simplicity, we have not distinguished these peaks in Table 3.1. Next, we looked more closely at the paths between two peaks separated by the average Hamming distance (six mutations). In particular, we considered the peaks atgcPK and tgcaSQ (respectively, peaks 9 and 3 in Table 3.1). These two peaks have repression values of 200 and 325, respectively. The peak tgcaSQ is the optimum that is closest to the wildtype sequence tgcaYQ. Note that for simplicity we have excluded the cases of reverse mutations and restricted our analysis to direct paths between the peaks.

For a Hamming distance of  $h$  between two peaks, one can follow  $h!$  different direct paths. Figure 3.3 presents the histogram of the smallest repression values encountered along each of the  $6! = 720$  paths going from peaks acgtPK to tgcaSQ. The vast majority of paths ( $>600$ ) decreases down to repression values of 2 or less, which represents more than a 100-fold reduction in repression. The 12 most optimal paths, i.e., the paths that involve the least drastic dips, still decrease down to repression values of around 20. Thus, to evolve from one peak to the other, the system has to overcome a loss of at least tenfold in repression. A number of typical paths are illustrated in Fig. 3.4, where the respective mutations and repression values at each step are indicated. In this graph, path 1 is one of the least likely paths, exhibiting a 200-fold drop in repression value at step 3 where the repression decreases to a value of 1. In this path, the operator is mutated first, resulting in disruption of its palindromic symmetry, and decreasing the repression value to about 80. At the second and third steps, the operator experiences additional mutations that bring it closer to the final sequence, although still maintaining the sequence asymmetry initially introduced. Ultimately, the repression shrinks to 1 at the third step. Subsequently, in steps 4 and 5 two mutations occur in the repressor. The first of these mutations, lysine (L) to glutamine (K), compensates for the mutations in the operator and restores the repression level to about 100, while the second mutation in the repressor, a proline (P) to serine (S) transition, is almost neutral. Finally, the last  $a$  to  $t$  mutation from step 6 restores the symmetry of the operator, bringing the repression value to 350 at the tgcaSQ peak. A close alternative to path 1 would be path 2, where all mutation steps are the same as in path 1 except for a permutation of the mutations occurring at steps 3 and 4, affecting respectively the operator and the repressor (see Fig. 3.4b, paths 1 and 2, outlined steps). With this new mutation order, instead of a decrease at step 3 followed by a restoration of the repression level at step 4, now both mutations (K to Q at steps 3 and  $g$  to  $c$  at step 4) increase the repression level, thus making this path more favorable.



**Figure 3.3 Histogram of the minimal repression values along direct paths between peaks *acgtPK* and *tgcaSQ*.** Dotted lines indicate the repression levels at the beginning and at the end of the mutational path (repression values of 200 and 350 for genotypes *acgtPK* and *tgcaSQ*, respectively).

These two alternative paths show that the *g* to *c* mutation in the second half of the operator with the K or Q amino acid in the second key residue of the repressor exhibit a sign epistatic interaction. The repression level then stays almost constant during the next two mutation steps that occur in the operator.

The most likely path between optima *acgtPK* and *tgcaSQ* is path 3 depicted in Fig. 3.4, which exhibits the smallest dip among all possible paths. Here, the first mutation occurs in the repressor with the transition from P to S, which brings the repression level to about 100. The repression level then stays almost constant during the next two mutation steps that occur in the operator. Interestingly, in this pathway the palindromic symmetry, initially broken by the *a* to *t* mutation in the operator sequence at step 2, is immediately restored at step 3 with a *t* to *a* mutation in the second half of the operator. The following mutation is the K to Q transition in the repressor at step 4, which reduces the repression level to 20. This is the lowest repression level along this path, constituting a tenfold drop relative to the initial repression value at the peak. The repression level is then progressively restored as

Peak rank	Genotype		Repression level
	Operator	Repressor	
1	tg...ca	SM	520
2	tt...aa	HM	500
3	tg...ca	SQ	325
4	gt...ac	KS; KT; K M <sup>a</sup>	300
5	aa...tt	IS	300
6	aa...tt	SS	225
7	ta...ta	SG; IG; PG <sup>b</sup>	220
8	tg...ac/gt...ca	KM	219
9	ac...gt	PK	200
10	ac...gt	MK	200
11	aa...ac/gt...tt	SS	184
12	at...at	VM	160
13	ag...ac/gt...ct	PS	150
14	ag...ct	PS	150
15	tg...ac/gt...ca	KQ	100
16	gt...ac	KQ	100
17	tg...ca	SN	91
18	ga...tc	QT	90
19	gt...ac	VQ	50

<sup>a</sup>S, T, and M are connected,so 1 optimum.

<sup>b</sup>S, I, and P are connected,so 1 optimum.

**Table 3. 1 Genotypes and repression values for the nineteen independent peaks in the phenotype landscape of lac operator-repressor pair variants.** Non-palindromic operators are indicated by their two equivalent reverse-complement sequences.

the two remaining symmetric *g* and *c* key bases of the operator mutate, respectively, to *c* and *g* to give the final palindromic operator.

In fact, path 3 belongs to a group of 12 best paths that are essentially equivalent. Indeed, due to the sequence symmetry of the operator, mutations of base pairs 5 and 5' at steps 2 and 3 can occur indistinctively in the reverse order, as well as mutations at base pairs 4' and 4 at step 5. Additionally, the P to S transition at step 1 can occur at step 2 or 3 with only a negligible decrease in repression values at some steps along the respective paths. Combining all these possible permutations produces a group of 12 paths, all having the same shape as path 3 with their minima in repression level at step 4.

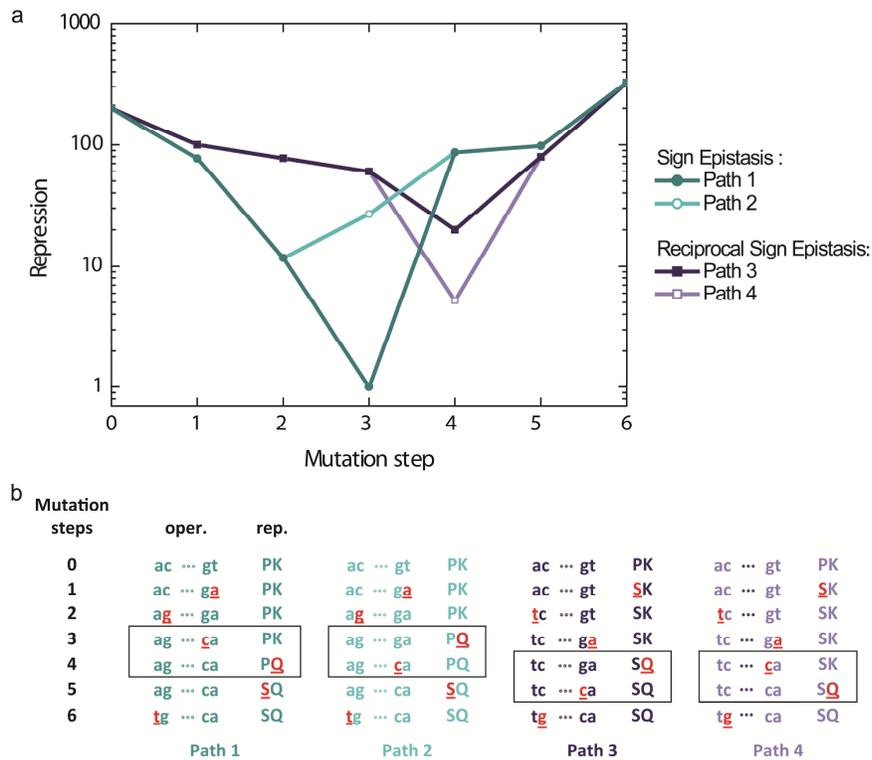
The best alternative path to path 3, apart from the 12 aforementioned paths, is path 4, which differs only in the order of the mutations leading to, and following, the deepest drop in repression value at step 4. Permuting the order of these mutations—

that is, the K to Q transition in the repressor protein and the first *c* to *g* mutation in the operator (Fig. 3.4b, outlined steps) -results only in a deeper global dip in repression value at step 4 compared to path 3. Similar to paths 1 and 2, paths 3 and 4 also show that the effect of the *g* to *c* mutation in position 4' can change sign, depending on residue 18 of the operator (K or Q). Additionally, however, the effect of this K to Q mutation now also changes sign depending on the position 4' of the operator (*g* or *c*). Thus, the two mutations exhibit a reciprocal sign epistatic interaction (Fig. 3.1a).

Reciprocal sign epistasis occurs when two mutations are individually deleterious but jointly result in a positive effect (Fig. 3.1a). Such a situation captures, at the level of individual mutation steps, the constraints created by a multipeaked landscape. Our analysis of paths 3 and 4 shows that the choice between alternative best paths between two peaks reduces to a choice between two routes in a reciprocal sign epistasis pattern that is located where these paths encounter their deepest drop in repression values. In other words, the lowest point in the optimal path between two peaks results from a reciprocal sign epistasis interaction. This observation illustrates that reciprocal sign epistatic interactions stand at key locations of a multipeaked landscape, and is in line with a theoretical investigation of ours, which indicates that reciprocal sign epistasis is an essential ingredient for the existence of multiple peaks [146].

From Table 3.1, we can also identify several peaks that are in close proximity to each other, being separated by a Hamming distance of only two. Three different situations can be discerned among the 13 cases. First, two different peaks can have the same repressor, while their operators differ by two mutations. This is for instance the case for peaks 6 and 11 or 13 and 14. The opposite situation also exists, where several peaks share the same operator sequence but the associated repressors differ by two mutations. This holds, for instance, for peaks 1, 3, and 17. The intermediate situation, where each of the operator and repressor variants differs by only one mutation between two peaks, also exists. This special case is encountered between peaks 11 and 13, both carrying a non-palindromic operator.

Examination of the direct paths between these proximal peaks reveals an interesting pattern. When two peaks differ only by their operators or by one mutation in the repressor and one in the operator (which is the case only between peaks 11 and 13), they are only weakly separated, with the minimal dip among the different paths being less than a factor of 2. Notably, when two proximal peaks differ by their operators, at least one of them is non-palindromic. Thus, we do not observe two proximal peaks that differ only by their operators with both of them being palindromic. This observation might explain why those peaks are only weakly separated.



**Figure 3.4** Examples of direct evolutionary paths between peaks acgtPK and tgcaSQ.

a) Repression values of the intermediate mutants along each path. b) Intermediate sequences along the mutational paths. Mutations are underlined. Steps exhibiting epistatic interactions are outlined.

In contrast, when two proximal peaks differ only by their repressor sequence, which occurs in half of the cases of proximal peaks, the minimum drop in repression is larger than a factor of 5, or even a factor of 10 or 100 in three of the cases (between peaks 1 and 3, 3 and 17, and 8 and 15). The sole exception concerns the paths between peaks 16 and 19, for which the minimal drop in repression is small (i.e., less than a factor of 2). Therefore, while some peaks are close to each other in sequence space, they can still be separated by genotypes having substantial reduced repression values.

Note that due to the degeneracy of the genetic code, there exist silent mutations for most of the codons of the amino acids. Therefore, all peaks in the landscape form in fact a small plateau of neutral variants (see, however, Kimchi-

Sarfaty *et al.* [147] for an example of phenotypic effect due to “silent” mutations). The presence of neutral variants at the peaks results in the existence of parallel identical groups of paths between peaks. For instance, for the acgtPK to tgcaSQ transition depicted in Fig. 3.4, because P and S can be encoded, respectively, by the triplet *ccn* and *ucn* (where *n* can be any base), and K and Q, respectively, by the triplets *aar* and *car* (where *r* can be either *a* or *g*), there are in fact six identical “channels” of direct paths to go from one peak to the other (one channel for each combination of sequences at *n* and *r*). Each of these channels is therefore constituted of the same group of 720 paths described previously, differing only by the base sequence at positions *n* and *r* in the codons. For example, in the paths of Fig. 3.4, *n* and *r* have been arbitrarily chosen to be *c* and *a*—although this choice is not apparent and could have been different without altering the result. Since these channels are independent from each other, the validity of our previous discussion on epistasis and constraints is unaltered.

Interestingly, several of the peaks in the genotype-phenotype map of the *lac* repressor-operator occur for non-palindromic operator sequences (see Table 3.1). Considering the symmetry constraint imposed by the homodimeric repressor protein, this finding might seem counterintuitive as for the given repressor sequence, one would expect to find a better optimum with a symmetrized palindromic operator. Closer inspection reveals that all the non-palindromic optima we have found need more than just one mutation for their operator sequence to become palindromic, making them just sufficiently isolated to constitute local optima (see Table 3.1). Furthermore, these non-palindromic optima are not global optima and their symmetrized versions always have a higher or at least equal repression value, thereby forming higher or equivalent peaks unless they are themselves buried into another higher peak.

### 3.5 Discussion

In recent years a number of adaptive landscapes have been determined empirically through the genetic reconstruction of neighboring genotypes. These efforts have identified sign epistatic interactions, either at the genotype-phenotype level [61] or at the genotype-fitness level [71], thereby showing that paths can be selectively inaccessible (Fig. 3.1b). Nevertheless, some paths to the global optimum remained selectively accessible, indicating that the landscapes were single peaked. [61,71] Here we report the presence of multiple peaks in the landscape detailing the repression value of the *lac* regulation system as a function of key operator base pairs and repressor residues. The peaks are distinct: they consist of repressor-operator pairs capable of high repression values, which are surrounded by genotypic variants

of lower repression levels. Our assumption of complete additivity between mutations in the key residues of the repressor might lead to an underestimate of the ruggedness of the landscape. Relaxing this assumption would only lead to a more rugged landscape. However, despite this assumption, distinct peaks are identified in the genotype-phenotype space.

A rationale for the existence of multiple peaks in the case of the *lac* regulatory system can be found by considering the analogy between the operator-repressor interaction and a key fitting a lock. Forming a new lock and matching key by stepwise mutations presents a dilemma: mutating the key first is not viable because it does not fit the old lock, and vice versa. This dilemma can arise for a recognition function between two components that can change both, in contrast, for instance, with an enzymatic reaction, where only one component changes by evolution. However, it is not a necessary consequence. The dilemma can in principle be resolved by the molecular equivalent of a master key: an intermediate transcription factor that is able to bind intermediate operator sequences, thus bridging two peaks [96]. Our study shows that such a master key does not exist for the *lac* repressor-operator system.

A multipeaked landscape reflects the widespread presence of epistatic interactions across the genotypic space. Indeed, among the mutations that bring the system to an optimum, there must necessarily be some that have a decreasing effect if introduced from another optimum. Otherwise the system would be single peaked. In other words, some of the mutations in one binding partner will only be beneficial when the other partner has already been modified, and vice versa. The requirement of such a reciprocal sign-epistatic interaction for multipeaked landscapes, which can also be theoretically addressed in a more rigorous manner [146], is supported by our analysis: as predicted, such interactions appeared present along paths exhibiting the highest minimum.

It has frequently been recognized that a multipeaked landscape can constrain a stepwise Darwinian evolution process by trapping the evolving population in local suboptima—i.e., peaks lower than the global optima. Given the existing diversity of recognition within the GalR-LacI family of transcription factors [148], the results suggest that evolution has been able to overcome entrapment on suboptimal peaks. Different scenarios may be considered for escaping suboptima. First, certain environments may free the system from a selective pressure temporarily, allowing new recognitions to be achieved through neutral drift. Alternatively, the participation of the system within a larger network of interacting components may alleviate the constraints. For instance, a duplication event may allow one of the duplicate repressors to compensate repression-decreasing mutations in the diverging copy [67]. One might also hypothesize the existence of hidden paths, involving substitutions

beyond the key residues. However, this implies longer paths in an expanded genotypic space, which also occurs at the expense of reduced probability [149].

Finally, we also would like to discuss the limitations of our approach. First, our analysis is based on phenotypic rather than fitness data. In order to address the evolutionary dynamics in a quantitative manner, the relation between repression characteristics and fitness should be determined, which also involves the nature of the environmental changes. Second, not all evolutionary intermediates have been directly characterized, but rather have been interpolated using the assumption that the two residues contribute additively to the repression value. While this assumption does not change our main conclusion that the *lac* repressor-operator system exhibits a multi-peaked landscape, it will be of interest to reconstruct all intermediates between two peaks.

Specific molecular interactions are ubiquitous in biological systems and essential to their complexity and their ability to survive. One may therefore expect that multiple peaks in phenotype and fitness, as well as the underlying reciprocal sign epistatic interactions, be equally pervasive. It will be intriguing to explore how these elementary interactions shape the course of evolution of more elaborate biological functions.



# Chapter 4

## Tradeoffs and optimality in the evolution of gene regulation

*Cellular regulation is believed to evolve in response to environmental variability. However, this has been difficult to test directly. Here, we show that a gene regulation system evolves to the optimal regulatory response when challenged with variable environments. We engineered a genetic module subject to regulation by the lac repressor (LacI) in Escherichia coli, whose expression is beneficial in one environmental condition and detrimental in another. Measured tradeoffs in fitness between environments predict the competition between regulatory phenotypes. We show that regulatory evolution in adverse environments is delayed at specific boundaries in the phenotype space of the regulatory LacI protein. Once this constraint is relieved by mutation, adaptation proceeds toward the optimum, yielding LacI with an altered allosteric mechanism that enables an opposite response to its regulatory ligand IPTG. Our results indicate that regulatory evolution can be understood in terms of tradeoff optimization theory.*

## 4.1 Introduction

The capability to regulate behavior and physiological state in response to the environment is a fundamental property of all living systems. How novel regulatory phenotypes emerge and adapt in populations challenged by the conflicting demands of variable environments has long fascinated biologists [150,151,152]. The general prediction from theory is that regulatory responses are favored to evolve when the selective environment fluctuates more slowly than the typical timescale of responses [151,153]. However, the outcome of evolution in variable environments may depend on various unknown factors, such as constraints of physical, biochemical, and genetic origin [154]; the competition between different regulatory phenotypes [155]; and the precise strength and direction of selection on regulation [151,153,156]. Therefore, even though experiments have shown that regulation can be beneficial [151] and can be altered by phenotypic screening [153], most experimental work on the evolution of regulation by mutation and natural selection has remained indecisive and difficult to explain in terms of the causal selective forces and constraint [157].

Conceptually, addressing this issue is straightforward. Selection at the phenotypic level can be revealed by quantifying the dependence of fitness on the relevant phenotypic parameters [61,71,96]. Constraints may be identified by experimental evolution: an evolutionary response according to selection indicates adaptation, whereas conversely, a lack of such a response points to a constraint [92]. In practice, however, it is nontrivial to determine the relation between regulation and fitness. The overall growth rates of two known *Escherichia coli* regulatory mutants have been measured in a variable environment [158], but our limited insight into the genetic basis of regulatory changes hampers extending this approach to systematically assay the full range of possible regulatory responses.

To overcome these obstacles, we have developed a synthetic approach. Synthetic systems [159] can be engineered to contain the two core elements of regulatory evolution: a cellular phenotype that confers a benefit in one environment and a burden in another and a regulation system that senses the environment and modulates the phenotype. By designing a sensed environmental cue that can be varied separately from the environmental factor that confers the burden or benefit, one can quantify the relation between expression regulation and fitness, prior to adaptation and without the need for a comprehensive library of regulatory mutants. Using this approach, we present a case study of optimality in evolution in variable environments throughout the various levels of biological organization, from the environment down to molecular mechanisms and genotype.

## 4.2 Experimental system and fitness in constant environments

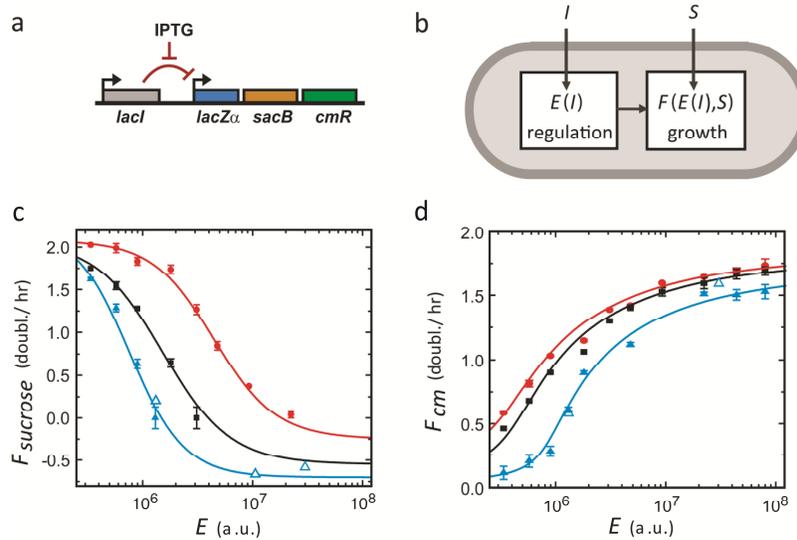
To quantify the selective forces and evolutionary change of regulation in variable environments, we constructed an experimental system in *E. coli* consisting of two genetic modules (Fig. 4.1a and 4.1b). The first module comprises an operon that harbors the *sacB* and *cmR* genes, which affect the growth rate, as well as the *lacZ $\alpha$*  gene, which is used to quantify the operon expression level  $E$ .  $E$  is controlled by the *lac* repressor, LacI, which constitutes the regulatory module. The system as constructed (termed WT hereafter) responds to increasing concentrations of the environmental cue isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG) by increasing  $E$ .

In media containing sucrose, expression of the operon by induction with IPTG leads to a reduced growth rate (Fig. 4.1c and Fig. 4.6-materials and methods), as determined by monitoring the optical density of growing populations. Before inoculation in the sucrose media, the cells were grown without sucrose but with the IPTG concentration of interest in order to achieve the steady-state expression levels. The observed negative effect on growth rate is due to the sucrose-polymerizing activity of levansucrase encoded by *sacB*, which leads to the accumulation of large polysaccharide chains that interfere with cell wall formation [160]. At the highest levels of operon expression, we observed negative growth rates. The corresponding decrease in the number of viable cells was quantified using competition and plating assays (materials and methods). We note that the cells lack the genes to metabolize sucrose and use it as a carbon and energy source. Thus, sucrose-containing media confer a selective pressure to decrease operon expression.

For media that contain the antibiotic chloramphenicol (Cm), the growth rate is suppressed at low operon expression levels (Fig. 4.1d and Fig. 4.6a-materials and methods). When operon expression is increased by induction, the growth rate is progressively restored. This beneficial effect of operon expression is due to the inactivation of the antibiotic by the CmR gene product. Cm media thus confer a selective pressure to increase expression. Increasing the Cm or sucrose concentrations predominantly shifts the point of half-maximum growth along the expression axis and does not significantly affect the maximum growth rate. We find that the growth expression data in sucrose and in Cm media are well-described by the functions  $F_{sucrose}(E)$  and  $F_{Cm}(E)$ , which are based on a simple model of the underlying reaction kinetics (materials and methods).

## 4.3 Tradeoffs for fixed expression phenotypes in variable environments

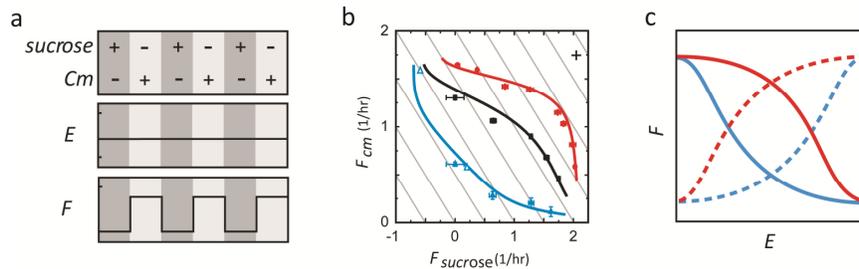
Environments that vary in time between sucrose and Cm result in fluctuating demands on operon expression. The simplest phenotype is one that exhibits a single



**Figure 4.1 Phenotype and fitness characterization**

a) Schematic of the synthetic operon and regulation system. b) Functional system representation.  $F(E,S)$  describes the dependence of fitness (growth rate) on operon expression and concentration of selective agent (sucrose or Cm).  $E(I)$  is the dependence of expression on the concentration of the environmental cue IPTG. c) Measured  $F(E,S)$  relations for sucrose media. Sucrose concentrations (w/v): 0.15% (red), 0.25% (black), and 0.40% (blue). Error bars are standard errors ( $n = 3$ ). (Closed symbols) Expression and fitness values as achieved by IPTG induction. (Open symbols) Data obtained by competition assays (materials and methods). Expression standard error is smaller than the symbol size. d) Measured  $F(E,S)$  relations for Cm media. Cm concentrations: 25  $\mu\text{g/ml}$  (red), 40  $\mu\text{g/ml}$  (black), and 80  $\mu\text{g/ml}$  (blue). Curves in c) and d) are fits to a growth model (materials and methods). See also Figure 4.6 (materials and methods).

fixed expression level in both the sucrose and the Cm environment (Fig. 4.2a). Variable environments confront such unregulated phenotypes with a tradeoff: high fitness in sucrose media—resulting from low operon expression levels—will be at the expense of low fitness in Cm media. Conversely, high fitness in Cm media—due to high expression levels—will entail low fitness in sucrose media. To gain insight into the optimization of the total fitness of fixed expression phenotypes in sucrose-Cm variable environments, we plotted  $F_{\text{sucrose}}(E)$  versus  $F_{\text{Cm}}(E)$  for a range of  $E$  (Fig. 4.2b). We find that, for the lower Cm and sucrose concentrations, this so-called tradeoff curve (Fig. 4.2b, red line) is concave and bulges outward toward optimal growth in both environments (Fig. 4.2b, cross), whereas for the high concentrations, it is convex and bulges inward away from optimal growth (Fig. 4.2b, blue line).



**Figure 4.2 Tradeoffs for fixed expression phenotypes in variable environments.**

a) Variation of environmental and system parameters. Here, sucrose alternates with Cm, and the operon expression level  $E$  is constant across both environments, whereas the growth rate  $F$  may vary.

b) Tradeoff curves in sucrose-Cm variable environments for different sucrose and Cm concentrations. For any  $E$ , the resulting growth rate in Cm  $F_{Cm}(E)$  (Figure 4.1d) is plotted against the growth rate in sucrose  $F_{sucrose}(E)$  (Figure 4.1c). (Red line) 0.15% sucrose, 25  $\mu\text{g/ml}$  Cm. (Black line) 0.25% sucrose and 40  $\mu\text{g/ml}$  Cm. (Blue line) 0.40% sucrose and 80  $\mu\text{g/ml}$  Cm. The cross indicates the optimal growth in both environments at low sucrose and Cm concentrations (red line). The diagonal isoclines indicate equal overall growth rate  $F_{tot} = (F_{sucrose}(E) + F_{Cm}(E))/2$  when alternating between sucrose and Cm for equal periods of time. Error bars are standard errors ( $n=3$ ).

c) Schematic diagram illustrating the origin of the switch from concave to convex tradeoff curves, as observed in b). The lines represent the expression growth relations  $F(E)$  for sucrose (solid lines) and for Cm (dashed lines). Red indicates lower concentrations of sucrose and Cm, and blue indicates higher concentrations. At fixed medium expression  $E$ , the red curves show a near optimal growth in both environments resulting in a tradeoff curve that bulges outward. For the blue growth conditions, such a condition does not exist. Here, the highest fitness is attained by phenotypes with a fixed high or low expression.

Shapes of tradeoff curves and their consequences for selection have been extensively analyzed on a theoretical level. As pointed out by Levins, who originally introduced the idea in the 1960s (Levins, 1968), concave tradeoff curves favor intermediate expression phenotypes that do averagely well in both environments, whereas convex tradeoff curves favor extreme expression phenotypes that perform well just in one environment (Fig. 4.2c). In our study, these contrasting cases depend on the Cm and sucrose concentrations that may be viewed as the amplitude of the environmental variations. The observed change in convexity is a direct consequence of the nonlinear biochemical processes that underlie growth and may well be a general feature of fitness tradeoffs in biological systems.

#### 4.4 Fitness of regulated phenotypes in variable environments

Regulated phenotypes, which can sense changes in the environment and respond by adjusting their expression, can escape from tradeoff curves for unregulated phenotypes (Fig. 4.2c) and acquire higher fitness. When Cm is supplemented with the cue IPTG (1 mM; see Fig. 4.3a), the WT phenotype in fact exhibits a nearly optimal response: with IPTG, the high expression level  $E_I$  is favorable in Cm, whereas without IPTG, the resulting low expression level  $E_\theta$  is favorable in sucrose. As observed in the previous section, however, the best unregulated phenotypes perform almost as well for the lower Cm and sucrose concentrations due to the concave tradeoff curve (Fig. 4.2b, red line). The resulting limited advantage of regulation is not caused by weak selection in each medium separately, as evidenced by the large growth rate differences (Fig. 4.1c and 4.1d, red curves) but, rather, by the precise shape of the growth expression curves  $F_{sucrose}(E)$  and  $F_{Cm}(E)$ . For higher sucrose and Cm concentrations though, the tradeoff curve becomes convex (Fig. 4.2b, blue curve), thus lowering the maximally attainable fitness for unregulated phenotypes and consequently increasing the selective advantage of regulated phenotypes.

To compare the performance of different phenotypes systematically, we consider their overall growth rate in variable environments as a function of  $E_\theta$ , the expression level in the environment without IPTG (and with sucrose), and  $P = E_I/E_\theta$ , the fold change in expression when switching to the environment with 1 mM IPTG (and with Cm). The separate growth rates in the sucrose and Cm media are then found by substituting the value for  $E_\theta$  in the function  $F_{sucrose}(E)$  and  $E_I$  in  $F_{Cm}(E)$  (Fig. 4.1c and 1d).

We determine the overall growth rate,  $F_{tot}$ , as the arithmetic mean of the separate growth rates weighted by the fraction of time spent in each environment. We use the arithmetic mean (as opposed to the geometric mean) of the local growth rates because growth rates—unlike the number of offspring—are Malthusian fitness parameters [161]. In Fig. 4.3b, we plot  $F_{tot}$  as a function of  $E_\theta$  and  $P$  for the high sucrose and Cm concentrations. This fitness function contains a single optimum near WT (low  $E_\theta$  and high  $P$ ). The peak is well separated from  $P = 1$ , consistent with the strong selection for regulated over non-regulated phenotypes.

#### 4.5 Competition in variable environments

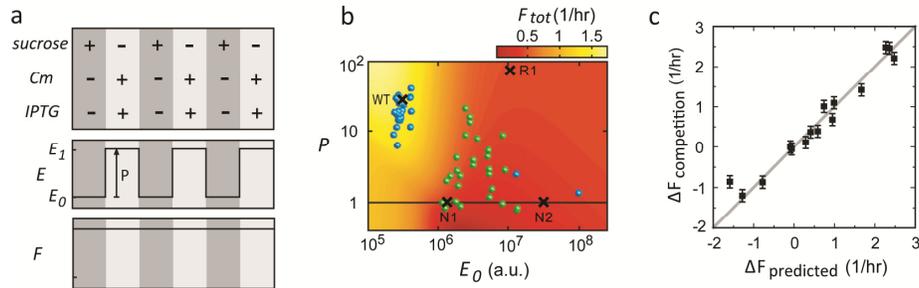
Having determined the relation between expression phenotype and fitness (Fig. 4.3b), the question is whether it accurately captures the competition between different regulatory mutants in variable environments. We tested this in two ways.

First, we competed *lacI* mutants one to one. We employed *lacI* mutants that were generated using error-prone polymerase chain reaction (PCR) and displayed contrasting expression phenotypes (Fig. 4.3b, crosses). The different alleles were cloned into our plasmid, and pairs of them were grown in sucrose and Cm environments, either with or without IPTG, whereas their relative abundance was followed by plating. Again, cells were grown non-selectively before and after selection, and care was taken that growth was not limited by other growth factors such as carbon or oxygen. These competition results were compared to fitness differences predicted by the fitness function  $F_{tot}(E_0, P)$ , based on the uninduced ( $E_0$ ) and induced ( $E_i$ ) expression levels of each mutant, which showed a good agreement (Fig. 4.3c).

Second, we investigated the competition within a large population of cells with randomly mutated regulatory systems. For this purpose, we again mutated the *lacI* coding sequence with error-prone PCR. Sequencing indicated an average mutation rate of  $\sim 3 \times 10^3$  per basepair. The mutated *lacI* coding regions were placed back into the original plasmid, resulting in a diverse population of  $\sim 10^6$  variants. Note that only the regulatory system was genetically diversified. The rest of the plasmid, including the selection genes, as well as the chromosomal background, remained identical. To characterize the phenotypic diversity prior to selection, 35 clones were randomly picked from the population and mapped in phenotype space based on their measured  $E_0$  and  $P$  values. The phenotypes did not appear distributed equally throughout the  $E_0$ - $P$  plane but were, rather, located toward higher  $E_0$  and lower  $P$  down to  $P = 1$  (Fig. 4.3b), consistent with the expectation that most random mutations in the repressor will deteriorate the ability to repress. The original WT phenotype was not present in this sample of 35 variants.

We then exposed the population of  $\sim 10^6$  random variants to a variable-purifying selection. In short, the cells were first grown in the sucrose medium (0.4% w/v) for 6 hr. From the end population,  $1/500^{\text{th}}$  was taken and then grown non-selectively at 1 mM IPTG for 3 hr to induce the cells. Next, Cm was added (at 80  $\mu\text{g}/\text{ml}$ ), and the population continued to grow for another 6 hr. Experiments indicated that the 3 hr of nonselective growth was sufficient to reach steady-state expression and did not significantly influence the overall fitness (Fig. 4.6-materials and methods). From the final population, we again took isolates randomly, assayed their  $E_0$  and  $E_i$  values, and mapped them onto the  $E_0$ - $P$  plane. The group of isolates was clustered at the WT phenotype (Fig. 4.3b).

These results indicate that the simple fitness function  $F_{tot}(E_0, P)$  captures the competition between diverse regulatory variants, which involves the integration of the different fitness gains and losses experienced by the competing variants as they experience the environmental changes, into an overall total fitness.



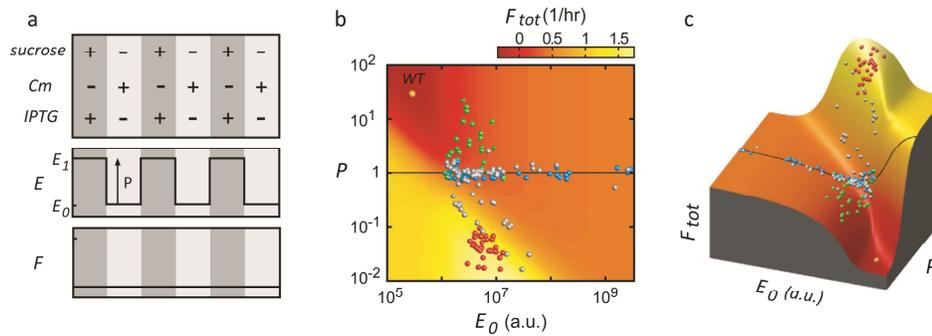
**Figure 4.3 Fitness and evolution in a variable environment with stabilizing selection.**

a) Variation of environmental and system parameters. Here, sucrose alternates with Cm plus IPTG. In this environment, the operon expression level of the system as constructed (WT) will vary between a low  $E_0$  in sucrose and a high  $E_1$  in Cm plus IPTG, resulting in high growth rates in both environments.  $P$  is the fold change in operon expression ( $P = E_1 / E_0$ ). Note that the selection experiments (dots, b) involve one sucrose/Cm cycle. b) Selective landscape and mapped regulatory variants. Color indicates the overall growth rate in the variable environment defined in a), determined with the data from Fig. 4.1c and 1d ( $F_{tot} = (F_{sucrose}(E_0) + F_{Cm}(E_1))/2$ ), as a function of system parameters  $E_0$  and  $P$ . Sucrose, Cm, and IPTG concentrations are 0.40%, 80  $\mu\text{g/ml}$  Cm, and 1 mM, respectively. Different *lacI* regulatory variants are mapped on the landscape based on their measured  $E_0$  and  $P$  values. (Crosses) WT and three previously isolated *lacI* mutants. (Green spheres) 35 randomly chosen isolates from a diverse population obtained by *lacI* mutagenesis. (Blue spheres) After selection in the corresponding variable environment. Growth time is 6 hr in each medium. Prior to inoculation in new medium, the cells are grown non-selectively (neither sucrose nor Cm) to adjust to the new IPTG level. c) Fitness differences from competition experiments between genetic *lacI* variants against fitness differences predicted by the fitness landscape (materials and methods).

They also confirm that the expression  $E$  is the key phenotypic parameter in determining fitness in our system.

#### 4.6 Evolution under directional selection in variable environments

In order to study adaptation to a novel regulatory function, we defined a variable environment in which the WT phenotype is maladapted and can improve by changing the regulatory response. We exploit the decoupling in the system, which allows one to impose a controlled mismatch between cue and selective agent. Here, a medium with sucrose plus IPTG (1 mM) alternates with a medium with Cm (Fig. 4.4a). In the presence of sucrose and IPTG, the high induced operon expression is burdensome, whereas in Cm media without IPTG, a high operon expression would be beneficial but is repressed.



**Figure 4.4 Fitness and evolution in a variable environment with directional selection.**

a) Variation of environmental and system parameters. Here, sucrose plus IPTG alternates with Cm. In this environment, the operon expression level of the system as constructed (WT) will vary between a high  $E_1$  in sucrose plus IPTG and a low  $E_0$  in Cm, resulting in low growth rates in both environments.  $P$  is the fold change in operon expression ( $P = E_1 / E_0$ ). Note that the selection experiments (dots, b and c) involve one sucrose/Cm cycle. b) Selective landscape and mapped regulatory variants. Color indicates the overall growth rate in the variable environment, determined with the data from Figures 4.1c and 4.1d ( $F_{tot} = (F_{sucrose}(E_0) + F_{Cm}(E_1))/2$ ), as a function of system parameters  $E_0$  and  $P$ . Sucrose, Cm, and IPTG concentrations are 0.40%, 80  $\mu\text{g/ml}$  Cm, and 1 mM, respectively. Different *lacI* regulatory variants are mapped on the landscape based on their measured  $E_0$  and  $P$  values. (Green spheres) 35 randomly chosen isolates from a diverse population obtained by *lacI* mutagenesis. (Blue spheres) after selection in the corresponding variable environment. (Gray spheres) after a second cycle of *lacI* mutagenesis and selection in both media. (Red spheres) after a third cycle. Growth time is 6 hr in each medium. Prior to inoculation in new medium, the cells are grown nonselectively (neither sucrose nor Cm) to adjust to the new IPTG level. c) Idem, in three-dimensional representation, rotated for visibility.

Mirroring the case of the unregulated phenotypes, changing expression overall (in both  $E_0$  and  $E_1$ ) gives rise to a tradeoff: an increase in expression leads to gains in Cm but losses in sucrose, whereas conversely, a decrease in expression yields gains in sucrose but losses in Cm. Improvements in both environments may be realized by changes in the regulatory response, namely when IPTG would lead to repression, and an absence of IPTG to expression.

Such an inverse regulatory response is, in fact, observed for the LacI homolog PurR, which represses in the presence of the co-repressor guanine [162], suggesting that such inversions have occurred in evolutionary history.

The fitness function corresponding to this variable environment is displayed in Fig. 4.4b and 4.4c.  $E_1$  and  $E_0$  here, respectively, denote the expression level in the medium with sucrose plus IPTG and the medium with Cm. The WT phenotype here

maps onto a fitness valley at  $P > 1$ , indicating its poor performance in both environments. The peak is located at  $P < 1$ , which reflects the inverse nature of the corresponding phenotype ( $E_I < E_\theta$ ).

Next, we monitored evolution of the WT system using the same directed evolution approach. In brief, we prepared randomly mutated *lacI* variants and then grew a starting population of  $\sim 10^6$  in a sucrose medium (0.4% w/v) with IPTG (1 mM) for 6 hr. From the end population, 1/500<sup>th</sup> was taken and grown non-selectively without IPTG for 3 hr. Then Cm was added (to reach 80  $\mu\text{g/ml}$ ), and the population was grown for an additional 6 hr.

After selection, a random sample of 35 isolates was mapped within the  $E_\theta$ - $P$  plane, which showed that the population mean had shifted toward higher fitness as expected (Fig. 4.4b and 4.4c, blue spheres). However, no isolates were found below  $P = 1$ , suggesting that there exists an adaptive constraint that renders crossing of the  $P = 1$  boundary a rare or impossible event. Note that  $P < 1$  phenotypes might be present in the population at low numbers and absent in the isolates due to insufficient sampling. However, the fitness function does indicate strong selection for  $P < 1$  phenotypes: the growth rate of the optimal phenotypes is more than 1.4 dbl/hr higher than the best isolates at  $P = 1$ , making their initial presence very unlikely ( $p < 0.02$ , given the relative enrichment factor of  $2^{(1.4 \times 12)} \sim 1.1 \times 10^5$ ).

Selection also affected the shape of the phenotypic distribution, which was now spread out broadly along the  $E_\theta$  axis ( $P = 1$ ) and had become narrow along the  $P$  axis (Fig. 4.4b and 4.4c). The expansion along  $E_\theta$  can be understood from the constraint: with paths to  $P < 1$  poorly accessible, the paths offering higher fitness lead along the  $E_\theta$  axis toward high or low expression values (Fig. 4.4b, black line). These fixed expression phenotypes may represent evolutionary dead ends, as they can be achieved by simply abolishing repressor-operator binding or IPTG binding capabilities.

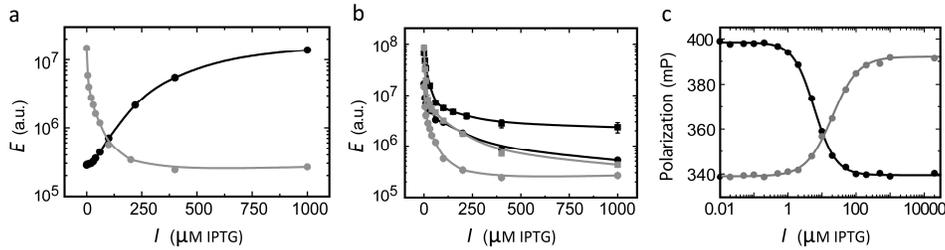
To test whether  $P < 1$  phenotypes could still emerge before the population was taken over by evolutionary dead ends, we performed a new cycle: 1/500<sup>th</sup> of the previous culture was taken, and the plasmid DNA was extracted; *lacI* was randomly mutated as before and placed back into fresh plasmids and hosts, which were subsequently grown in the same sucrose-Cm variable environment. The resulting isolates indicated that, whereas about half of the population remained at  $P = 1$ , the other half emerged below (Fig. 4.4b and 4.4c, gray spheres), showing that  $P = 1$  constituted not a global but a local (breakable) constraint. These results indicate a capacity of the system to evolve a fundamentally altered response to inducer before getting trapped in specialization. The evolved inverse phenotypes were distributed along a downward diagonal, indicating that they all had similar  $E_I$  and differed mainly in  $E_\theta$ .

After a third cycle of the same *lacI* mutagenesis and variable selection, the mean fitness of the population had further increased, first of all through decreased frequencies of fixed expression ( $P = 1$ ) phenotypes, which were no longer observed among the isolates (Fig. 4.4b and 4.4c, red spheres). In addition, the subpopulation of inverse phenotypes showed further fine-tuned improvements of the response toward smaller  $E_0$  and smaller  $P$ . Fig. 4.5a and 4.5b show induction profiles of evolved inverse phenotypes. The isolates were now distributed near the effective fitness optimum, which is taken as the average of the maximum observed growth rates in the two environments (1.85 db/hr). Some evolved phenotypes (4 out of 35) had a fitness within 10% of the optimum, which is comparable to the variability that we observed between the predicted fitness and results from competition experiments (Fig. 4.3c). Also note that small fitness increases near an optimum entail long fixation times, given the limited number of isolates that can be assayed. For example, a single mutant with a 10% fitness improvement over the rest of the population (of  $10^6$  cells) would require more than 80 hr of growth to be detected in the sampled isolates, as the required enrichment factor is about  $10^4$  (which is equal to  $2^{(0.1 \cdot 1.85 \cdot 80)}$ ). In conclusion, we find no evidence for global constraint in  $E_0$  or  $P$ , nor a rigid correlation between the two, that prevents access to the landscape optimum.

#### 4.7 Genetic basis of local constraint

Tracking evolution within phenotype fitness landscapes can identify local adaptive constraints but does not reveal the genetic architecture from which they originate. We sequenced several inverse *lacI* alleles to provide some insight into the genetic causes of the observed constraint at  $P = 1$  (Table 4.1). S97P appeared to be a key substitution: it occurred in the majority of genotypes and was essential for achieving the optimal inverse response. In a WT background, however, S97P yielded a  $P = 1$  phenotype, as was determined by constructing this mutant using site-directed mutagenesis, and as was also corroborated by previous studies [140,163]. It has been suggested that this serine residue, which is located at the dimer interface of the repressor, is central to the IPTG-induced allosteric transition in LacI [163,164]. This may explain the observed interference with induction by the S97P mutation.

Unlike the recurring S97P substitution, other genetic changes found in the inverse *lacI* alleles were diverse. For instance, one inverse phenotype harbored the mutations S69Y, Q131P, M242I, and a stop at L346 in addition to S97P, whereas another contained K108N, E235D, and Q352E together with S97P (Table 4.1). Though we could not identify a specific pattern or known functional effect for these additional mutations, their diversity indicated the presence of different possible mutational routes from S97P and  $P = 1$  to the optimum inverse phenotype.



**Figure 4.5. Induction curves**

a) Measured induction curves for wild-type LacI (black) and an evolved inverse LacI phenotype M-inv1 (gray). b) Induction profiles of a number of evolved inverse regulation mutants (at 1 mM from low to high expression: M-inv1, M-inv2, M-inv3, M-inv4; sequences given in Table 4.1). Standard errors ( $n = 3$ ) are smaller than the symbol size. c) Fluorescence polarization profiles of wild-type (black) and the inverse regulation mutant M-inv-1 of Figure 5a (grey). Fluorescence polarization (see materials and methods) here is a measure for the binding of purified repressor protein to TAMRA-labeled 18 base pair symmetric lac operator DNA. Polarization is recorded as a function of IPTG concentration, and higher polarization indicates a higher fraction of operator DNA bound by repressor. The data is fitted with a Hill curve with  $K_D = 6.0 \mu\text{M}$  and Hill coefficient 1.4 for wild-type and  $K_D = 20 \mu\text{M}$  and Hill coefficient 1.15 for M-inv-1. Error bars are standard errors over three or four measurements and smaller than the symbol size. See also Table 4.1.

To study the functional relevance of the observed mutations and their interactions, we engineered *lacI* variants containing subsets of the mutations found in one evolved inverse phenotype (M-inv-5; see Table 4.1). We found that, out of the six non-synonymous mutations, three (S97P, R207L, and T258A) were sufficient in a WT background to confer an inverse phenotype ( $P \sim 0.02$ ). Next, we engineered the double mutants S97P-R207L and S97P-T258A, which both gave a value for  $P$  of order 1, as did the single mutant S97P.

Thus, in the S97P background, R207L and T258A individually were nearly neutral but together conferred inversion, which indicates epistasis between R207L and T258A. R207L also exhibited epistasis with S97P; in the T258A background that displayed a WT phenotype ( $P$  of order 10), R207L was again neutral, but in the S97P-T258A background, R207L transformed a  $P = 1$  phenotype into the inverted phenotype. Overall, the results suggest that S97P is a rare and unique gateway at  $P = 1$  from WT at  $P > 1$  to the optimum at  $P < 1$  and support the notion that the phenotypic clustering at  $P = 1$  was due to genetic constraint.

M-inv1	M-inv2	M-inv3	M-inv4	M-inv5
S69Y	S97P	A72V	S97P	K84I
S97P	G315D	A92V	K108N	S97P
Q131P	P339H	C107S	E235D	R207L
M242I		S120R	Q352E	T258A
L346stop		L295M		L307H
		V301A		L349P
		T334M		
		L346stop		

**Table 4 1. Amino acid substitutions for evolved inverse phenotypes**, related to figure 4.5  
P-values of these mutants with 1 mM ONPF instead of IPTG were determined to be 0.34, 0.25, 0.45, 0.26, and 0.59, respectively (values in the order of the mutants in the table).

#### 4.8 Molecular mechanism of evolved phenotypes

Next, we questioned what molecular mechanism could underlie the regulatory change. One might imagine that the transcription factor now activates rather than represses by binding elsewhere in the regulatory region. However, no mutations were found in the DNA recognition domain of LacI, which would be required for this scenario. The inversion could also be explained by a mechanism in which the non-specific affinity to DNA is increased, as has been shown previously for LacI inversion [165,166]. An increased overall affinity then leads to binding to other cellular DNA, titrating transcription factor away from the operator and thus producing expression without IPTG.

In this scenario, the allosteric effect of IPTG on LacI actually remains the same as for WT, namely a decreased affinity for DNA. With IPTG, the mutant LacI would thus be liberated to start the operator search, which can result in repression if binding affinity to the operator is significant. However, given the recurrence of mutation S97P and its supposed role in the allosteric mechanism, we surmised that the inversion might be based on a modification of the allosteric transition. In this scenario, the evolved LacI would bind the operator with IPTG and have a low affinity for the operator in the absence of IPTG.

To discriminate these two scenarios, we first measured the expression response to the compound ONPF. ONPF is an anti-inducer for WT LacI that leads to increased DNA affinity, as seen by a lower expression level than achieved by WT repression. For inverted phenotypes with unchanged allostery but increased non-specific DNA affinity, addition of ONPF can be expected to further increase non-specific affinity and hence should not reduce expression, as has indeed been shown previously [167]. In contrast,

we find that our evolved inverted phenotypes do show reduced expression in response to ONPF, with values for  $P$  ranging between 0.25 and 0.59 for the five isolates indicated in Table 4.1. This result is in agreement with the allosteric scenario. In order to provide a direct test of changes in operator affinity, we performed operator binding assays with purified WT and an evolved inverse LacI protein (as assayed in Fig. 4.5) in the absence of nonspecific DNA (Fig. 4.5c). These data support the allosteric model: the evolved LacI is able to bind the operator but only in the presence of IPTG.

#### 4.9 Discussion

Recent decades have brought important advances to our understanding of how cellular regulation evolves, for instance by rewiring of regulatory networks [168], cooption of existing transcription factors to new regulatory regions [169] and ligands [170], evolving transcription factor protein functions [171], recombination of regulatory protein domains [172], and by the convergent evolution toward generic network motifs [173].

These efforts at the genetic, structural, and functional levels are complemented by studies of fitness consequences of regulatory changes in constant environments [120,121,174] and some in variable environments [158,175]. Fitness in variable environments has been extensively studied within evolutionary ecology, wherein the ability to alter the expressed phenotype in response to environmental changes is typically referred to as phenotypic plasticity [151,152]. However, direct experimental study of the adaptive evolution of phenotypic plasticity by mutation and natural selection [176] is a challenge because of limited insight into the genetic basis of the studied regulatory phenotypes and the selective forces in variable environments [157].

Here, we find that that the evolutionary trajectory in variable environments is altered by internal constraint of the regulatory system, but the evolutionary outcome is the optimal solution of a fitness tradeoff problem and hence determined by selection. The measured expression fitness relations of a founding genotype in constant environments predicted the optimal regulatory response to a challenging variable environment, whereby expression changes that are beneficial in one environment are detrimental in another. Cells were subsequently shown to evolve to the optimum response, in which the sensed cue (originally an inducer) acts as a corepressor. These findings indicate that regulatory evolution by selection in variable environments may be understood within the framework of multiobjective optimization theory [177], which addresses the maximization of overall performance in the presence of tradeoffs between conflicting objectives. It will be intriguing to test

the limits of prediction by optimization principles, for instance by studying more elaborate regulatory phenotypes and environments, and by determining the fundamental conflicts between objectives that are imposed by physico-chemical constraints.

The observed absence of global constraints in the synthetic system studied here raises the question of whether transcriptional regulation phenotypes found in microorganisms similarly are near-optimal biological functions. Insight into how transcriptional regulation systems achieve optimality would help to explain the adaptive origins of the highly specialized repertoire of signal detection and transmission capabilities found in transcription factor families [178], as well as the observed convergence toward generic regulatory network motifs [173].

Local constraints within the *lac* regulatory system that could be broken by selection affected the evolutionary trajectory. Whereas the population predominantly followed the direction of strongest selection within phenotype space, genetic constraint caused delay and deviation at a distinct boundary defined by a single parameter (at  $P = 1$ ). Upon the fixation of mutations that provide the functional innovation—in which IPTG acts as a corepressor rather than an inducer—a diverse set of genetic changes offered a fine-grained variation in both  $P$  and  $E_0$  that is essential to developing optimality in a novel function. The picture emerging from these observations is a genotype space consisting of multiple regions that confer different regulatory functions, which are distinct and connected by just a few mutations. The data highlight the potential of regulatory evolution by changes in protein coding regions, which complement other mechanisms such as the co-option of existing regulatory proteins to new regulatory regions.

Central to our strategy was the engineering of a model system that allowed us to independently measure the cross-environmental tradeoffs, which in turn informed us of the selective forces and constraint in regulation. This genetic engineering approach may offer a starting point for quantitative models describing the adaptive evolution of biological complexity, relating environmental conditions, regulatory architecture, adaptive constraint, and competition dynamics.

## 4.10 Materials and methods

### 4.10.1 Strains

In all selection experiments, *E. coli* K12 strain MC1061 (Casadaban and Cohen, 1980) was used, which carries a deletion of the complete *lac* operon. Genotype of MC1061: F<sup>+</sup>DlacX74 mcrB1 e14<sup>-</sup> (mcrA0) rpsL150 (StrR) galE15 galK16 D(ara,leu)7697 araD139  $\Gamma$  hsdR2(rk<sup>-</sup>,mk<sup>-</sup>) spoT1. This strain was obtained from Avidity LLC, Denver CO, USA, as electrocompetent strain EVB100 (containing an additional chromosomal birA). All growth and expression measurements, as well as the selection and competition experiments, were performed in Defined Rich medium (Teknova, Hollister, CA, USA), with 0.2% glucose as carbon source, and supplemented with 1mMthiamine HCl. For protein expression, we used the BLIM/pTara system [179] with an arabinose-inducible T7 polymerase and lacking a native lac repressor. After transformation of the pRSET-B (Invitrogen) plasmid expressing *lacI* into BLIM/pTara cells, all growth was performed in M9T medium [179] containing 0.5% glucose and the appropriate antibiotics.

### 4.10.2 Plasmids

Two plasmids were constructed based on the pZ vector system [180] in which the expression of the selection module is regulated by *lacI* (pRD007). The selection module consists of the co-expressed genes *lacZ $\alpha$* , *cmR*, and *sacB* under control of the pTrc promoter from pTrc99A [181] (which is amplified until base pair~300 before start). Reporter gene *lacZ $\alpha$*  consists of the first 364 base pairs of *lacZ*, amplified from the chromosome of strain MG1655 (CGSC stock center). Chloramphenicol resistance gene *cmR* originates from the pZ vector system. The levan sucrose 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 (pReplacZ $\Omega$  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 PlacIq-*lacZ $\Omega$*  fragment or by deleting pTrc99A for *lacI* and pTrc and inserting the MG1655 *plac-lacZ* fragment. Between *cis* and *trans* expression levels, an empirical relation was observed  $E_{cis} = 2.3310^4 * E_{trans}^{0.32}$ . For the production of wild-type and mutant *lac* repressor protein, the *lacI* coding sequence was inserted directly downstream of the enterokinase cleavage site of expression plasmid pRSET-B (Invitrogen). Plasmids and sequences are available upon request.

#### 4.10.3 Mutagenesis

Mutants were created using the Stratagene Genemorph II Random Mutagenesis kit. Mutagenized product was restricted and ligated into the (nonmutated) selection vector and subsequently transformed into *E. coli* strain MC1061 by electroporation. Pool sizes were routinely between  $5 \times 10^5$  and  $1 \times 10^7$ .

Throughout this experiment, mutagenesis conditions were constant. In order to determine the mutation rate, a random sample of mutants was sequenced after one mutagenesis round, yielding an average mutation rate of 0.003/bp ( $n = 9$ ).

#### 4.10.4 In vitro binding assay

Fluorescence polarization measurements were performed in the dialysis buffer with addition of BSA to 0.05%. Oligonucleotides containing the 18 base pair symmetric *lac* operator (ATTGTGAGCGCTCACAAT) and containing a 3'-carboxytetramethylrhodamine (TAMRA) fluorophore (Integrated DNA Technologies) were hybridized in a 10 mM Tris-Cl buffer at pH 8.5, by cooling down overnight in a water bath from 95°C to room temperature. The polarization assay was performed in a 384 well plate in a Victor 3V plate reader (Perkin Elmer) at 531 nm excitation and 595 nm emission. Each well contained 50  $\mu$ l dialysis buffer and 10 nM of *lac* operator. IPTG was added at appropriate concentrations. The amount of repressor protein was such that saturating binding could be observed (without IPTG for wild-type and with IPTG for M-inv-1). Each measurement was performed with 3 or 4 replicates

#### 4.10.5 Growth conditions during selection

Growth was performed at 37°C in 100 ml erlenmeyer flasks, under vigorous shaking. Culture medium was 20 or 40 ml EZ Rich Defined medium + 0.2% 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 hr of pre-selection, after which the cultures were grown for 6 hr. The 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  $10^6$ ). The pre-selection time of 3 hr (~6 generations for unselective growth) was chosen in order to be long enough to reach steady state enzyme expression, either due to protein production (in the order of 3 generations), or due to dilution (in which case 6 generations amounts to a dilution factor of 64). Optical density was monitored at 550 nm and whenever an  $OD_{550}$  of 0.1

is reached, a dilution was made into fresh pre-warmed selective medium. After selection, cultures were washed, and flash frozen. When transferred to the next environment (without mutagenesis), a threshold dilution of  $5 \times 10^2$  is applied, which determines the minimum growth rate for mutants to effectively increase in number in the previous environment.

#### 4.10.6 Growth conditions during measurement of the fitness landscapes

To measure growth rates for determination of the fitness landscape, wells of a 96-well plate containing 200 ml of Defined Rich glucose medium with the appropriate amount of IPTG were inoculated with a  $2 \times 10^4$  x dilution of an O/N (LB) culture, and grown for three hours (pre-selection) at 37°C in a Perkin & Elmer Victor3 plate reader until an optical density at 550 nm of around 0.0005 was reached (in the plate reader, which corresponds to an  $OD_{550}$  of 0.002). As this OD is too low to be determined, the same plate contains 6 wells that were inoculated with a mere  $5 \times 10^2$  times dilution, which reached a measurable OD of around 0.02 at the same time. At that moment sucrose or chloramphenicol was added.

Optical density at 550 nm was recorded every 4 min, and every 29 min 9 ml 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 (see examples in Fig. 4.6) the growth rate was obtained by determining for each well the increase in cell density at  $t = 6$  hr. From this the effective exponential growth rate (or Malthusian parameter, see e.g., [161]) was obtained according to  $F = (\log(OD_{t=6}/OD_{t=0})/\log 2)/6$ , in doublings per hour. In case the growth rate was high and stationary phase was reached within 6 hr, the initial slope of the growth curve was taken, since in the selection experiments the cultures were always diluted long before reaching stationary phase.

In the main text the fitness  $F$  is identified with the Malthusian parameter (2). This is the appropriate measure of the fitness for clonal organisms with variable generation times and overlapping generations, such as bacteria. The average growth rate over multiple environmental conditions is  $F = \sum_i p_i F_i$ , where the  $p_i$ 's denote the time fraction spent in environmental condition  $i$ , and  $F_i$  is the fitness for that condition. For the two conditions used in the main text at equal dwelling times, the fitness therefore can be expressed as  $F_{tot} = (F_{sucrose} + F_{cm})/2$  (see caption Fig. 4.2).

#### 4.10.7 Growth conditions during competition assays

Two mutants were mixed in a known ratio and subjected to selective environments (see above). After 6 hr of growth for a certain environment (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 population 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 with or without IPTG. As our selection module contains a *lacZ $\alpha$*  gene, complementation with the chromosomally expressed *lacZ $\omega$*  allowed for discrimination between the mutants and determination of their ratio, on the basis of their differential expression of *lacZ $\alpha$* . For example, WT and mutant N1 (Fig. 4.3b) can be distinguished on plates containing Xgal and IPTG, where WT forms blue colonies and N1 remains white. We can then calculate their fitness difference  $\Delta F = \log(A^{1,2}) / (6 \log 2)$ , where  $A^{1,2}$  is the factor by which mutant 1 out-grows mutant 2 during 6 hr of growth. The graph in Fig. 4.3c contains 15 data points for 3 competitions: WT versus N1, N1 versus R1, WT versus N2, each in 5 different environments: 0.4% sucrose, 0.25% sucrose, 0.15% sucrose, 80  $\mu\text{g}/\text{ml}$  cm + 1 mM IPTG, 25  $\mu\text{g}/\text{ml}$  cm + 1 mM IPTG.

#### **4.10.8 Growth conditions during the induction/de-induction assays**

To measure the rates of induction and de-induction after a switch in IPTG concentration, exponentially growing cultures expressing WT *lacI* were grown for four hours in the presence or absence of 1mM IPTG. Subsequently a time series for 3.5 hr was performed. At each time point 1 mM IPTG was added to a culture that had grown previously without IPTG. For cultures that had previously grown with IPTG, the medium was exchanged for medium without IPTG at each time point (after washing the cells). Growth was monitored such that the OD at 600 nm never exceeded 0.05. After 3.5 hr the expression level for each culture experiencing induction or de-induction for a different period of time was measured by an FDG assay (Fig. 4.6c, left).

#### **4.10.9 Assay of growth recovery after selective periods**

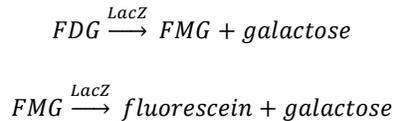
Growth during the periods of non-selective growth was assayed by plating cultures over time and counting the number of colony forming units (CFU). Cultures grown selectively for 6 hr (see above) were washed and grown in Defined Rich medium + glucose for 3.5 hr. At specific time points a part of the growing culture was taken, appropriately diluted, and spread on an LB plates. Plates were grown overnight at 37°C, after which colonies were counted (Fig. 4.6, right).

#### 4.10.10 $\beta$ -Galactosidase assay conditions

A reporter plasmid expressing either *lacZ* or *lacZ $\omega$*  was cotransformed into the mutant population or clone that is to be assayed. Cell cultures were grown at 37\_C in a Perkin & Elmer Victor3 plate reader, at 200  $\mu$ l 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 min, and every 29 min 9 sterile water is 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. Cells were fixed 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<sub>600</sub> of 0.05 to 0.23), by adding 20  $\mu$ l FDG-fixation solution (109  $\mu$ M fluorescein di- $\beta$ -D-galactopyranoside (FDG, MarkerGene Technologies Inc, Eugene, OR, USA, cat. nr M0250), 0.15% formaldehyde, and 0.04% DMSO in ddH<sub>2</sub>O). Fluorescence development was measured every 8 min (exc. 480 nm, em. 535 nm), as well as the optical density at 600 nm. Shaking and dispensing conditions as above. When cells are induced with IPTG, directly before or after fixation, an appropriate amount of inhibitive IPTG was added. Analysis of fluorescence trace is described below.

#### 4.10.11 Analysis of the FDG $\beta$ -galactosidase assay

Using the fluorogenic substrate fluorescein di- $\beta$ -D-galactopyranoside (FDG) allows for an accurate determination of the  $\beta$ -galactosidase activity over at least 4.5 orders of magnitude. FDG contains two galactose groups that both have to be cleaved in order to release fluorescein.



In ref. [182] a model for the FDG-FMG hydrolysis is proposed. In the concentration range of LacZ and FDG used in our experiments, the increase in fluorescence  $F$  is given by (Equation 7 in ref. [182]):

$$\frac{d}{dt}F = k_2E \frac{S_0}{K_m + S_0} (\alpha_p + (\alpha_M - \alpha_p)e^{-Rt}) \quad (S. 4.1)$$

where  $R$  is the relaxation constant (timescale 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  $\alpha$ 's are proportionality factors between the products and their fluorescence,  $F = \alpha_P P + \alpha_M M$  ( $P$  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 is proportional to  $E$ , though with different proportionality constants (first  $\alpha_M$ , then  $\alpha_P$ ).

Ref. [182] gives measured values for  $\alpha_M = 5.3 \mu\text{M}^{-1}$  and  $\alpha_P = 150 \mu\text{M}^{-1}$ . Although assigning arbitrary units to the fluorescence counts, they are relevant as relative quantities between FMG and fluorescein. Thus at  $t = 0$ , equation (S. 4.1) reduces to:

$$\frac{d}{dt} F = \alpha_M k_2 E \frac{S_0}{K_m + S_0} \quad (\text{S. 4.2})$$

To determine the enzyme concentration per cell, fitted slopes at  $t = 0$  are divided by the cell density. We use here  $\varepsilon \propto E/\text{OD}_{600}$ , where  $\varepsilon$  is the LacZ concentration per cell.

FDG expression measurements were compared to the standard Miller assay for  $\beta$ -galactosidase activity [144]. We measured an induction curve of wild-type LacI (as expressed from plasmid pRD007), both by using the Miller assay and the FDG assay described above (Fig. 4.6b).

#### 4.10.12 Growth models and interpolation of expression-growth curves

In order to interpolate the measured points on the expression-growth relations in Fig. 4.1c and 4.1d, we use models for the selective action of chloramphenicol and sucrose.

#### 4.10.13 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 [183],  $\mu \sim KX/X + K$ , where  $\mu$  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  $K/X + K$ . Therefore we adopt the following simple functional form for growth in the presence of chloramphenicol

$$\mu([cm])_{ext} = \frac{\mu_0}{c_1[cm]_{int} + 1} \quad (S. 4.3)$$

where  $c_1$  is a constant,  $\mu_0$  the growth rate in absence of  $cm$ , and  $[cm]_{ext}$  and  $[cm]_{int}$  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

$$C_{bar,cm}([cm]_{ext} - [cm]_{int}) = r_{acet,cm} \quad (S. 4.4)$$

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  $C_{bar,cm}$  times the concentration difference between inside and outside. The acetylation rate  $r_{acet,cm}$  is given by

$$r_{acet,cm} = k_{cat,cm}[E*cm] = k_{cat,cm} \frac{E_{tot}}{1 + \frac{K_m E_{cm}}{[cm]_{int}}} \quad (S. 4.5)$$

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

$$C_{bar,cm}([cm]_{ext} - [cm]_{int}) = k_{cat,cm} \frac{E_{tot}}{1 + \frac{K_m E_{cm}}{[cm]_{int}}} \quad (S. 4.6)$$

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

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

(S.4.7)

Expression-growth data for media containing cm in Fig. 4.1c were fitted with this equation, using the following parameter set in table 4.2.

#### 4.10.14 Sucrose growth

Sucrose selection is based on the formation of sugar chains (levan) in the periplasmic domain of gram-negative bacteria [184]. The enzyme catalyzing this polymerization reaction is levansucrase (SacB) from *Bacillus subtilis*. In the gram-positive *B. subtilis* 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. 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, we can write a modified Monod form

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

(S. 4.8)

where  $\mu_0$  is the growth rate in absence of sucrose and  $c_0$  is a constant. (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]). This can be solved to yield

$$\mu([sucrose]) = \mu_0 - c_0 r_{levan} \quad (\text{S. 4.9})$$

Analogous to the chloramphenicol selection, we write for the rate of levan formation and the equilibrium governing the transport of sucrose through the outer membrane

$$r_{levan} = k_{cat,sucr}[E * sucrose] = k_{cat,sucr} \frac{E_{tot}}{1 + \frac{K_m E_{sucr}}{[sucrose]_{int}}} \quad (\text{S. 4.10})$$

And

$$C_{bar,sucr}([sucrose]_{ext} - [sucrose]_{int}) = k_{cat,sucr} \frac{E_{tot}}{1 + \frac{K_m E_{sucr}}{[sucrose]_{int}}} \quad (\text{S. 4.11})$$

we can solve equation (S. 4.11) for  $[sucrose]_{int}$ , substitute this into equation (S. 4.9), which in its turn can be substituted into equation (S. 4.10) 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 (7,8). 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

$$\mu([sucrose]) = \mu_0 - c_0 r_{levan}^n \quad (\text{S4.12})$$

The obtained function provides good fits for the low-enzyme regime of the expression-growth data, with the parameter set in Table 4.3.

However, at the high  $[E]$  end, we observe a saturation at higher growth rates than equation (S. 4.12) can account for. There are at least three saturation effects (see also (7,8)) 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 (Fig. 4.1d), we used

$$\mu([sucrose]) = \frac{\mu_0 + \mu_{sat}}{c_0 r_{levan}^n + 1} - \mu_{sat} \quad (\text{S. 4.13})$$

which approaches growth rate  $\mu_{sat}$  for high expression levels.

#### 4.10.15 Repressor protein expression conditions and fluorescence polarization

For the production of wild-type and mutant *lac* repressor protein, the *lacI* coding sequence was inserted into expression plasmid pRSET-B (Invitrogen), in which a T7 promoter drives expression of the His-tagged repressor. The BLIM/pTara system (9) with an arabinose inducible T7 polymerase, and lacking a native *lac* repressor, was used for all protein expression. After transformation of the pRSET-B plasmid into BLIM cells containing pTara, all growth was performed in M9T medium (9) containing 0.5% glucose. 100 mg/l ampicillin and 40  $\mu$ g/l chloramphenicol was used to retain pRSET-B and pTara respectively, except when 1L cultures were grown for expression, in which case the chloramphenicol concentration was lowered to 15  $\mu$ g/l. Protein expression was induced at absorbance between 0.9 and 1.1 at 600 nm, by addition of 0.25% arabinose, after lowering the temperature to 17°C. Cells were harvested by centrifugation after 16 hr of induction. Subsequently the cell pellet was resuspended in a 50mM sodium phosphate buffer at pH 8.0 containing 500mMNaCl, 20mMimidazole, 2.5% glycerol, 1 mM DTT, 10 mM MgCl<sub>2</sub>, 0.1% tween 20, 20 mg lysozyme, and one tablet of protease inhibitor cocktail (Roche). Cells were lysed by sonication, ~2000U of DNase was added and allowed to incubate for 10 min at 4°C. The suspension was cleared by centrifugation at 4°C for one hour at 4.8\*10<sup>4</sup>g in a Sorvall SS-34 rotor. 0.5 ml of Ni-NTA agarose (QIAGEN) was added and incubated for 30 min at 4°C. The

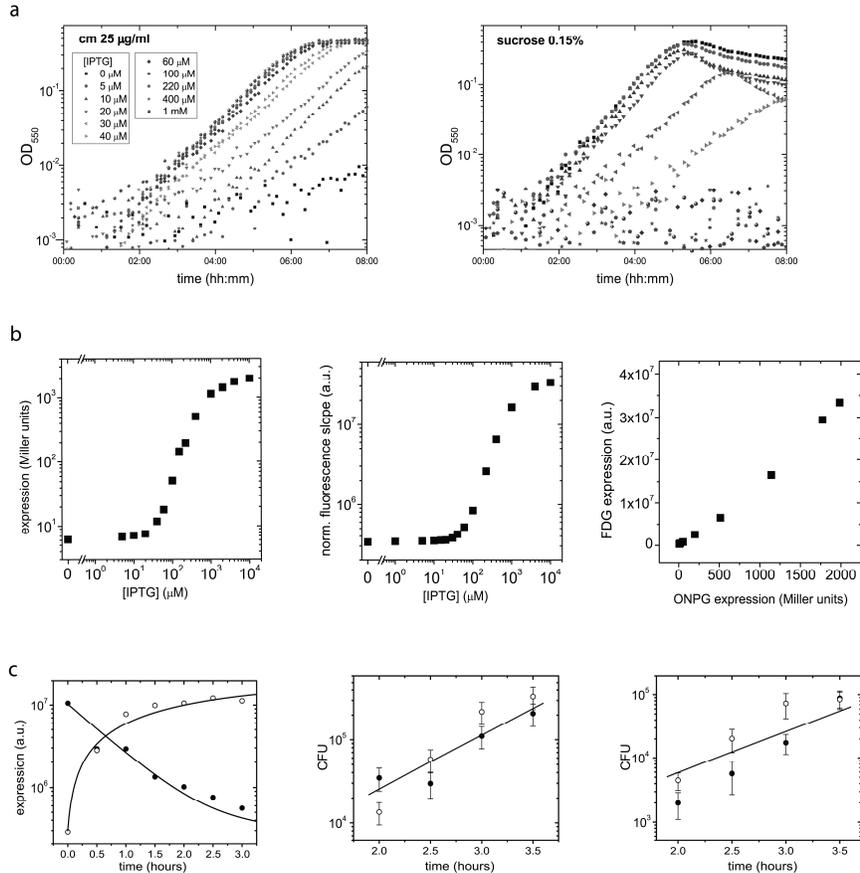
agarose was batch washed 5x in a 50 mM sodium phosphate buffer at pH 8.0 containing 500 mM NaCl, 20 mM imidazole, 2.5% glycerol, 1 mM DTT, and the protein was eluted with 250 mM imidazole. The protein solution was dialyzed overnight into a 50mMHEPES buffer at pH 8.0 containing 200mMNaCl, 20mMimidazole, 1% glycerol and 1mMDTT.

[sucrose]	0.4%	0.25%	0.15%
$\mu_0$	2.18	2.14	2.18
$c_0$	86	86	86
$n$	4	4	4
$10^7 * k_{cat}$	4.8	6.4	4.9
$K_m Esucr$	0.35	0.35	0.23

**Table 4.2 Parameter set used for the fitting of expression-growth relation in Fig. 4.1c**

[cm]	80 $\mu\text{g/ml}$	40 $\mu\text{g/ml}$	25 $\mu\text{g/ml}$
$\mu_0$	1.7	1.8	1.8
$c_1$	0.17	0.17	0.17
$n$	1	1	1
$10^5 * k_{cat}$	10.8	6.8	4.5
$K_m Ecm$	0.44	0.44	0.44

**Table 4.3 Parameter set used for the fitting of expression-growth relation in Fig. 4.1d**



**Figure 4.6 Growth rate and expression measurements**, Related to Figure. 4.1

a) Examples of measured growth curves of cells harboring the selection module, with 25 µg/ml chloramphenicol (left) or 0.15% sucrose (right). Optical density of the culture is recorded as a function of time in a 96-well plate reader. Different concentrations of the inducer IPTG are indicated.

b) Comparison of Miller assay and fluorescein di-β-D-galactopyranoside (FDG) assay. Shown is the expression in response to the inducer IPTG as determined by the Miller assay (left) and the FDG assay (center), and the FDG expression values against the Miller expression values (right).

c) left: induction and de-induction of WT LacI as a function of time. At  $t = 0$  either 1mM IPTG is added to a culture growing without IPTG (open symbols) or a culture previously growing at 1 mM IPTG is grown further in the absence of IPTG (solid symbols). See Supplemental Experimental Procedures A4 for details. The curves are theoretical expectations based on a growth rate of 2.15 doublings per hour. Center and right: Recovery after selective periods, as a function of time after transfer to non-selective medium. CFU is shown after 80 µg/ml chloramphenicol (center) with 1 mM (solid symbols) and 25 mM IPTG (open symbols), and 0.4% sucrose (right) with 1 mM IPTG (open symbols) and without IPTG (solid symbols). Straight lines indicate 2.15 doublings per hour.



## Chapter 5

# Environmental dependence of genetic epistasis in the *E. coli lac* repressor

*Evolutionary constraints have mainly been studied in constant environments, whereas evolution occurs mostly in heterogeneous environments. To investigate genetic constraints in heterogeneous environments, we have systematically reconstructed genetic intermediates between the wild type lac repressor, LacI, and three different evolved LacI variants with an inverse response to inducer. We find that one mutation (S97P) controls how the other mutations interact with inducer: in the absence of S97P, the other mutations change expression predominantly in the presence of inducer, and in the presence of S97P they change expression predominantly in the absence of inducer. The data implies that epistatic interactions within the LacI protein are profoundly altered by interactions with inducer molecules. These observed higher-order interactions between genotypes and the environment should fundamentally affect evolutionary dynamics and phylogeny, and show that epistasis perceived in constant environments does not properly inform on evolutionary constraints in the heterogeneous natural environment.*

## 5.1 Introduction

As pointed out by Wright in the 1930's, the genetic makeup of a biological system should determine not only current functionality but also affect future evolutionary change [33]. How the genetic architecture constrains evolution is only now starting to be addressed experimentally [61,71,96]. By systematically constructing single-mutant neighbors and assaying their function or fitness, proteins ranging from  $\beta$ -lactamases [71] to steroid receptors [170] have been shown to exhibit sign epistasis, in which one mutation can be beneficial or detrimental depending on the presence of a second mutation. This form of epistasis leads to a reduction in the number of trajectories that can be followed by fixing one adaptive mutation after another by positive selection [62]. Two mutations may individually be deleterious but jointly beneficial, as has been observed for mutations in MTH1 and HXT6/HXT7 that are involved in glucose metabolism in *Saccharomyces cerevisiae* [64] and between *arg* and *pyr* mutants in *Aspergillus niger* [185]. Such reciprocal sign epistasis and associated multi-peaked landscapes [146] represent a constraint that precludes the one-by-one fixation of positively selected mutations. Because of their ability to delay, divert, and arrest evolution, genetic interactions have been speculated to play a central role in speciation [186,187], the maintenance of biodiversity [188], and developmental evolution [92].

So far, epistatic interactions have been studied predominantly for environments that are constant in time and favor a single function or phenotype. However, natural environments are characterized by temporal variations and unpredictability, which rather impose temporally changing demands on the expressed phenotypes. Indeed, much of the complex regulatory systems are considered to have evolved in response to environmental heterogeneity [151,152]. Experimentally, mutational effects are commonly observed to differ in different environments [127] and mutations in alternative genetic backgrounds constitute different epistatic interactions depending on the environment in which the mutants are assayed [189,190].

These observations raise pressing questions on how environmental changes affect evolutionary constraint. If epistatic interactions between mutations that are required for functional innovation are turned 'on' and 'off' by environmental change, evolutionary dynamics could be drastically affected. For instance, evolutionary change that is constrained by adaptive valleys in constant environments could be opened up to positive selection in variable environments. Conversely, evolutionary changes that display positively selected paths in constant environments [61,71], could be hampered by environment-induced sign epistasis, thus blocking once accessible paths,

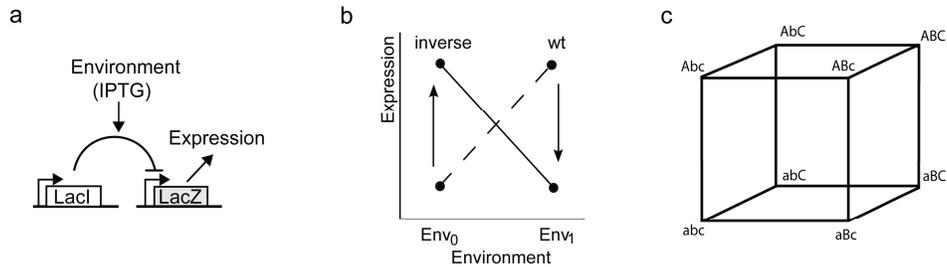
driving adaptation to unproductive regions of genotype space, or making positive selection dependent on the precise pattern of environmental variations.

These elementary issues are best addressed with a simple phenotype that responds to the environment. We focused on one of the best-understood model systems of environmentally regulated gene expression: the *Escherichia coli lac* repressor, LacI. We considered the evolutionary transition to a new function that constitutes the inverse of the wild type function: suppressing expression with IPTG rather than inducing it, and promoting expression in the absence of IPTG. We have previously isolated LacI variants with such inverse phenotypes in evolutionary experiments [191]. Here we reconstructed the intermediate mutants for three variants, assayed their expression level in the presence and absence of inducer IPTG, and analyzed the higher-order interactions between genetic and environmental changes.

We find that these three way genotype-environment interactions play a major role in the accessibility of the inverse LacI. A key mutation in the core of the repressor has a large beneficial effect in one environment in the *wild type* background, and opens up adaptive trajectories in the other environment when this mutation is present. The majority of genetic combinations display an environmentally dependent effect, which leads to the emergence of higher-order genotype-environment interactions. The phenotypic effect of these mutations is dependent on the internal molecular pleiotropy in the separate environments, thus the accessibility of the adaptive trajectory is determined by the exact sequence of both genetic and environmental changes.

## 5.2 Environmental dependence of sign epistasis

To investigate the interplay between the environment and the presence of epistasis we used the *E. coli lac* repressor as a model system (Fig. 5.1a). The wild type *lac* repressor, LacI<sub>wt</sub>, suppresses the expression of  $\beta$ -galactosidase, LacZ, in the absence of inducer, and relieves this repression in the presence of the inducer IPTG. Inverse *lac* repressors, LacI<sub>inv</sub>, do the opposite: they suppress expression in the presence of IPTG, and relieve this repression in the absence of IPTG (Fig. 5.1b). We investigate three such inverse LacI variants that were generated in evolutionary experiments [191], and contained three to six point mutations compared to LacI<sub>wt</sub>. For each variant, three mutations appeared essential for the inverse function, as was determined by engineering *lacI* variants that contained sub-sets of these mutations. We denote these three inverse variants as LacI<sub>inv1</sub> (S97P, T258A, R207L), LacI<sub>inv2</sub> (S97P, L349P, L307H) and LacI<sub>inv3</sub> (S97P, P339H, G315D). Note that all share the mutation S97P. Next, we constructed all single and double mutants for each variants,



**Figure 5.1 Functional description and schematic representation of genetic variants in the *lac* system.** a) Schematic representation of the genetic system in *E. coli*. The *lac* repressor, *LacI*, controls expression of *LacZ*. The system responds to IPTG. b) Environmental dependence of the expression level of *lacZ*. The expression level is measured in two environments: in  $Env_0$ , in the absence of IPTG and in  $Env_1$ , in the presence of IPTG. The wild type *LacI* (wt) is induced by IPTG in  $Env_1$  and has a low expression level in  $Env_0$  (dashed line). The functionally inverse *LacI* variant (inverse) exhibits high expression when IPTG is absent ( $Env_0$ ) and low expression levels in its presence ( $Env_1$ ) (black line). We consider the functional inversion from the wt to the inverse variant (arrows). c) Schematic representation of the genotype space of mutational intermediates from wt to the inverse variants,  $LacI_{inv}$ . *abc*- wt *LacI*, *ABC*-  $LacI_{inv}$ . Each allele *a*, *b*, *c* changes to *A*, *B*, *C* via  $3! = 6$  possible trajectories and six possible intermediates. There are three inverse variants that each contain three mutations:  $LacI_{inv1}$ - *a* to *A*: S97P, *b* to *B*: R207L, *c* to *C*: T258A.  $LacI_{inv2}$ - *a* to *A*: S97P, *b* to *B*: L307H, *c* to *C*: L349P.  $LacI_{inv3}$ - *a* to *A*: S97P, *b* to *B*: G315D, *c* to *C*: P339H.

and assayed their *LacZ* expression phenotypes in an environment without IPTG ( $Env_0$ ) and in an environment with 1mM IPTG ( $Env_1$ ) (Table 5.1), using a fluorogenic  $\beta$ -galactosidase reporter assay (materials and methods).

The three mapped genotype spaces can be depicted as cubes, where vertices denote genotypes and edges mutational steps (Fig. 5.1 c). The six faces of the cubes represent the pairs of mutations that may interact. Interactions are categorized either as magnitude (M), sign (S), or reciprocal sign (R) epistasis (Table 5.2). In the case of sign epistasis, we indicate in brackets the mutation whose effect on expression alters in sign. Mutations with measured effects smaller than the measurement error are considered neutral.

We find that seven out of the eighteen pairs of mutations show the same category of epistasis in both environments. For instance, P339H and S97P in a  $LacI_{wt}$  background (Table 5.2,  $LacI_{inv3}$ ) exhibit sign epistasis in both environments. Note that for all these seven pairs, the *magnitude* of the mutational effect does depend on the environment (see Table 5.2). However, the signs of the mutational effects are not altered by the environment, and so the category does not change.

Variant	Expression in Env <sub>0</sub> (a.u)	Expression in Env <sub>1</sub> (a.u)
Wild type	16.0 ± 0.5	256.0 ± 11.5
<b>LacI<sub>inv1</sub></b>		
R207L	18.2 ± 1.4	352.2 ± 39.4
T258A	15.9 ± 1.0	264.8 ± 29.0
R207L-T258A	20.5 ± 0.2	708.0 ± 43.0
S97P	30.7 ± 1.7	19.4 ± 3.4
S97P-R207L	25.2 ± 1.3	30.1 ± 6.3
S97P-T258A	93.5 ± 17.7	35.3 ± 5.3
S97P-R207L-T258A	1846.3 ± 140.4	31.4 ± 1.4
<b>LacI<sub>inv2</sub></b>		
L307H	22.4 ± 1.3	1400.0 ± 122.9
L349P	19.1 ± 2.7	494.3 ± 50.1
L307H-L349P	24.2 ± 5.3	1633.3 ± 342.7
S97P	30.7 ± 1.7	19.4 ± 3.4
S97P-L307H	3616.7 ± 624.4	72.1 ± 7.9
S97P-L349P	48.2 ± 3.9	34.4 ± 7.4
S97P-L307H-L349P	383.0 ± 56.0	21.8 ± 1.9
<b>LacI<sub>inv3</sub></b>		
G315D	15.0 ± 0.5	232.7 ± 8.3
P339H	16.1 ± 1.0	216.1 ± 6.0
G315D-P339H	14.7 ± 1.1	399.1 ± 37.8
S97P	30.7 ± 1.7	19.4 ± 3.4
S97P-G315D	399.30 ± 150.2	8.6 ± 0.8
S97P-P339H	63.9 ± 2.2	23.4 ± 3.2

**Table 5. 1 Expression level of genetic variants in two environments.** The expression level of LacZ was measured in two environments by a fluorogenic reporter assay. Env<sub>0</sub>, in the absence of IPTG and Env<sub>1</sub> in the presence of IPTG. Errors are standard deviations, n=3.

The other eleven mutation pairs display a different category of epistasis in the two environments (Table 5.2). Some sign epistatic interactions are switched ‘off’ by adding IPTG: in the LacI<sub>wt</sub> background for instance when R207L interacts with S97P, IPTG induces a sign change in the effect of R207L, thus transforming the sign-epistasis without IPTG into magnitude epistasis with IPTG (Table 5.2). Sign epistasis is turned ‘on’ between other mutations. For instance, in a S97P background, L349P and L307H exhibit sign epistasis without IPTG and reciprocal sign epistasis with IPTG (Table 5.2).

Thus, the presence or absence of sign epistatic interactions between residues involved in LacI inversion appears to be highly environment-dependent.

Genetic Interaction			Lacl <sub>inv1</sub>		Genetic Interaction			Lacl <sub>inv2</sub>		Genetic Interaction			Lacl <sub>inv3</sub>	
S97P	R207L	T258A	Env <sub>0</sub>	Env <sub>1</sub>	S97P	L307H	L349P	Env <sub>0</sub>	Env <sub>1</sub>	S97P	G315D	P339H	Env <sub>0</sub>	Env <sub>1</sub>
X	X	0	S (R207L)	M	X	X	0	M	M	X	X	0	S (G315D)	M
X	0	X	S (T258A)	M	X	0	X	S (L349P)	M	X	0	X	S (P339H)	S (P339H)
1	X	X	S (R207L)	M	1	X	X	S (L349P)	R	1	X	X	M	M
X	1	X	M	M	X	1	X	M	S (L349P)	X	1	X	S (P339H)	M
0	X	X	S (T258A)	M	0	X	X	S (L307H)	M	0	X	X	M	R
X	X	1	M	M	X	X	1	S (L307H)	S (L307H)	X	X	1	S (G315D)	S (G315D)

**Table 5. 2 Genetic interactions in Lacl variants and their environmental dependence in two environments.** The genetic interactions are drawn for three inverse Lacl variants. Each row details the interactions between two mutations, indicated by an x for each mutation, in a Lacl<sub>wt</sub> background (0), or a single mutant background (1). We consider three types of interactions: S, sign epistasis; M, magnitude epistasis; R, reciprocal sign epistasis. The mutation responsible for the sign change is indicated in between brackets.

### 5.3 Visualizing higher-order genotype-environment interactions

The observed dependence of epistasis on the environment (Table 5.2) does not show generic trends, in the sense that the pattern of epistatic interactions and overall dependence on the environment are different for each variant. To gain more insight in the underlying higher-order interactions we introduce a graphical method (Fig. 5.2a). In this method, mutations are represented as vectors in a two-dimensional coordinate system, where the axes indicate the corresponding changes in expression phenotype in both environments.

Each mutational vector points to one of the four quadrants of a coordinate system that has its origin at the vector's initial point. A vector pointing to quadrant I signifies functional improvements in both environments, quadrants II and IV denote improvement in one environment and deterioration in the other, and quadrant III denotes deterioration in both (Fig. 5.2a). The probability of fixing neutral mutations is low when compared with mutations that confer functional improvements [62,192]. Mutations that are neutral in both environments are therefore classified as in quadrant III, while mutations that are neutral in one but confer functional improvements in the other are classified as in quadrants II or IV. Thus, vectors in quadrants II and IV indicate mutations that exhibit genotype-environment interactions, while those in quadrants I and III do not.

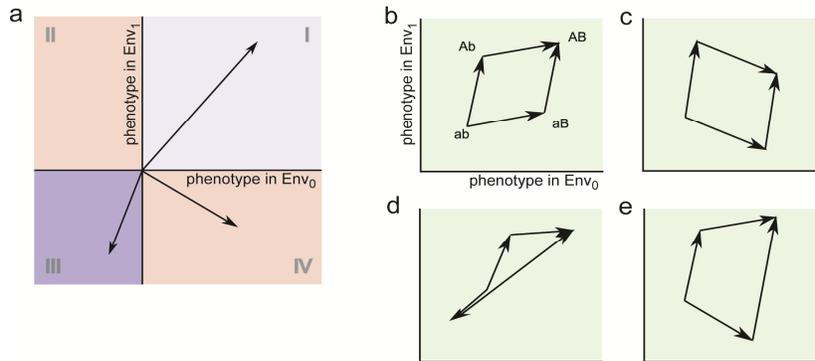
Higher-order epistatic interactions between mutations and the environment can be visualized by paths composed of two or more mutational vectors. Within the two-dimensional coordinate system, the two mutational paths from genotype *ab* to *AB* (*via* *Ab* or *via* *aB*) form a four-sided polygon. Opposite sides of the polygon represent the same mutation in different genetic backgrounds (for instance *a* to *A* in background *b* and *B*). In the absence of any genetic interactions, the polygon is a simple parallelogram (Fig. 5.2b and c). Its parallel and equally long opposing sides reflect that the mutations have the same phenotypic effect in both environments, regardless of the genetic background. When both mutations display phenotypic improvements in both environments, all vectors are positioned in quadrant I (Fig. 5.2b). Genetic interactions may be absent even as the mutations do interact with the environment: at least one pair of the opposing sides is then positioned in quadrant II or IV (Fig. 5.2c).

Deviations from the parallelogram indicate genetic interactions, or epistasis. Vectors of opposing sides that have different angles but point in the same quadrant signal magnitude epistasis. Sign-epistasis is represented by vectors of opposing sides pointing in different quadrants: the corresponding mutation then confers improvements in one genetic background and deteriorations in another. When the sign-change occurs irrespective of the environment (Fig. 5.2d), one can consider it to be a two-way interaction, between the two mutations only. Sign-change triggered by the environment in the opposing sides of the polygon (Fig. 5.2e), reflects three-way interactions between two mutations and the environment. Taken together, higher-order interactions between mutations and the environment are signaled by polygons that have opposing sides in different quadrants, where one or more sides lie in quadrants II and IV (Fig. 5.2e).

#### 5.4 Higher-order genotype-environment interactions in *Lacl*

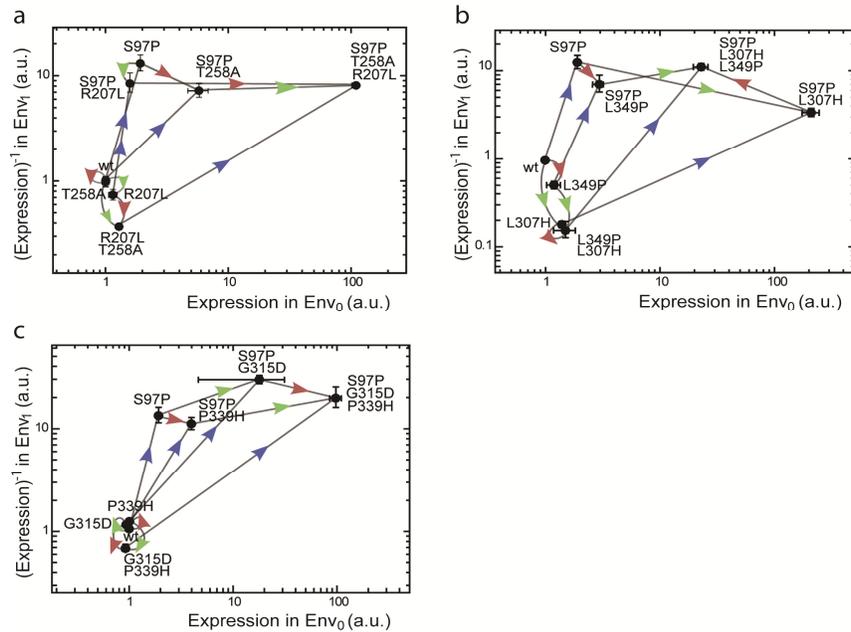
The higher order genotype-environment interactions in *Lacl* are analyzed by displaying the expression data as mutational vectors in Figure 5.3a, b and c. The functional transition from a wild type to an inverse response involves increasing expression in the absence of IPTG ( $Env_0$ ), and decreasing expression in the presence of IPTG ( $Env_1$ ). For clarity, we therefore ranked the expression values by *increasing* in  $Env_0$ , and by *decreasing* order in  $Env_1$ . In this manner, inversion progresses when moving towards the upper-right corner of Fig. 5.3.

While details differ, the overall pattern displayed by the three variants is strikingly similar when represented with this method, which contrasts with the diversity seen in Table 5.2.



**Figure 5.2 Method for analyzing higher order genotype-environment interactions**

a) Schematic representation of the effect of mutations on phenotype in two environments. Mutations are represented as vectors with the start in the origin of the coordinate system. Mutations that improve the phenotype in both environments,  $Env_0$  and  $Env_1$ , are located in quadrant I. Mutations that improve the phenotype in one environment and deteriorate the phenotype in the other environment, are located in quadrant II or IV. Mutations that are detrimental in both environments are located in quadrant III. b, c, d, e) Schematic categorization of the interactions of two mutations in two environments. The two mutational paths from genotype  $ab$  to  $AB$  form a polygon. Opposite sides of the polygon represent the same mutation in different genetic backgrounds (eg.  $a$  to  $A$  in background  $b$  and  $B$ ). b) No epistasis or genotype-environment interactions. The polygon is a simple parallelogram, with all vectors positioned in quadrant I or III, when genetic and environmental interactions are absent. Here, both mutations display phenotypic improvements in both environments. c) Genotype-environment interactions. The polygon is a parallelogram with opposing sides in the same quadrant, of which at least one pair of opposing sides lie in quadrant II or IV, when mutations exhibit only genotype-environment interactions. Here, mutation 'B' exhibits a genotype-environment interaction and is located in quadrant IV. d) Sign epistasis in both environments. The polygon contains non-parallel sides, one vector is turned to quadrant III in the presence of sign epistasis but the absence of environmental interactions. Here mutation 'B' in the 'a' background exhibits sign epistasis in both environments. e) Higher-order genotype-environment interactions. The polygon contains non-parallel sides of which at least one vector is located in quadrant II or IV, in the presence of higher-order genetic-environment interactions. Here, mutation 'B' in the 'a' background exhibits sign epistasis in  $Env_1$ .



**Figure 5.3 Mutational vector plot of three inverse LacI variants.**

The three inverse LacI variants all contain three mutations. Each mutation is represented by a vector (see Fig. 5.2). The axes indicate expression without IPTG in  $Env_0$  and expression with IPTG in  $Env_1$ . Expression along the Y-axis is represented as  $1/\text{expression}$  and expression levels in both environments are normalized to the *wt* level. The inverse, triple mutant, is located in the upper right corner of the plot. a)  $LacI_{inv1}$  S97P (blue), R207L (green), T258A (red). b)  $LacI_{inv2}$ , S97P (blue), L307H (green), L349P (red). c)  $LacI_{inv3}$ , S97P (blue), G315D (green), P339H (red).

In all variants, S97P affects expression in most genetic backgrounds and in both environments, though most strongly in  $Env_1$  and the  $LacI_{wt}$  background (Fig. 5.3, blue vectors). On the other hand, the green and red vectors corresponding to the other two mutations appear to predominantly point downward along the  $Env_1$  expression axis (and are thus nearly neutral in  $Env_0$ ) or to the right along the  $Env_0$  expression axis (and are thus nearly neutral in  $Env_1$ ).

Mutation S97P appears responsible for the rotation of the red and green vectors: in the absence of S97P, all red and green vectors point along  $Env_0$  while in the presence of S97P the red and green vectors point along  $Env_1$ . Thus, the polygons formed by the blue and red or green vectors deviate from a parallelogram, which would have for instance resulted from red and green vectors pointing along  $Env_0$  both before and after S97P.

This deviation from a parallelogram indicates higher-order interactions between two genetic changes and the environment. In other words, S97P represents a 'switch' that changes the interaction of these mutations with the environment.

Some pairs of mutations show a lack of higher-order interactions. For instance in Fig. 5.3c, the red and green vectors form a parallelogram in the S97P background, with opposite sides pointing in the same direction. The polygon is tilted, with the red vectors pointing in quadrant IV, which indicates environment-genotype interactions. Most mutational pairs do exhibit higher order interactions however, of different types. For instance, in Fig. 5.3b in the S97P background, L307H rotates the red vector (L349P) from quadrant IV to II. Overall, the vectorial rotations indicate pervasive higher-order interactions between genetic and environmental changes in LacI.

How do the observed higher-order interactions affect the development of inversion in the wild type *lac* repressor? In the *wild type* background, the red and green vectors along the Env<sub>1</sub> axis point towards higher expression, which actually distances them from inverse phenotypes (Fig. 5.3, upper-right corner). For environments favoring inversion, the corresponding mutations would therefore not be fixed by positive selection in Env<sub>1</sub>. In the LacI<sub>wt</sub> background, S97P does yield functional gains towards inversion, but predominantly in Env<sub>1</sub>. In Env<sub>1</sub> S97P thus yields a strong selective benefit. Once fixed, one or both of the other mutations provide functional gains, but now mainly in Env<sub>0</sub>. Indeed, most paths sustain functional losses in either Env<sub>0</sub> or Env<sub>1</sub>. Thus, the observed higher-order interactions between multiple mutations and the environment make access to inverse phenotypes dependent on a specific order of mutations and environmental changes.

### 5.5 Relation between higher-order interactions and system architecture

Given the generic nature of the pattern seen in the mutational vector plots, we surmised that the underlying interactions might stem from general characteristics of the system architecture. The functioning of the *lac* regulatory system has been modeled by the classic Monod-Changeux-Wynman (MCW) description of allosteric transitions [193]. Here, the *lac* repressor exhibits equilibrium fluctuations between a 'binding' state, which has high affinity for the operator, and a 'non-binding' state with low affinity. In the absence of ligand, the repressor is predominantly in the binding state, which results in tight repression. Interaction with the inducing ligand shifts the balance to the non-binding state, resulting in increased expression levels.

The three variants studied exhibit an inverse response compared to that of the wild type *lac* repressor, with a high affinity for the *lac* operator in the presence of ligand and a low affinity without it [191]. Within the MWC framework, the nature of

the states would then be swapped: in the absence of ligand, the repressor is predominantly in the binding state, and interaction with the ligand shifts the balance to the non-binding state.

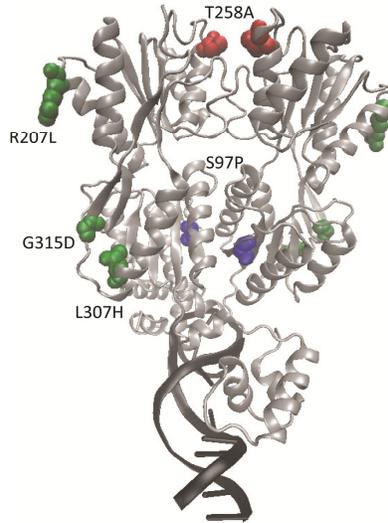
We find that the addition of the two additional mutations in the S97P background not greatly affects the expression in the presence of IPTG in all three variants. Also in the absence of IPTG and S97P, expression is only little affected when the two additional mutations are added (Fig. 5.3 a,b,c), which suggests the mirroring of the effects of the two additional mutations in the binding states in  $\text{LacI}_{\text{inv}}$  and  $\text{LacI}_{\text{wt}}$  in the two environments.

In the absence of IPTG and the presence of S97P, expression is greatly increased upon addition of the two additional mutants in all three variants. Expression by the addition of the two mutations in the  $\text{LacI}_{\text{wt}}$  background in the presence of IPTG is also affected, though not as strong as the former (Fig. 5.3 a,b,c). Although the observed effect is smaller in  $\text{Env}_1$  than in  $\text{Env}_0$ , this also suggests a mirroring of the effects of the two additional mutations on the non-binding states in the presence of IPTG in the  $\text{LacI}_{\text{wt}}$ , and  $\text{LacI}_{\text{inv}}$  in the absence of IPTG. The mirroring of the 'binding' and 'non-binding' states may either be brought about by an inverted allosteric change in the inverse variants or another mechanism.

An alternative mechanism that has been proposed [194] is a swap between the wild type 'binding' state and an unfolded state in the absence of IPTG for the inverse variants, complemented with a swap between the wild type 'unbinding' state and the IPTG stabilized 'binding' state for the inverse variants. Which mechanism is responsible for the inversed phenotype has not been resolved, yet. However, we do find that also on structural level, the three mutations in the three inverse variants show similarities, which suggests a common mechanism of inversion.

The functional domains in which the mutations occur do reveal similarities between the three inverse variants. Three properties of a functional repressor are expected to give rise to pleiotropic compromises or trade-offs when mutating  $\text{LacI}$ : stability, allosteric transition and IPTG recognition. Since we do not alter the head piece we naively expect no changes in the DNA binding properties. Here we relate the structural information of the mutations to the published literature, and we speculate about the role of these mutations regarding the stability, allosteric transition and IPTG recognition. The three mutations in all three  $\text{LacI}_{\text{inv}}$  seem all to serve one, or a combination of these functions and all have pleiotropic effects on the functionality of the protein in the two environments.

For instance, residue 258 in  $\text{LacI}_{\text{inv}1}$  is positioned within the dimerisation interface that is solvent accessible. Replacing the hydrophilic threonine for a hydrophobic alanine at this site could decrease the dimerization stability, which is



**Figure 5. 4 Mutations from the three inverse variants mapped on the LacI crystal structure (1EFA, pdb).** The mutated amino acids of all three inverse variants are depicted as space filling residues in the wild type structure. They are color coded on basis of their functional grouping. The tetramerization domain is absent in this structure. Red residue, involved in multimerisation of the protein. Green residues, connects to a residue involved in IPTG binding via an alpha helix. Blue residue, involved in allosteric transition.

consistent with the measured expression increases (in the S97P and S79P-R207L backgrounds). Residues 349 in  $\text{LacI}_{\text{inv}2}$  and 339 in  $\text{LacI}_{\text{inv}3}$  are located within the tetramerization domain. The former is part of a leucine heptat repeat, and the proline for leucine replacement at that site is expected to distort the leucine heptat repeat in [195] tetramerisation helices [195,196], hence altering tetramer stability. The latter residue is located just in front of the tetramerization helices [196], and substitutions may thus alter its orientation. These changes are consistent with the observed expression increases for L349P and P339H in the S97P background.

Similarly, each inverse variant contains one mutation near an alpha-helix that interacts with IPTG (Fig. 5.4, green residue). R207L in  $\text{LacI}_{\text{inv}1}$  is part of a loop attached to an alpha-helix that harbors residue S193, which interacts with D149 and IPTG. The interaction between S193 and D149 causes a contraction of the N-subdomain during the allosteric transition [163,197,198]. L307H in  $\text{LacI}_{\text{inv}2}$  is located in the alpha-helix that harbors F293 which contacts IPTG. The replacement of leucine for histidine is known to destabilize an alpha helix [199]. In the presence of S97P it does increase expression in both environments in our data, which is in agreement with a destabilization of the protein. G315D in  $\text{LacI}_{\text{inv}3}$  is linked to this same helix by a

loop and it is adjacently connected to another helix in the core pivot that undergoes a local displacement during the allosteric transition [200,201].

Finally, all inverse variants contain the mutation S97P that is positioned in the core of the dimer interface [164,196]. In  $\text{Lacl}_{wt}$ , the nearby K84 undergoes a movement that allows the formation of a beta-sheet which involves both monomers upon IPTG binding [163,164,202,203,204,205]. During this transition the K84 side-chain moves to a transient position close to the side chain of S97, before contacting other residues [163]. This transient bond cannot be formed when S97 is mutated to proline because it lacks serine's amide hydrogen. Consequently, the conformational changes following the K84 movements and thus induction would be impaired. Indeed we find S97P to be quite insensitive to inducer (Fig. 5.3). These molecular interactions in the  $\text{Lacl}$  structure, in the presence or absence of IPTG, lie at the basis of the observed higher-order genotype-environment interactions.

## 5.6 Conclusion and discussion

Recently developed approaches to systematically map the adaptive landscapes of specific biological functions have provided a first view on the causes of constraint in functional innovations. Proteins have been shown to exhibit epistatic interactions between mutations that limit the number of mutational trajectories that can be fixed by positive selection in constant environments [61,71]. Directed evolution of proteins analyzed in the context of measured selective pressures have revealed genetic constraints that delay or restrict evolutionary adaptation [92,191], whereas tradeoffs demonstrated constraints of adaptation to different environments [206,207]. Here we show that the  $\text{Lacl}$  protein contains pervasive higher-order interactions between the environment and multiple mutations that together confer functional innovation.

The interactions appeared characterized by two interlocking switches, one that is flipped by the environment, and one by a genetic change. The IPTG-switch turned (nearly) neutral mutations into mutations that confer significant functional gains, thus opening-up a pathway to strong positive selection. The genetic S97P-switch flipped this environmental interaction upside-down, with IPTG now turning advantageous mutations into nearly neutral ones, thus restricting pathways by reducing the selective pressure in the presence of IPTG. Overall, the higher-order interactions between genetic structure and the environment awards the external environment critical influence on internal constraints of the biological system.

The high amount of higher-order genotype-environment interactions in our data indicates that the beneficial effect of mutations depends not only on the required genetic background but also on the environment. This suggests that there is a

deviation from a constant rate of accumulation of adaptive mutations in variable environments since both genetic and environmental backgrounds may need to coincide for the mutation to be fixed by positive selection. This dependence can lead to longer waiting times for beneficial mutations, and thus induces heterotachy, in which specific evolutionary rates change across the tree [208,209]. Heterotachy, originating from the ignorance of co-dependencies between both genetic as well as environmental interactions, can result in topological inaccuracies in phylogenetic trees [210] such as long-branch biases [211] and a lack of phylogenetic resolution [211,212].

It is striking that although two of the three mutations differ between the three inverse variants, they all show a similar phenotype. This means that there is redundancy in the genetic architecture, which can generate the same functional phenotype. This feature is also known as parallelism, in which descendants from a common ancestor develop similar traits, yet contain other genetic substitutions. Parallelism in evolving microbial populations has been observed in other microbial experiments, albeit in constant environments [174,213,214]. Here we find that also in variable environments the same phenotype can be obtained via different genetic trajectories.

It will be of interest to explore whether the higher-order interactions between environment and genotype observed here in a simple model system are more generally present in biological systems. The ability to control internal developmental programs and to alter phenotype in response to the environment is one of the defining properties of living systems. The intrinsic functional interdependence among regulatory proteins, and between regulatory systems and the environment, within these biological functions and the complex regulatory networks that underlie them, suggests that such higher-order interactions are not confined to model systems or individual proteins. We thus expect both genetic architecture and environment to be crucial for shaping the evolutionary potential of the future.

## 5.7 Materials & Methods

### 5.7.1 Strains

Escherichia coli K12 strain MC1061 [215], which carries a deletion of the *lac* operon was used in all experiments. This strain was obtained from Avidity LLC, Denver CO, USA, as electrocompetent strain EVB100 (containing an additional chromosomal *birA*). Plasmid pRD007 was constructed based on the pZ vector system [216] and contains *lacI*, driven by the PL01-tet promoter. The reporter plasmid pReplacZ was created by deletion of *lacI* and *ptrc* in *ptrc99A* [181] followed by insertion of the *plac-lacZ* fragment of MG1655 [87].

### 5.7.2 Media

In all experiments EZ defined rich medium (Tecknova, Hollister, CA, USA) with 0.2% glucose and 1 mM thiamine HCL (Sigma) was used. IPTG was purchased from Sigma.

### 5.7.3 Reconstruction of (intermediate) mutants

Non-synonymous base pair mutations were introduced into the coding region of *lacI* by site-directed mutagenesis with the QuickChange® II-E Site-Directed Mutagenesis Kit (Stratagene, USA) according to the manufacturer's protocol, followed by DpnI digestion of the native plasmid pRD007 [191](or pRD007 plus mutation(s) for construction of double or triple mutants) that was used as template. With primers:

A772G\_fwd: GCGGCCATTGCCGAGTCCGGGCTG;  
A772G\_rev: CCGGACTCGGCAATGGC GGCATTG;  
C1016A\_fwd: CAAACCGCCTCTCACCGCGGTTGG;  
C1016A\_rev: CCAACGCG CGGTGAGAGGCGGTTTG;  
G620T\_fwd: CATAAATATCTCACTCTCAATCAAATTC;  
G620T\_rev: GAATTTGATTGAGAGTGAGATATTTATG;  
G944A\_fwd: GGCGGTGAAGGACA ATCAGCTGTTG,  
G944A\_rev: GCAACAGCTGATTGTCCTTCACCGCC;  
T289C\_fwd: CAGCG TGGTGGTGCCGATGGTAGAACGAAG;  
T289C\_rev: GTTCGTTCTACCATCGGCACCACCACGCTG;  
T1046C\_fwd: GATTCATTAATGCAGCCGGCAGCACAGGTTTCCC;  
T1046C\_rev: GGGAAACCTGTCGTGCCGGCTGCATTAATGAATC

#### 5.7.4 Expression measurements

Cultures were grown at 37°C in a Perkin & Elmer Victor3 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 min, and every 29 min 9 sterile water is 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. Cells were fixed after the cultures had reached an optical density of at least 0.015 and at most 0.07, by adding 20 µl FDG-fixation solution (109 mM fluorescein di-b-D-galactopyranoside (FDG, Enzo Life sciences, NL), 0.15% formaldehyde, and 0.04% DMSO in MilliQH2O). Fluorescence development was measured every 8 min (exc. 480 nm, em. 535 nm), as well as the OD<sub>600</sub>. Shaking and dispensing conditions as above. When cells are induced with IPTG, directly before or after fixation, an appropriate amount of inhibitive IPTG was added. Analysis of the fluorescence trace is as described in chapter 4 and [191].

## Chapter 6

# Crossing multipeaked landscapes in variable environments

*Sub-optimal peaks in adaptive landscapes are recognized for their ability to constrain the progress of evolution, even as empirical evidence for their existence is scarce. Here we experimentally reconstruct a multipeaked adaptive landscape and investigate a counter-intuitive suggestion: can the valleys nonetheless be crossed by positive selection in a variable environment? We systematically mapped the sequence space in between a sub-optimal lac repressor-operator variant and the wild type lac system, which involved 6 intermediates and 720 different mutational trajectories. We measured the expression of the downstream genes regulated by the constructed lac-operator variants in the absence and presence of inducer IPTG. We find that the corresponding constant-environment landscapes are multipeaked: without IPTG, all trajectories from the sub-optimum to the wild-type exhibited decreases in the ability to repress expression, and with IPTG, all trajectories exhibited decreases in expression of the downstream genes. We find that the adaptive landscapes in the two environments are negatively correlated, mutations that increase repression without IPTG decrease expression with IPTG, and vice versa, which leads to the occurrence of tradeoffs in all trajectories. Counter-intuitively, these tradeoffs can allow an evolving population to follow trajectories that alternate between the fixation of a beneficial mutation in one environment, and the fixation of the next mutation in the second environment, ultimately arriving at the global optimum. Under the assumption that mutations co-occur with their suitable environment, there are 273 out of 720 trajectories that can cross this multipeaked fitness landscape by fixing adaptive mutations one-by-one. This indicates that tradeoffs not only constrain evolution in variable environments, but can also aid the adaptive progress in variable environments.*

## 6.1 Introduction

The evolutionary implications of multiple fitness peaks have been debated ever since Wright introduced this idea in the 1930's [33]. Multi-peaked fitness landscapes have been speculated to underlie a diverse array of phenomena, ranging from speciation [57] to transitions in developmental evolution [217]. Central to the effects of multiple fitness peaks are the limitations they impose on the adaptive process. For single-peaked landscapes, optimal phenotypes can evolve rapidly by fixing one adaptive point mutation after the other. In contrast, such mutational trajectories can become trapped at sub-optima on multi-peaked fitness landscapes. Mutations conferring fitness losses are then essential for access to the global optimum, thus preventing evolution by fixing positively selected mutations one-by-one.

How adaptive constraints caused by multiple peaks can be overcome under specific selective regimes, and which requirements they impose on populations and evolutionary waiting times, has been the subject of many theoretical investigations. For example, it has been shown theoretically that fitness valleys may be crossed when populations expand, as this limits selection and allows maintenance of lower-fitness phenotypes [218,219]. For large populations and long waiting times, fitness valleys may be crossed by the joint fixation of multiple mutations [220], or through the recombination of different mutant alleles [70,221,222]. Other proposed non-adaptive mechanisms include drift [56,223] by which valleys can be traversed stochastically, partial penetrance [224] and non-heritable life time plasticity [225,226].

A recent computational study suggests that stasis at fitness plateaus can be overcome by temporal changes in the environment [227]. The mechanism exploits the notion that adaptive landscapes may be shaped differently in different environments, thus allowing the population to alternately climb different peaks in different environments and hence cross fitness valleys perceived in constant environments. We surmised that genetic constraints imposed by epistatic interactions, which may compose valleys in the adaptive landscapes, might similarly be overcome by environmental variations, as the relation between genotype and phenotype or fitness may equally be different in different environments. However, for this evolutionary mode to work, the fitness landscape must display highly specific changes in shape upon environmental change. It is unclear whether the fitness landscapes of actual biological systems exhibit the required properties and can support this evolutionary mode.

This question could be addressed by considering a phenotype that exhibits a multi-peaked adaptive landscape, and to determine its dependence on the environment. Adaptive landscapes have been mapped experimentally by systematically constructing neighboring genotypes that differ by one mutation, and

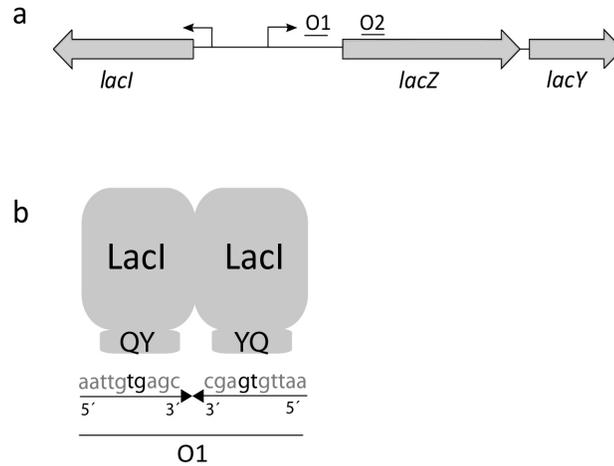
assaying their phenotype or fitness. These efforts have so far recovered single fitness peaks [60,61,71], and with exceptions [73] have focused on constant environments. Empirical observations of multi-peaked adaptive landscapes are scarce, although various studies have found indications for their existence [64,65,66,69,70,228,229].

Here we use the *lac* regulatory system in *Escherichia coli* as a model system to investigate the role of variable environments in overcoming genetic constraints in constant environments. We have recently argued on theoretical grounds that regulatory functions may exhibit multi-peaked landscapes [68,146], as the underlying lock-key molecular interactions give rise to reciprocal sign epistasis that is essential for the presence of multiple peaks. We systematically mapped the sequence space by constructing variants of LacI and the operator that are responsible for specific binding. These variants comprise substitutions in two amino acids in LacI and four base pairs in the operator. The ability of the LacI-operator variants to repress expression of the downstream gene products was assayed in the absence of the gratuitous inducer IPTG, while the ability to relieve repression was assayed in the presence of IPTG. A graphical method was developed to analyze the higher-order interactions between genotypes and the environment, and to assess the constraints in heterogeneous environments.

We find the landscape describing LacI-controlled expression to be highly rugged in the constant environments: all mutational trajectories starting from one sub-optimal variant, which is separated by six mutations from the optimal genotype, displayed transient decreases in the ability to repress before increasing to the optimum. Similarly, the trajectories in the environment with inducer all showed transient decreases in expression. The system displayed strong cross-environmental tradeoffs. Paths with limited decreases in repression without inducer showed strong decreases in expression with inducer, while *vice versa*, paths with limited expression decreases showed strong repression decreases. Counter-intuitively, the paths exhibiting the most mutations with tradeoffs were found to allow valley crossing. Along these particular paths, each mutation yielded functional improvements in at least one environment, even as functional deterioration was sustained in other environments. These results show that tradeoffs not only impose evolutionary constraint, but can also serve to accelerate evolution.

## 6.2 Genotype space for lock-key molecular recognition

In our system in *E. coli* the enzyme  $\beta$ -galactosidase, LacZ, and the *lac* permease, LacY, are co-regulated by LacI (Fig. 6.1a). In the presence of the inducer IPTG the downstream genes are expressed and in the absence of IPTG, the two dimers of the LacI tetramer suppress operon expression by binding to the two operators O1

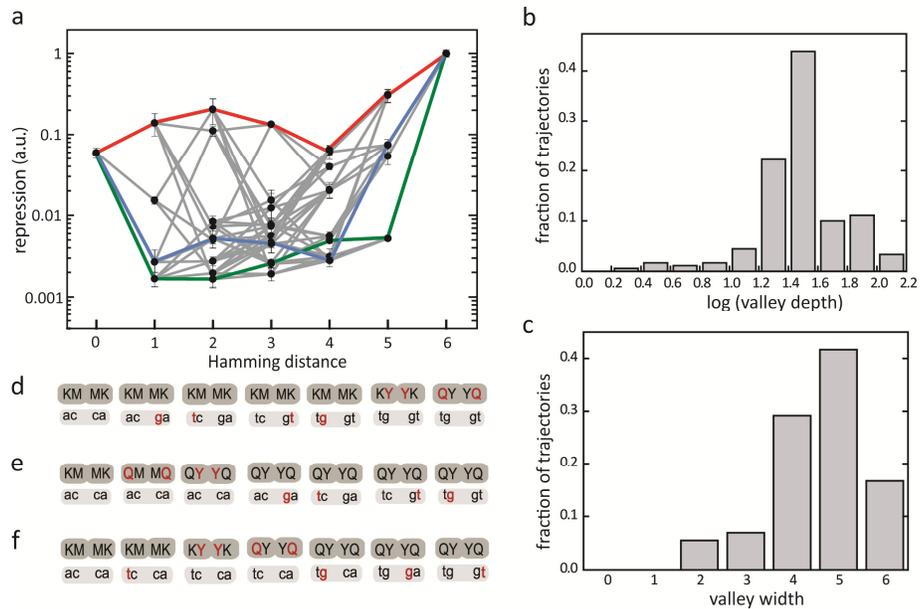


**Figure 6. 1 Schematic representation of the studied *E. coli lac* system.** a) Schematic depiction of the *lac* operon regulated by the *lac* repressor, LacI. The expression levels of  $\beta$ -galactosidase, LacZ and the *lac* permease, LacY, are co-regulated by the binding and unbinding of LacI to operator O1 and O2. b) Schematic depiction of the dimeric *lac* repressor binding to the palindromic operator. Indicated are the amino acids YQ (number 17 and 18) in the recognition helix in LacI and the base pairs tg-gt (45-5'4') in the two palindromic operator halves in O1, that are responsible for the specificity of binding of LacI to the operator.

and O2 [230,231]. O1 is located in the *lac* promoter region and the more weakly interacting O2 in the LacZ coding sequence (Fig. 6.1a). The operator O3 and the gene *lacA* in the natural *lac* system were found to be non-essential for the functioning of the *lac* operon [79,232] and have been omitted for simplicity.

A number of mutant *lac* repressors have previously been isolated that bind mutant operators specifically; the mutant repressor did bind the mutant operator but not the original operator. LacI binding specificity was found to be uniquely determined by two amino acid residues in the repressor (position 17 and 18) and four bases in the operator (5t and 4g, and 4g'and 5't in the two palindromic halves of O1 (Fig. 6.1b). Other repressor residues and bases in the regulatory region also affect binding and hence expression, but were not observed to play a role in specific binding [81,83,84,145,233,234,235]. Specific binding is a characteristic of lock-key interactions, and we therefore surmised that such specificity-determining loci could display reciprocal sign epistasis and multi-peaked adaptive landscapes [68].

We constructed several repressor and operator mutants with substitutions in the specificity-determining amino acids and base pairs (materials and methods), and assayed them on binding specificity by measuring expression of the downstream operon with a fluorogenic reporter that designates the beta-galactosidase



**Figure 6.2 The repression landscape of the *lac* system.**

a) Repression versus the Hamming distance. Repression is defined as  $1/\text{expression}$ , normalized to the YQ:tg-gt level. The Hamming distance denotes the number of substitutions relative to the *lac* repressor-operator pair MK:ac-ca. MK and YQ denote the amino acids in LacI and ac-ca and tg-gt the base pairs in the operator DNA. The trajectories from MK:ac-ca to YQ:tg-gt (grey lines) connect all intermediate *lac* repressor-operator variants (black points). The red trace connects the trajectory with the smallest decrease in repression. The green trace connects a trajectory in which first the operator is mutated and then the repressor. The blue trace connects a trajectory in which first the repressor is mutated and then the operator. Error bars are standard deviations,  $n=3$ . b) Histogram of the valley depth of all trajectories from MK:ac-ca to YQ:tg-gt. The valley depth is defined as the largest decrease in repression level along one mutational trajectory. c) Histogram of the valley width of all trajectories from MK:ac-ca to YQ:tg-gt. The valley width is defined as the number of mutations required to gain a repression level that is higher than the highest value before the largest decrease. d, e and f) Mutational trajectories of the green, blue and red traces in the repressor-operator repression landscape. Mutations in either repressor or operator are represented by red letters.

concentration in the cell (materials and methods). This procedure yielded a LacI variant carrying two mutations (Y17M and Q18K) that bound specifically and tightly to a *lac* operator variant carrying four mutations (t5a, g4c, g4'c and t5'a). The latter mutations are symmetric around the operator center such that the mutant and

original operators are palindromic. We denote these two LacI variants as YQ:tg-gt and MK:ac-ca, with e.g. YQ representing both repressor monomers that bind to the palindromic operator tg-gt.

The first question is whether the landscape describing the relation between LacI and the operator sequence and their ability to suppress expression has distinct peaks. More specifically, we aim to determine whether the isolated repressor-operator combination MK:ac-ca is separated from YQ:tg-gt by a valley.

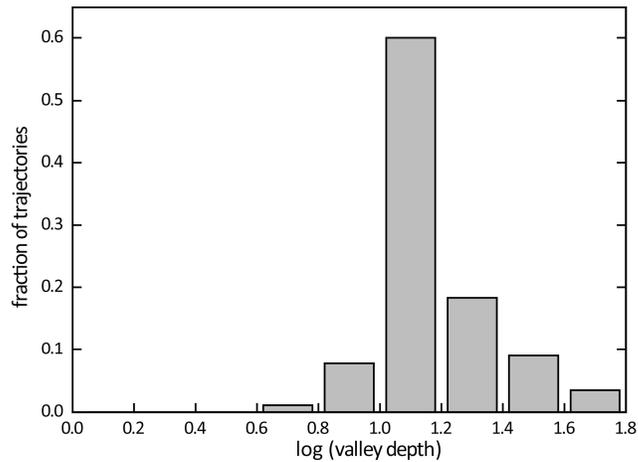
The sequence space of the direct trajectories between these two variants, which differ by six mutations, comprises  $2^6 = 64$  mutations. Because the repressors are homo-tetramers and thus symmetric, all non-palindromic operators are associated with an equivalent non-palindromic operator that is its mirror image.

With 12 of the in total  $2^4 = 16$  operators being non-palindromic, the total number of unique repressor-operator pairs in this sequence space is 40 ( $16 - (12/2) = 10$  operators times 4 repressors). We constructed these 40 variants harboring two mutations in the repressor and four in the operator, and first assayed their corresponding *lac* operon expression level in the absence of IPTG.

### 6.3 Multiple peaks in the *lac* repressor-operator recognition landscape

To visualize the relation between the genotype and the ability to repress, denoted repression, we plot one over the measured expression level without IPTG as a function of the Hamming distance from MK:ac-ca (Fig. 6.2a). The Hamming distance denotes the number of substitutions relative to MK:ac-ca. The repression is seen to vary over nearly three orders of magnitude, with YQ:tg-gt exhibiting the highest value and MK:ac-ca being about 15-fold lower. The plot shows that no path exists along which repression improves monotonically from MK:ac-ca to the optimum at YQ:tg-gt. Thus, for the repression level to improve from MK:ac-ca, it must first deteriorate before it reaches the optimal YQ:tg-gt.

To quantify the landscape ruggedness, we characterize the valley depth and width along all paths from MK:ac-ca to YQ:tg-gt. The valley depth is characterized as the largest total decrease in repression along one mutational trajectory. The corresponding histogram (Fig. 6.2b) shows that a large majority (95%) of all of the  $6! = 720$  trajectories exhibits a decrease of more than 10-fold. A small fraction ( $\sim 0.5\%$ ) displays the smallest observed decrease of a little more than two-fold. The valley width is taken as the number of mutations required to increase the repression to beyond that of the start of the largest decrease. We find that the smallest width is 2 mutations, which is observed for about 5% of the trajectories. A large majority of the trajectories (over 80%) has a valley width of 4 or more mutations.



**Figure 6.3 Histogram of the valley depth for the expression landscape.**

Histogram of the valley depth of all trajectories from MK:ac-ca to YQ:tg-gt. The valley depth is defined as the largest decrease in expression level along one mutational trajectory.

Overall, the data indicates that the two variants MK:ac-ca and YQ:tg-gt are separated by a valley, suggesting that the latter cannot be reached by fixing mutations one-by-one through positive selection.

#### 6.4 Genetic interactions underlying landscape ruggedness

The sequence of mutations along the trajectories can offer insight into the causes of the observed landscape ruggedness. For example, in the trajectory depicted in Fig. 6.2d, the four operator mutations occur first, which yields a decreasing repression level. Repression remains below that of the trajectory start at MK:ac-ca until the repressor is mutated to YQ (Fig. 6.2d and 6.2a, green trace). As a result, the valley along this path is both deep and wide. The same holds when the repressor is mutated first followed by mutations in the operator sequence (Fig. 6.2e and 6.2a, blue trace): repression is first abolished before it rises above the repression conferred by the starting sequence MK:ac-ca. The data shows that the repressor-operator interactions exhibit reciprocal sign epistasis: changing only the operator or only the repressor deteriorates repression, while changing both improves repression. Thus, these data are in line with the idea that the biological function of specific lock-key

interactions, as found in the *lac* repressor and operator, give rise to multiple peaks [146].

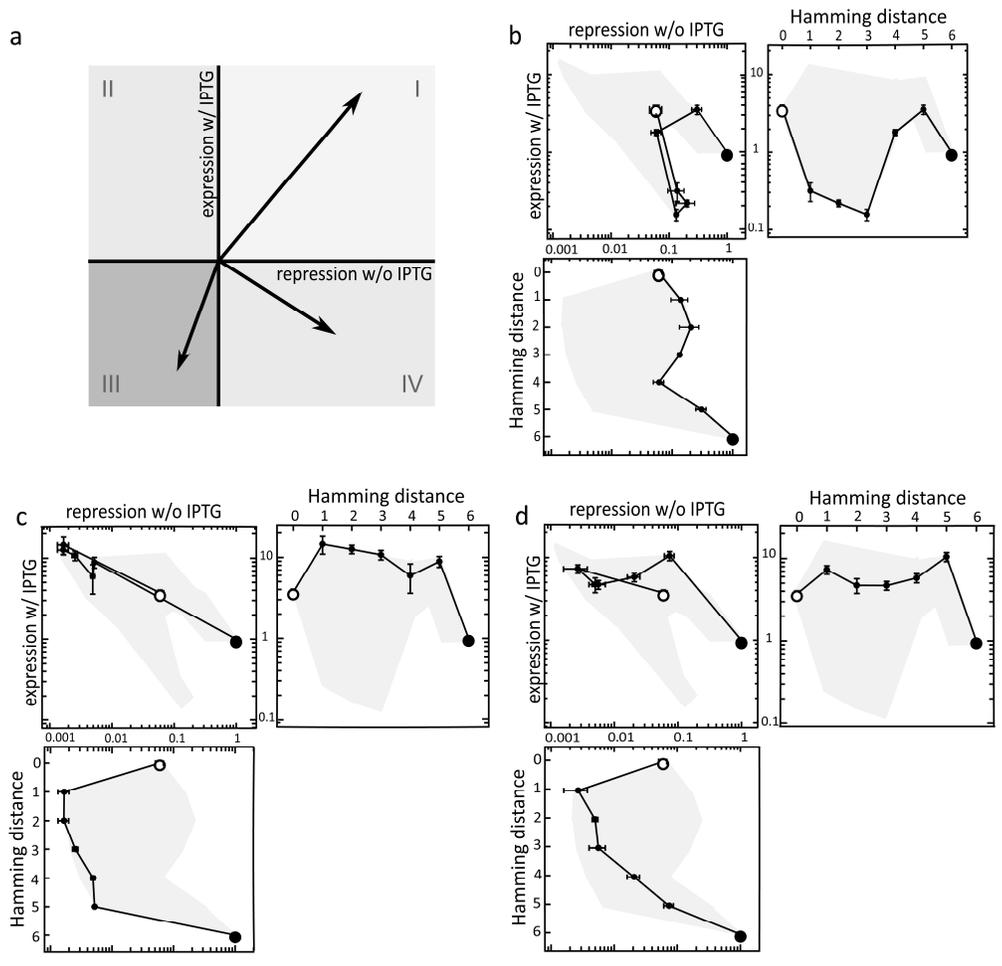
While the majority of trajectories are consistent with the above scenario, some rare cases are not. For instance, the mutational pathway depicted in Fig. 6.2f shows that operator tc-ca is recognized by both MK and YQ, as evidenced by the same repression of these two combinations ( $0.138 \pm 0.043$  and  $0.133 \pm 0.008$  respectively). This operator can be seen as a ‘master key’ that fits multiple locks. It has previously been argued that such master keys provide a mechanism to overcome the constraints associated with lock-key interactions [68] and allow the evolution of specific interactions by positive selection. However, the data shows that significant valleys remain regardless (Fig 6.2a, red trace, at a Hamming distance of 4). It appears that in a YQ:tc-ca background, two operator mutations (tc-ca to tg-ca, and tc-ca to tc-ga) individually decrease repression, but jointly increase repression (when forming tg-ga). Thus, reciprocal sign epistasis occurs here between parts that do not exhibit specific lock-key recognition.

### 6.5 Tradeoffs between expression and repression

The capability of the *lac* system to repress expression of the *lac* operon is advantageous in the absence of lactose, as it limits the costs of idly synthesizing the *lac* operon proteins [75,76,236]. Conversely, in the presence of lactose, the *lac* regulatory system allows the microbe to benefit from lactose metabolism by the *lac* enzymes.

To investigate how the mutations affect the latter objective we assayed LacZ expression in the presence of IPTG. We find this induced expression level to vary significantly repression without IPTG (about 2.5 orders). This large variation is notable, as one may expect mutations in the repressor-operator binding regions to affect predominantly the repressing state, and only marginally influence the expressing state. This expression landscape is also rugged: all paths to the optimum at YQ:tg-gt exhibit at least three-fold expression decreases (Fig. 6.3). Thus, YQ:tg-gt cannot be reached by fixing mutations one-by-one through positive selection in this environment. These data raise the question whether the changes in repression and expression are predominantly independent or correlated.

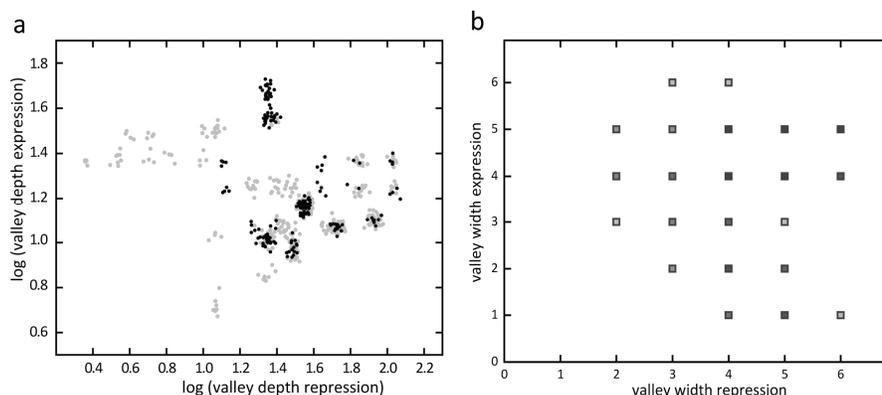
To analyze the relation between repression and expression in the *lac* system, we introduce a graphical method (Fig. 6.4a). In this method, mutations are represented as vectors in a two-dimensional coordinate system, where the axes indicate the repression in the environment without IPTG, and expression in the presence of IPTG. Trajectories from MK:ac-ca to the optimum at YQ:tg-gt display only



**Figure 6.4 Mutational trajectories in two environments.** Caption on next page.

**Figure 6. 4 Mutational trajectories in two environments.** a) Schematic visualization of the possible effects of mutations on phenotype in two environments. Mutations are represented as vectors with their start in the origin of the coordinate system. Mutations that improve the phenotype, repression or expression, in both environments, without IPTG or with IPTG, are located in quadrant I. Mutations that improve the phenotype in one environment and deteriorate the phenotype in the other environment, are located in quadrant II or IV. Mutations that are detrimental in both environments are located in quadrant III. Mutations that improve the phenotype, repression or expression, in one environment and simultaneously deteriorate the phenotype in the other environment, generate tradeoffs. These mutations are located in quadrant II or IV. Trajectories whose vectors lie exclusively in either I, II or IV allow an adaptive walk from MK:ac-ca to YQ:tg-gt, provided suitably varying environments. b, c, d) Single mutational trajectory from MK:ac-ca (white circle) to YQ:tg-gt (black circle) harbouring six mutations, represented as: repression without IPTG versus expression with IPTG (upper left); Hamming distance, relative to MK:ac-ca, versus expression with IPTG (upper right); and repression without IPTG versus Hamming distance, relative to MK:ac-ca.

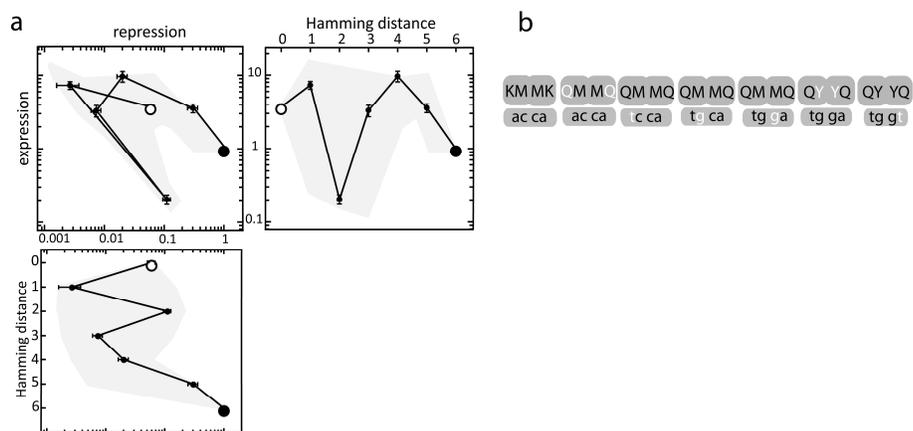
The grey surfaces outline all possible trajectories, either for both environments (upper left), for only expression with IPTG (upper right), or only for repression without IPTG (below). Error bars are standard deviations,  $n=3$ . b) The trajectory with the smallest decrease in repression. This particular trajectory, which corresponds to Fig. 6.2f is not selectively accessible due to a mutational step that deteriorates both repression and expression. c) A trajectory with the largest decrease in repression. In this trajectory, which corresponds to Fig. 6.2d, large decreases in repression are correlated with the maintenance of high expression levels. This trajectory is not selectively accessible in variable environments, since it contains neutral steps. d) A trajectory with a large decrease in repression. In this trajectory, which corresponds to Fig. 6.2e, a relatively large decrease in repression is balanced with relatively small decreases in expression. This trajectory is not selectively accessible in variable environments due to the presence of neutral mutations in the trajectory.



**Figure 6.5 Valley depth and valley width in both environments.** a) Valley depth repression without IPTG versus valley depth expression with IPTG. Grey points represent individual trajectories. Data points are offset by a random number between 0 and 0,1 for clarity. Black points are the subset of trajectories that are accessible by positive selection in variable environments. Trajectories that show a small valley depth in both repression and expression are absent. The selectively accessible trajectories are mainly located in the upper right corner. b) Valley width repression without IPTG versus valley width expression with IPTG. Individual trajectories are marked with a square, with the intensity as a measure of the occurrence. Trajectories that have a small valley width in both environments are absent.

few mutational steps that yield improvements in both repression and expression ability, and thus lie in quadrant I (around 10%). For the majority of mutational steps, improvement of one capability comes at the expense of deterioration of the other (over 75%). Mutations displaying such cross-environment tradeoffs have vectors that point either in quadrant II or IV.

Next, we considered mutational trajectories from MK:ac-ca to YQ:tg-gt. The paths with deep and wide valleys in repression without IPTG, displayed increases in expression with IPTG that decreased in the vicinity of YQ:tg-gt (Fig. 6.4c and d). Conversely, the trajectories displaying limited decreases in repression showed a strongly decreased expression with IPTG (Fig. 6.4 b). This negative correlation was generic: no paths were found with small valley depths in both repression and expression (Fig. 6.5a). These data indicate large tradeoffs between repression and expression ability.



**Figure 6.6 Tradeoff trajectory allows a walk on multi-peaked landscape.**

a) Mutational trajectory from MK:ac-ca (white circle) to YQ:tg-gt (black circle). Repression without IPTG versus expression with IPTG (upper left); Hamming distance, relative to MK:ac-ca, versus expression with IPTG (upper right); and repression without IPTG versus Hamming distance, relative to MK:ac-ca. The surfaces outline all possible trajectories in grey, either for both environments (upper left), for only expression with IPTG (upper right), or only for repression without IPTG (below). Error bars are standard deviations,  $n=3$ . This trajectory contains tradeoffs, in which all mutations are positively selected in at least one environment. b) The mutational representation of the trajectory with the mutational changes in white.

## 6.6 Valley-crossing in heterogeneous environments

Specific changes in the adaptive landscape upon environmental changes can offer routes to cross valleys that are not present in constant environments [227]. The minimal condition for such a scenario is that all mutations must yield functional improvements in at least one environment. This is not the case in the trajectories displayed in Fig. 6.4, as they contain mutations that are either detrimental or neutral in both environments. Indeed, the majority of all trajectories (62%) contain such mutations. However, a significant fraction of the paths (38%) contains exclusively mutations that yield improvements, either in repression in the absence of inducer or expression, in the presence of inducer.

An example of such a trajectory is depicted in Fig. 6.6. Strikingly, this pathway traverses the most rugged regions of the separate repression and expression landscapes, as evidenced by the large and frequent excursions in repression and expression. However, each decrease in repression in the environment without IPTG coincides with a matching increase in expression in the environment with IPTG, while

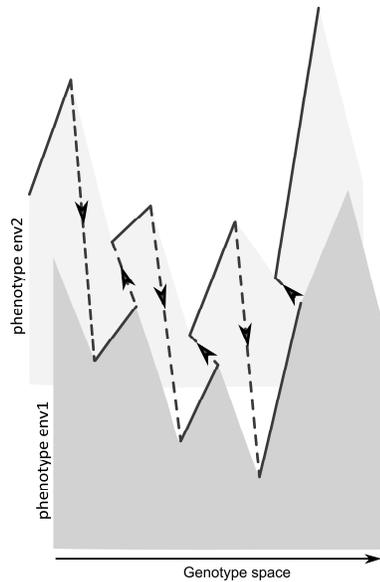
conversely, decreases in expression coincide with repression increases. Correspondingly, the trajectory is seen to make large detours in repression-expression space (Fig. 6.6 a), with the functional losses incurred by one mutation in one environment being compensated by different mutations in the other environment. We find that all these accessible paths localize in the upper-right corner of Fig. 6.5a and b, which is consistent with large variations in repression and expression along the trajectories. Overall, the data indicates that cross-environment tradeoffs are central to crossing the valley by positive selection.

## 6.7 Discussion

By systematically reconstructing evolutionary intermediates between a sub-optimal *lac* repressor-operator variant and the wild-type *lac* repressor and operator, we have shown that the landscapes describing repression in the absence of IPTG and expression in the presence of IPTG exhibit multiple peaks that are separated by valleys. Reciprocal sign epistasis is found to be associated with the ruggedness in the landscape in fixed environments. Negative, higher-order interactions between multiple genotypes and the environment are responsible for tradeoffs in the adaptive trajectories between environments.

Interestingly, cross-environmental tradeoffs, are typically discussed for their ability to constrain evolution in variable environments [158,206,237,238,239,240,241,242,243,244,245]. Here, they are found to play an important role in the crossing of valleys on a rugged landscape, such that stasis on sub-optimal peaks can be overcome. In the trajectories that could potentially cross the multi-peaked landscapes, each mutation yielded functional improvements in at least one environment, and adaptive mutations can be fixed one-by-one in a suitably fluctuating environment. This can be visualized by a trajectory zig-zagging over the multi-peaked landscapes in the separate environments (Fig. 6.7).

The probability of fixation of mutations in the proposed zig-zag trajectories in temporal fluctuating environments is dependent on the fluctuation frequency and dwelling time in the environment, as well as the selective pressure acting on mutations. It is beyond the scope of this current chapter to define the exact regimes in which this valley crossing is allowed, but this will be of interest to explore. Overall, we have assumed a simple scenario, under which the strong selection-weak mutation assumption holds [63], and the fluctuation frequency and the corresponding dwelling time in each environment is sufficiently long to provide for the (near) fixation of mutations.



**Figure 6. 7 Schematic of valley crossing mechanism in variable environments.** The mutational trajectory can cross valleys in both environments by adaptively fixing single mutations one-by-one by alternating between the environments, e.g. alternately fixing one beneficial mutation in one environment (e.g. env2), followed by the fixation of the next mutation in the other environment (e.g. env1), and so forth, until the global optimum is reached.

The strong dependence of the adaptive accessibility of these trajectories on the environments seems constraining, maybe even unlikely at first sight, since mutations need to coincide with the favorable environment to allow (near) fixation in the population. However, since there are many trajectories, 273 of 720, that contain mutations that increase function with each mutational step in at least one environment, and since all these trajectories show per mutational step different increases in performance in different environments the crossing of valleys may occur in a relatively large range of alternating environments.

Concluding, we find that the interplay between multi-peaked fitness landscapes in constant environments and tradeoffs in variable environments can promote adaptive evolution by the step-by-step fixation of positively selected mutations. This result suggests an even larger role to both ecology and life-history for the outcome of evolutionary processes as commonly assumed.

## 6.8 Materials and Methods

### 6.8.1 Strains

We used strain *Escherichia coli* strain BW2373 F<sup>-</sup>,  $\Delta(\text{argF-lac})169$ ,  $\Delta\text{uidA3::pir}^+$ , *recA1*, *rpoS396(Am)*, *endA9(del-ins)::FRT*, *rph-1*, *hsdR514*, *rob-1*, *creC510* [246] (Yale Coli Genetic Stock Collection #7837) for all experiments.

### 6.8.2 Plasmids

We used a single copy plasmid that can be induced to a multi-copy state for cloning purposes (pETcoco-2, Novagen), to mimic natural expression levels. We deleted *lacI* from the pETcoco-2 by SacI-NheI, and replaced it with the T0 and Trev terminator from pinv-110 [247] SacI-SpeI in SacI-NheI, by which the NheI site disappears.

*placI* is derived from an altered *placI-q* sequence from pTrc99A [181], in which the high promoter level was down-tuned to the wild type level by site directed mutagenesis to the wild type *placI* promoter. It is combined with *lacI* from pRD007 (without the O3-sequence) [191] and a spacer sequence derived from *Drosophila* Kinesin (pCascade5) by overlap pcr and inserted in pETcoco-2 by a ClaI-HindIII restriction and insertion.

The *plac* and *lacZ* and *lacY* wild type sequences are derived from *E. coli* strain Mg1655 (Yale Coli Genetic Stock Collection #6300). *plac* was amplified until the 9<sup>th</sup> codon in the sequence in *lacZ*, the amino acid was degenerated to produce the BmtI restriction site at base pair level, and was inserted HindIII-BmtI/NheI in the vector backbone. The rest of the *lacZ* sequence and *lacY* were amplified and inserted in this restriction site together with AvrII in the end of *lacY* and the polylinker sequence of Petcoco-2 to make pMdv53.

### 6.8.3 Mutants

Mutants of the *lacI* and *plac* wt sequence were constructed by overlap extension pcr, in which *lacI* mutants and *plac* mutants were constructed separately. Combinations of *lacI-plac* mutants were constructed by overlap pcr, and inserted in pMdv53 by KasI-NheI restriction and ligation.

#### **6.8.4 Media**

All growth and expression measurements were performed in Defined Rich medium (Teknova, Hollister, CA, USA), with 0.2% glucose as carbon source, and supplemented with 1mM thiamine HCl and uracil.

#### **6.8.5 Measurement of expression level**

Cultures were grown at 37°C in a Perkin & Elmer Victor3 plate reader, at 200 ml 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, uracil and the appropriate antibiotics. Optical density at 600 nm was recorded every 4 min, and every 29 min 9 sterile water is 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. Cells were fixed after the cultures had reached an optical density of at least 0.015 and at most 0.07, by adding 20 ml FDG-fixation solution (109 mM fluorescein di-b-D-galactopyranoside (FDG, Enzo Life sciences, NL), 0.15% formaldehyde, and 0.04% DMSO in MilliQH2O). Fluorescence development was measured every 8 min (exc. 480 nm, em. 535 nm), as well as the OD<sub>600</sub>. Shaking and dispensing conditions as above. When cells are induced with IPTG, directly before or after fixation, an appropriate amount of inhibitive IPTG was added. Analysis of the fluorescence trace is as described in chapter 4 and [191].

## Bibliography

1. Ridley M (2004) *Evolution*: Blackwell Science Ltd.
2. Chernikova D, Motamedi S, Csuros M, Koonin EV, Rogozin IB A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol Direct* 6: 26.
3. Saltzman MR, Young SA, Kump LR, Gill BC, Lyons TW, et al. Pulse of atmospheric oxygen during the late Cambrian. *Proc Natl Acad Sci U S A* 108: 3876-3881.
4. Canfield DE, Poulton SW, Narbonne GM (2007) Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. *Science* 315: 92-95.
5. Narbonne GM, Gehling JG (2003) Life after Snowball: the oldest complex Ediacaran fossils. *Geology* 31: 27-30.
6. Bowring SM, Myrow P, Landing E, Ramezani J (2002) Geochronological constraints on terminal Neoproterozoic events and the rise of Metazoans *Astrobiology* 2.
7. Gould SJ (1989) *Wonderful Life: The Burgess Shale and the Nature of History*. New York: W.W. Norton & Company.
8. Lofgren AS, Plotnick RE, Wagner PJ (2003) Morphological diversity of Carboniferous arthropods and insights on disparity patterns through the phanerozoic. *Paleobiology* 29: 349.
9. Nielsen C, Parker A Morphological novelties detonated the Ediacaran-Cambrian "explosion". *Evol Dev* 12: 345-346.
10. Peterson KJ, Cotton JA, Gehling JG, Pisani D (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc Lond B Biol Sci* 363: 1435-1443.
11. Redecker D, Kodner R, Graham LE (2000) Glomalean fungi from the Ordovician. *Science* 289: 1920-1921.
12. Fortey RA, Briggs DEG, Wills MA (1996) The Cambrian evolutionary 'explosion': decoupling cladogenesis from morphological disparity. *Biol J Linn Soc* 57: 13-33.
13. Marshall CR (2006) Explaining the Cambrian explosion of "animals". *Annu Rev Earth Planet Sci* 34: 355-384.
14. Zhuravlev A, Riding R (2001) *The Ecology of the Cambrian Radiation*. New York: Columbia Univ. Press. In: A.Y. Z, Riding R, editors. New York: Columbia Univ. Press. pp. 525.
15. Bergstrom J, Hou X-G (2005) Early palaeozoic non-lamellipeian arthropods; Koenemann S, Jenner RA, editors. Oxford: Taylor & Francis.
16. Powell JACIR, Vincent H.; Cardé, Ring T.. *Encyclopedia of Insects* (2 (illustrated) ed.). Academic Press. (2009) Coleoptera. In: Resh VH, Cardé RT, editors. *Encyclopedia of Insects* (2 (illustrated) ed: Elsevier/Academic Press. pp. 1132.
17. Williams BA, Kay RF, Kirk EC New perspectives on arthropod origins. *Proc Natl Acad Sci U S A* 107: 4797-4804.
18. Clarke JA, Tambussi CP, Noriega JI, Erickson GM, Ketchum RA (2005) Definitive fossil evidence for the extant avian radiation in the Cretaceous. *Nature* 403: 305-308.

19. Brunet M, Guy F, Pilbeam D, Lieberman DE, Likius A, et al. (2005) New material of the earliest hominid from the Upper Miocene of Chad. *Nature* 434: 752-755.
20. Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, et al. (2002) A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* 418: 145-151.
21. Leakey R, editor (1994) *The origin of humankind*. 6 ed. London: Weidenfeld & Nicholson.
22. Cuvier G (1800) *Mémoires sur les espèces d'éléphants vivants et fossiles*.
23. Darwin E (1796) *Zoonomia, Vol. I Or, the Laws of Organic Life*. LONDON: ST. PAUL'S CHURCH-YARD.
24. Pearson PN (2003) In retrospect. *Nature* 425.
25. Hutton J (1794) *An Investigation of the Principles of Knowledge and of the Progress of Reason, from Sense to Science and Philosophy*.
26. Lamarck JB (1809) *Philosophie Zoologique. Philosophie Zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux*. Paris.
27. Darwin C (1859) *The origin of species by means of natural selection or The preservation of favoured races in the struggle for life*. London.
28. Mendel G (1866) *Experiments in Plant Hybridization*.
29. Huxley J (1942) *Evolution: The modern synthesis*. London: Allen & Unwin.
30. Fisher RA (1930) *The genetical theory of natural selection*: Clarendon press.
31. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb* 52.
32. Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16: 97-159.
33. Wright S (1932 ) The roles of mutation, inbreeding, crossbreeding and selection in evolution. . *Proceedings of the Sixth International Congress of Genetics* 1: 356-366.
34. Dobzhansky T (1937) *Genetics and the Origin of Species*. New York: Columbia University Press.
35. Ayala FJ, Fitch WM (1997) Genetics and the origin of species: an introduction. *Proc Natl Acad Sci U S A* 94: 7691-7697.
36. Mayr E (1942) *Systematics and the Origin of Species*. New York: Columbia University.
37. Wallace AR (1895) The method of organic selection. *Fortnightly Rev* 57: 435-445.
38. Simpson GG (1944) *Tempo and Mode in Evolution*. . New York: : Columbia Univ. Press.
39. Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* 36: 39-56.
40. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *nature* 171: 737-738.
41. Johannsen W, editor (1909) *Elemente der exakten Erblchkeitslehre* 2nd ed., 1913; 3rd ed., 1926 ed. Jena: Gustav Fischer.
42. Churchill FB (1974. ) William Johannsen and the genotype concept. *Journal of the History of Biology* 7 5-30.
43. Beadle GW, Tatum EL (1941 ) The genetic control of biochemical reactions in *Neurospora* *Proceedings of the National Academy of Science* 27: 499-506.

44. Bradshaw AD (1965) Evolutionary significance of phenotypic plasticity in plants  
Advances in Genetics 13: 115-155.
45. Via S (1993) Adaptive Phenotypic Plasticity: Target or By-Product of Selection in a  
Variable Environment? The American Naturalist 142: 352-365.
46. Scheiner SM, Lyman RF (1989 ) The genetics of phenotypic plasticity I.  
Heritability. Journal of Evolutionary Biology 2: 95-107.
47. Scheiner C, Lyman RF (1991) The genetics of phenotypic plasticity II. Response to  
selection. . Journal of Evolutionary Biology 3,23-50.
48. Scheiner SM, Caplan RL, Lyman RF (1991) The genetics of phenotypic plasticity III.  
Genetic correlations and fluctuating asymmetries. . Journal of Evolutionary  
Biology 4: 51-68.
49. Dawkins R (1976) The selfish gene. Oxford: Oxford press.
50. Conway-Morris S, Gould SJ (1998) Showdown on the Burgess Shale Natural  
History magazine 107: 48-55.
51. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution:  
the functional synthesis. Nat Rev Genet 8: 675-688.
52. Yokoyama S, Zhang, H., Radlwimmer, F. B. & Blow, N. S. (1999) Adaptive evolution  
of color vision of the Comoran coelacanth (*Latimeria chalumnae*). . Proc Natl  
Acad Sci USA 96: 6279–6284.
53. Yokoyama S (2000) Color vision of the coelacanth (*Latimeria chalumnae*) and  
adaptive evolution of rhodopsin (RH1) and rhodopsin-like (RH2) pigments. . J  
Hered 91: 215–220.
54. Yokoyama S, Tada T (2000) Adaptive evolution of the African and Indonesian  
coelacanths to deep-sea environments. . Gene 261: 35–42.
55. Whitlock MC, Phillips PC, Moore FB-G, Tonsor SJ (1995) Multiple fitness peaks and  
epistasis. Annu Rev Ecol Syst 26: 601–629.
56. Kimura M (1983) The Neutral Theory of Molecular Evolution.; Kingman JFC, editor.  
Cambridge: Cambridge University Press.
57. Wright S (1982) The shifting balance theory and macroevolution. Annu Rev Genet  
16: 1-19.
58. Fisher RA (1950) The "Sewell Wright Effect". Heredity 4: 117-119.
59. Gavrillets S (2004) Fitness landscapes and the origin of species. Princeton and  
Oxford: Princeton University press.
60. Carneiro M, Hartl DL Colloquium papers: Adaptive landscapes and protein  
evolution. Proc Natl Acad Sci U S A 107 Suppl 1: 1747-1751.
61. Lunzer M, Miller SP, Felsheim R, Dean AM (2005) The biochemical architecture of  
an ancient adaptive landscape. Science 310: 499-501.
62. Weinreich DM (2005) The rank ordering of genotypic fitness values predicts  
genetic constraint on natural selection on landscapes lacking sign epistasis.  
Genetics 171: 1397-1405.
63. Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic  
constraint on evolutionary trajectories. Evolution Int J Org Evolution 59:  
1165-1174.

64. Kivitek DJ, Sherlock G (2011) Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. *PLoS Genet* 7: e1002056.
65. 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 U S A* 91: 9037-9041.
66. Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, et al. (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc Natl Acad Sci U S A* 106: 12025-12030.
67. 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.
68. Dawid A, Kiviet DJ, Kogenaru M, de Vos M, Tans SJ (2010) Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos* 20: 026105.
69. Burch CL, Chao L (1999) Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics* 151: 921-927.
70. de Visser JA, Park SC, Krug J (2009) Exploring the effect of sex on empirical fitness landscapes. *Am Nat* 174: S15-S30.
71. 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.
72. Poelwijk FJ, Tanase-Nicola S, Kiviet DJ, Tans SJ Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* 272: 141-144.
73. Longzhi T, Stephen Serene S, Chao HX, Gore J (2011) Hidden randomness between fitness landscapes limits reverse evolution. *Phys Rev Lett* 106: 198102.
74. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3: 318-356.
75. Stoebel DM, Dean AM, Dykhuizen DE (2008) The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. *Genetics* 178: 1653-1660.
76. Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436: 588-592.
77. Mossing MC, Record MT, Jr. (1986) Upstream operators enhance repression of the lac promoter. *Science* 233: 889-892.
78. Eismann ER, Muller-Hill B (1990) lac repressor forms stable loops in vitro with supercoiled wild-type lac DNA containing all three natural lac operators. *J Mol Biol* 213: 763-775.
79. Oehler S, Eismann ER, Kramer H, Muller-Hill B (1990) The three operators of the lac operon cooperate in repression. *Embo J* 9: 973-979.
80. Boelens R, Lamerichs RM, Rullmann JA, van Boom JH, Kaptein R (1988) The interaction of lac repressor headpiece with its operator: an NMR view. *Protein Seq Data Anal* 1: 487-498.

81. Kisters-Woike B, Lehming N, Sartorius J, von Wilcken-Bergmann B, Muller-Hill B (1991) A model of the lac repressor-operator complex based on physical and genetic data. *Eur J Biochem* 198: 411-419.
82. Lehming N (1990) Regeln für Protein/DNA-Erkennung. PhD thesis.
83. Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, Muller-Hill B (1990) Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. *Embo J* 9: 615-621.
84. Lehming N, Sartorius J, Niemoller M, Genenger G, v Wilcken-Bergmann B, et al. (1987) The interaction of the recognition helix of lac repressor with lac operator. *Embo J* 6: 3145-3153.
85. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update. *Nucleic Acids Res* 32: D23-26.
86. Otto SP (1997) Unravelling gene interactions. *Nature* 390: 343.
87. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science* 277: 1453-1462.
88. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, et al. (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 123: 507-519.
89. Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 7: R63.
90. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446: 806-810.
91. Zhu G, Golding GB, Dean AM (2005) The selective cause of an ancient adaptation. *Science* 307: 1279-1282.
92. Miller SP, Lunzer M, Dean AM (2006) Direct demonstration of an adaptive constraint. *Science* 314: 458-461.
93. Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128: 11-45.
94. Arnold SJ, Pfrender ME, Jones AG (2001) The adaptive landscape as a conceptual bridge between micro- and macroevolution. *Genetica* 112-113: 9-32.
95. Bateson W (1907) Facts Limiting the Theory of Heredity. *Science* 26: 649-660.
96. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 445: 383-386.
97. Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution* 55: 1085-1094.
98. Kondrashov AS (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature* 336: 435-440.
99. Kondrashov FA, Kondrashov AS (2001) Multidimensional epistasis and the disadvantage of sex. *Proc Natl Acad Sci U S A* 98: 12089-12092.
100. Kouyos RD, Otto SP, Bonhoeffer S (2006) Effect of varying epistasis on the evolution of recombination. *Genetics* 173: 589-597.
101. de Visser JA, Elena SF (2007) The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat Rev Genet* 8: 139-149.

102. Gandon S, Otto SP (2007) The evolution of sex and recombination in response to abiotic or coevolutionary fluctuations in epistasis. *Genetics* 175: 1835-1853.
103. Masel J (2005) Evolutionary capacitance may be favored by natural selection. *Genetics* 170: 1359-1371.
104. Martin G, Elena SF, Lenormand T (2007) Distributions of epistasis in microbes fit predictions from a fitness landscape model. *Nat Genet* 39: 555-560.
105. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6: 678-687.
106. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* 103: 5869-5874.
107. Francino MP (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* 37: 573-577.
108. Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res* 13: 2052-2058.
109. Lynch M (2005) Simple evolutionary pathways to complex proteins. *Protein Sci* 14: 2217-2225; discussion 2226-2217.
110. Elena SF, Lenski RE (1997) Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390: 395-398.
111. Burch CL, Turner PE, Hanley KA (2003) Patterns of epistasis in RNA viruses: a review of the evidence from vaccine design. *J Evol Biol* 16: 1223-1235.
112. Elena SF (1999) Little evidence for synergism among deleterious mutations in a nonsegmented RNA virus. *J Mol Evol* 49: 703-707.
113. Wloch DM, Szafraniec K, Borts RH, Korona R (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. *Genetics* 159: 441-452.
114. De Visser JA, Hoekstra RF, van den Ende H (1997) Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. *Evolution* 51: 1499-1505.
115. Sanjuan R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A* 101: 8396-8401.
116. Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell* 20: 1199-1216.
117. Lenski RE, Ofria C, Collier TC, Adami C (1999) Genome complexity, robustness and genetic interactions in digital organisms. *Nature* 400: 661-664.
118. You L, Yin J (2002) Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage t7. *Genetics* 160: 1273-1281.
119. Tischler J, Lehner B, Fraser AG (2008) Evolutionary plasticity of genetic interaction networks. *Nat Genet* 40: 390-391.
120. Cooper TF, Remold SK, Lenski RE, Schneider D (2008) Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*. *PLoS Genet* 4: e35.

121. Cooper TF, Rozen DE, Lenski RE (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci U S A* 100: 1072-1077.
122. Gosset G, Zhang Z, Nayyar S, Cuevas WA, Saier MH, Jr. (2004) Transcriptome analysis of Crp-dependent catabolite control of gene expression in *Escherichia coli*. *J Bacteriol* 186: 3516-3524.
123. Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, et al. (2004) Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. *Genetics* 167: 1721-1737.
124. Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL (2006) Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* 440: 87-90.
125. Moore JH (2005) A global view of epistasis. *Nat Genet* 37: 13-14.
126. Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* 37: 77-83.
127. Remold SK, Lenski RE (2001) Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *Proc Natl Acad Sci U S A* 98: 11388-11393.
128. Kauffman SA, Johnsen S (1991) Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *J Theor Biol* 149: 467-505.
129. Kauffman SA, Weinberger ED (1989) The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J Theor Biol* 141: 211-245.
130. Kim Y (2007) Rate of adaptive peak shifts with partial genetic robustness. *Evolution* 61: 1847-1856.
131. Buckling A, Wills MA, Colegrave N (2003) Adaptation limits diversification of experimental bacterial populations. *Science* 302: 2107-2109.
132. Whitlock MC, Phillips PC, Moore FBG, Tonsor SJ (1995 ) Multiple Fitness Peaks and Epistasis. *Annu Rev Ecol Syst* 26: 601-629
133. Bell CE, Lewis M (2001) The Lac repressor: a second generation of structural and functional studies. *Curr Opin Struct Biol* 11: 19-25.
134. Adler K, Beyreuther K, Fanning E, Geisler N, Gronenborn B, et al. (1972) How lac repressor binds to DNA. *Nature* 237: 322-327.
135. Barker A, Fickert R, Oehler S, Muller-hill B (1998) Operator search by mutant Lac repressors. *J Mol Biol* 278: 549-558.
136. Calos MP, Galas D, Miller JH (1978) Genetic studies of the lac repressor. VIII. DNA sequence change resulting from an intragenic duplication. *J Mol Biol* 126: 865-869.
137. Coulondre C, Miller JH (1977) Genetic studies of the lac repressor. III. Additional correlation of mutational sites with specific amino acid residues. *J Mol Biol* 117: 525-567.
138. Fickert R, Muller-Hill B (1992) How Lac repressor finds lac operator in vitro. *J Mol Biol* 226: 59-68.

139. Miller JH, Ganem D, Lu P, Schmitz A (1977) Genetic studies of the lac repressor. I. Correlation of mutational sites with specific amino acid residues: construction of a colinear gene-protein map. *J Mol Biol* 109: 275-298.
140. Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, et al. (1996) Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol Biol* 261: 509-523.
141. Weickert MJ, Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem* 267: 15869-15874.
142. Swint-Kruse L, Matthews KS (2009) Allosterity in the LacI/GalR family: variations on a theme. *Curr Opin Microbiol* 12: 129-137.
143. Sadler JR, Sasmor H, Betz JL (1983) A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc Natl Acad Sci U S A* 80: 6785-6789.
144. Miller JH (1972) *Experiments in Molecular Genetics*. New York: Cold Spring Harbor Laboratory.
145. Hollis M, Valenzuela D, Pioli D, Wharton R, Ptashne M (1988) A repressor heterodimer binds to a chimeric operator. *Proc Natl Acad Sci U S A* 85: 5834-5838.
146. Poelwijk FJ, Tanase-Nicola S, Kiviet DJ, Tans SJ (2011) Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* 272: 141-144.
147. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315: 525-528.
148. Nguyen CC, Saier MH, Jr. (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett* 377: 98-102.
149. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461: 515-519.
150. Agrawal AA (2001) Phenotypic plasticity in the interactions and evolution of species. *Science* 294: 321-326.
151. Pigliucci M (2001) *Phenotypic plasticity. Beyond nature and nurture.*; Scheiner SM, editor. Baltimore and London: The Johns Hopkins University Press.
152. DeWitt TJ, Scheiner SM (2004) *Phenotypic plasticity. Functional and conceptual approaches*. Oxford: Oxford University Press.
153. Scheiner SM (1993) Genetics and Evolution of phenotypic plasticity. *Annu Rev Ecol Syst* 24: 35-68.
154. Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, et al. (1985) Developmental constraints and evolution: a perspective from the MountainLake conference on development and evolution. *QRevBiol* 60: 265-287.
155. van Tienderen PH (1997) Generalists, specialists, and the evolution phenotypic plasticity in sympatric populations of distinct species. *Evolution* 51: 1372-1380.

156. Via S, Gomulkiewicz R, De Jong G, Scheiner SM, Schlichting CD, et al. (1995) Adaptive phenotypic plasticity: consensus and controversy. *Trends Ecol Evol* 10: 212-217.
157. Scheiner SM (2002) Selection experiments and the study of phenotypic plasticity. *J Evol Biol* 15: 889-898.
158. Suiter AM, Banziger O, Dean AM (2003) Fitness consequences of a regulatory polymorphism in a seasonal environment. *Proc Natl Acad Sci U S A* 100: 12782-12786.
159. Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* 6: 533-543.
160. 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.
161. Roff DA (2001) Life history evolution. Sunderland, MA: Sinauer Associates.
162. Choi KY, Zalkin H (1992) Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *J Bacteriol* 174: 6207-6214.
163. Flynn TC, Swint-Kruse L, Kong Y, Booth C, Matthews KS, et al. (2003) Allosteric transition pathways in the lactose repressor protein core domains: asymmetric motions in a homodimer. *Protein Sci* 12: 2523-2541.
164. Zhan H, Camargo M, Matthews KS Positions 94-98 of the lactose repressor N-subdomain monomer-monomer interface are critical for allosteric communication. *Biochemistry* 49: 8636-8645.
165. Pfahl M (1976) lac Repressor-operator interaction. Analysis of the X86 repressor mutant. *J Mol Biol* 106: 857-869.
166. Miller JH, Schmeissner U (1979) Genetic studies of the lac repressor. X. Analysis of missense mutations in the lacI gene. *J Mol Biol* 131: 223-248.
167. Chamness GC, Willson CD (1970) An unusual lac repressor mutant. *J Mol Biol* 53: 561-565.
168. Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, et al. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452: 840-845.
169. True JR, Carroll SB (2002) Gene co-option in physiological and morphological evolution. *Annu Rev Cell Dev Biol* 18: 53-80.
170. Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312: 97-101.
171. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61: 995-1016.
172. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300: 1701-1703.
173. 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.
174. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE (2006) Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci U S A* 103: 9107-9112.

175. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, et al. (2009) Adaptive prediction of environmental changes by microorganisms. *Nature* 460: 220-224.
176. Vasi F, Travisano M, Lenski RE (1994) Long-term experimental evolution in *Escherichia coli* II. Changes in life-history traits during adaptation to a seasonal environment. *Am Nat* 144: 432-456.
177. Sawaragi Y, Nakayama H, Tanino T (1985) Theory of multi-objective optimisation. Orlanda: Academic press.
178. Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Research* 31: 1234-1244.
179. Wycuff DR, Matthews KS (2000) Generation of an AraC-araBAD promoter-regulated T7 expression system. *Anal Biochem* 277: 67-73.
180. Lutz R, Bujard H (1997) Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/11-12 regulatory elements. *Nucleic Acids Res* 25: 1203-1210.
181. 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.
182. Huang ZJ (1991) Kinetic fluorescence measurement of fluorescein di-beta-D-galactoside hydrolysis by beta-galactosidase: intermediate channeling in stepwise catalysis by a free single enzyme. *Biochemistry* 30: 8535-8540.
183. Monod J (1949) The growth of bacterial cultures. *Annu Rev Microbiol* 3: 371-394.
184. Chambert R, Treboul G, Dedonder R (1974) Kinetic studies of levansucrase of *Bacillus subtilis*. *Eur J Biochem* 41: 285-300.
185. de Visser JA, Park SC, Krug J (2009) Exploring the effect of sex on empirical fitness landscapes. *Am Nat* 174 Suppl 1: S15-30.
186. Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proc Natl Acad Sci U S A* 106 Suppl 1: 9955-9962.
187. Via S (2002) The ecological genetics of speciation. *Am Nat* 159 Suppl 3: S1-7.
188. Wade MJ, Goodnight CJ (1998) Genetics and adaptation in metapopulations: When nature does many small experiments. *Evolution* 52: 1537-1553.
189. Latta RG, Gardner KM, Staples DA Quantitative trait locus mapping of genes under selection across multiple years and sites in *Avena barbata*: epistasis, pleiotropy, and genotype-by-environment interactions. *Genetics* 185: 375-385.
190. Remold SK, Lenski RE (2004) Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. *Nat Genet* 36: 423-426.
191. Poelwijk Frank J, de Vos Marjon GJ, Tans Sander J (2011) Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell* 146: 462-470.
192. Gillespie JH (1984) Molecular evolution over the mutational landscape. *Evolution* 38: 1116-1129.
193. Monod J, Wyman J, Changeux JP (1965) On the Nature of Allosteric Transitions: A Plausible Model. *J Mol Biol* 12: 88-118.

194. Lewis M, Sochor M, Daber R (2011) Allostery via an Order-Disorder Transition. School of Medicine University of Pennsylvania.
195. Richardson JS (1981) The Anatomy and Taxonomy of Proteins. *Advances in Protein Chemistry* 34: 167-339.
196. 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.
197. Xu J, Matthews KS (2009) Flexibility in the inducer binding region is crucial for allostery in the Escherichia coli lactose repressor. *Biochemistry* 48: 4988-4998.
198. Daber R, Stayrook S, Rosenberg A, Lewis M (2007) Structural analysis of lac repressor bound to allosteric effectors. *JMB* 370: 609.
199. Armstrong KM, Baldwin RL (1993) Charged histidine affects alpha-helix stability at all positions in the helix by interacting with the backbone charges. *Proc Natl Acad Sci U S A* 90: 11337-11340.
200. Swint-Kruse L, Zhan H, Fairbanks BM, Maheshwari A, Matthews KS (2003) Perturbation from a distance: mutations that alter LacI function through long-range effects. *Biochemistry* 42: 14004-14016.
201. Swint-Kruse L, Zhan H, Fairbanks BM, Maheshwari A, Matthews KS (2003) Perturbation from a distance: mutations that alter LacI function through long-range effects. *Biochemistry* 42: 14004-14016.
202. Spronk CA, Folkers GE, Noordman AM, Wechselberger R, van den Brink N, et al. (1999) Hinge-helix formation and DNA bending in various lac repressor-operator complexes. *Embo J* 18: 6472-6480.
203. Spronk CA, Slijper M, van Boom JH, Kaptein R, Boelens R (1996.) Formation of the hinge helix in the lac repressor is induced upon binding to the lac operator. *Nat Struct Biol* 3 916-919.
204. Swint-Kruse L, Matthews KS, Smith PE, Pettitt BM (1998) Comparison of simulated and experimentally determined dynamics for a variant of the LacI DNA-binding domain, Nlac-P. *Biophys J* 74: 413-421.
205. Kalodimos CG, Boelens R, Kaptein R (2002) A residue-specific view of the association and dissociation pathway in protein-DNA recognition. *Nat Struct Biol* 9 193-197.
206. Jessup CM, Bohannon BJ (2008) The shape of an ecological trade-off varies with environment. *Ecol Lett* 11: 947-959.
207. Hawthorne DJ, Via S (2001) Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature* 412: 904-907.
208. Kolaczkowski B, Thornton JW (2008) A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol* 25: 1054-1066.
209. Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process in protein evolution. *Mol Biol Evol* 19: 1-7.
210. Lunzer M, Golding GB, Dean AM (2010) Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* 6: e1001162.

211. Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980-984.
212. Stefankovic D, Vigoda E (2007) Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst Biol* 56: 113-124.
213. Rainey PB, Travisano M (1998) Adaptive radiation in a heterogeneous environment. *Nature* 394: 69-72.
214. Travisano M, Mongold JA, Bennett AF, Lenski RE (1995) Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* 267: 87-90.
215. Casadaban MJ, Cohen SN (1980) Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *J Mol Biol* 138: 179-207.
216. 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.
217. Sole RV, Fernandez P, Kauffman SA (2003) Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion. *Int J Dev Biol* 47: 685-693.
218. Weissman DW, Desai MM, Fisher DS, Feldman MW (2009) The Rate at Which Asexual Populations Cross Fitness Valleys. *Theoretical Population Biology* 75: 286-300
219. Burton OJ, Travis JMJ (2008) Landscape structure and boundary effects determine the fate of mutations occurring during range expansions. *Heredity* 101: 329-340.
220. Weinreich DM, Chao L (2005) Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution Int J Org Evolution* 59: 1175-1182.
221. Watson RA, Weinreich DM, Wakeley J (2010) Genome structure and the benefit of sex. *Evolution* 65: 523-536.
222. Watson RA, Weinreich DM, Wakeley J (2006) Effects of intra-gene fitness interactions on the benefit of sexual recombination. *Biochem Soc Trans* 34: 560-561.
223. Lynch M (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* 8: 803-813.
224. Eldar A, Chary V, Xenopoulos P, Fontes ME, Loson OC, et al. (2009) Partial penetrance facilitates developmental evolution in bacteria. *Nature* 460: 510-514.
225. Crispo E (2007) The Baldwin effect and genetic assimilation: revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity. *Evolution* 61: 2469-2479.
226. Mills R, Watson RA (2006) On Crossing Fitness Valleys with the Baldwin Effect. *Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*: 493-499.
227. Kashtan N, Noor E, Alon U (2007) Varying environments can speed up evolution. *Proc Natl Acad Sci U S A* 104: 13711-13716.

228. Salverda ML, Dellus E, Gorter FA, Debets AJ, van der Oost J, et al. Initial mutations direct alternative pathways of protein evolution. *PLoS Genet* 7: e1001321.
229. Franke J, Klozer A, de Visser JA, Krug J Evolutionary accessibility of mutational pathways. *PLoS Comput Biol* 7: e1002134.
230. Besse M, von Wilcken-Bergmann B, Muller-Hill B (1986) Synthetic lac operator mediates repression through lac repressor when introduced upstream and downstream from lac promoter. *Embo J* 5: 1377-1381.
231. Krämer H, Niemöller M, Amouyal M, Revet B, von Wilcken-Bergmann B, et al. (1987 ) lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. *EMBO J* 6: 1481-1491.
232. Roderick SL (2005) The lac operon galactoside acetyltransferase. *C R Biologies* 328: 568-575.
233. Daber R, Sochor MA, Lewis M Thermodynamic Analysis of Mutant lac Repressors. *J Mol Biol* 409: 76-87.
234. Sartorius J, Lehming N, Kisters B, von Wilcken-Bergmann B, Muller-Hill B (1989) lac repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges. *Embo J* 8: 1265-1270.
235. Camas FM, Alm EJ, Poyatos JF (2011) Local gene regulation details a recognition code within the LacI transcriptional factor family. *PLoS Comput Biol* 6: e1000989.
236. Poelwijk FJ, Heyning PD, de Vos MG, Kiviet DJ, Tans SJ (2011) Optimality and evolution of transcriptionally regulated gene expression. *BMC Syst Biol* 5: 128.
237. Campbell LG, Snow AA, Sweeney PM (2009) When divergent life histories hybridize: insights into adaptive life-history traits in an annual weed. *New Phytol* 184: 806-818.
238. Knies JL, Izem R, Supler KL, Kingsolver JG, Burch CL (2006) The genetic basis of thermal reaction norm evolution in lab and natural phage populations. *PLoS Biol* 4: e201.
239. Bohannan BJ, Kerr B, Jessup CM, Hughes JB, Sandvik G (2002) Trade-offs and coexistence in microbial microcosms. *Antonie Van Leeuwenhoek* 81: 107-115.
240. Brockhurst MA, Rainey PB, Buckling A (2004) The effect of spatial heterogeneity and parasites on the evolution of host diversity. *Proc Biol Sci* 271: 107-111.
241. Buckling A, Kassen R, Bell G, Rainey PB (2000) Disturbance and diversity in experimental microcosms. *Nature* 408: 961-964.
242. Cooper VS, Lenski RE (2000) The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* 407: 736-739.
243. Hughes BS, Cullum AJ, Bennett AF (2007) Evolutionary adaptation to environmental pH in experimental lineages of *Escherichia coli*. *Evolution* 61: 1725-1734.
244. Roff DA, Gelinas MB (2003) Phenotypic plasticity and the evolution of trade-offs: the quantitative genetics of resource allocation in the wing dimorphic cricket, *Gryllus firmus*. *J Evol Biol* 16: 55-63.
245. Tschirren B, Richner H (2006) Parasites shape the optimal investment in immunity. *Proc Biol Sci* 273: 1773-1777.

246. Haldimann A, Wanner BL ( 2001. ) Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. . J Bacteriol 183: 6384-6393.
247. Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. Proc Natl Acad Sci U S A 99: 16587-16591.

## Summary

### Empirical adaptive landscapes in variable environments

Adaptive landscapes are used as a metaphor to describe the effect of genetic interactions on phenotype or fitness. Walks on the landscape can describe the progress of evolution. Genetic interactions determine the shape of the landscape, and multipeaked landscapes are known for their ability to constrain adaptation to an optimal solution. So far, empirical adaptive landscapes remain scarce. Previous studies that mapped adaptive landscapes have mainly revealed single peaked landscapes, and were performed in a constant environment.

In this thesis chapter 2 and 3 deal with genetic interactions in a constant environment, whereas in chapter 4, 5 and 6 the effect of variable environments on these interactions are discussed. In addition, some effects of variable environments on adaptive evolution are reviewed.

Chapter 2 gives an overview of the current literature on the genetic background dependent effect of mutations, epistasis, and the role of epistasis in adaptive evolution. We have made a distinction between intra and inter-genic epistasis and we have discussed its role in evolution in terms of the shape of fitness landscapes, the origins of robustness and modularity.

Chapter 3 describes the genotype-phenotype landscape describing the *lac* repressor-operator system in *Escherichia coli*. Two amino acids in the *lac* repressor protein and four base pairs in the operator DNA determine the phenotype, the repression value. We found that this landscape is multipeaked, containing in total nineteen distinct optima, direct evolutionary trajectories were found to contain significant decreases in the repression values. Consistent with earlier predictions we found the occurrence of reciprocal sign epistatic interactions at repression minima between the peaks.

In chapter 4 we show that a gene regulatory system in *E. coli* evolved to the predicted optimal regulatory response. We have challenged a synthetic system, regulated by the *lac* repressor with two alternating environments, whose expression-growth relations lead to tradeoffs in both environments. Analysis of the measured tradeoffs allowed the optimal response to be predicted, and evolution experiments showed the adaptation to the predicted response. We show that the trajectory towards the optimum was constrained in sequence space, which delayed adaptation. However, when this constraint was overcome by mutation, adaptation proceeded

towards the optimal *lac* repressor, which had an inverse allosteric response to the ligand IPTG.

Chapter 5 shows that evolutionary constraints, which have so far mainly been studied in a constant environment, are environment-dependent. We have used *lac* repressor mutants from chapter 4 as a case study, and have systematically reconstructed all intermediates between the wild type and the evolved mutants. We find that one key mutation controls how the other mutations interact with inducer IPTG. In the absence of this key mutation, the other mutations alter the expression level predominantly in the presence of inducer, while in the presence of this mutation the other mutations alter the expression predominantly in the absence of inducer. These observed higher-order interactions between genotypes and the environment, should have implications for evolution in variable environments, for instance they will have an effect on the evolutionary dynamics, which may go unnoticed in phylogenetic studies.

In chapter 6 the genotype-phenotype landscape spanning the sequence space between a suboptimal mutant and the optimal wild type *lac* repressor in *E. coli* is resurrected. Two amino acids in the *lac* repressor and four base pairs in the operator-DNA determine the specificity of binding of the *lac* repressor to the operator DNA, which determines the repression, in the absence of inducer IPTG. In the presence of inducer, the *lac* repressor is released from the DNA and expression can be measured. We find that both constant environment repression and expression landscapes are multipeaked. Further, we find that the landscapes are negatively correlated, such that mutations that decrease the repression in the absence of IPTG increase the expression in the presence of IPTG, and vice versa. This correlation leads to the presence of tradeoffs in many of the in total 720 trajectories from the suboptimal mutant to the wild type. Counter-intuitively, these tradeoffs facilitate adaptation by allowing an evolving population to follow trajectories that alternate between the fixation of a beneficial mutation in one environment, and the fixation of the next mutation in the second environment, ultimately arriving at the optimum. This indicates that tradeoffs not only constrain evolution in variable environments, but that they can also aid the adaptive progress in variable environments.

## Samenvatting

### Empirische adaptieve landschappen in variabele omgevingen

Adaptieve landschappen worden in evolutionaire studies gebruikt als een metafoor om het effect van genetische interacties op het fenotype of op fitness te beschrijven. Genetische interacties bepalen de vorm van het landschap; landschappen met meerdere pieken kunnen de evolutie naar de optimale oplossing beperken. Mutationale paden, die de verschillende genotypes met elkaar verbinden, kunnen het verloop van de evolutie weergeven. Tot nu toe zijn nog weinig experimentele adaptieve landschappen bestudeerd. De onderzoeken die wel systematisch de nabijliggende genotypes in kaart hebben gebracht, hebben vooral landschappen gevonden met een enkele piek, en deze studies werden vooral uitgevoerd in een constante omgeving.

Dit proefschrift beschrijft in hoofdstukken 2 en 3 de genetische interacties in constante omgevingen, terwijl in hoofdstukken 4, 5 en 6 ook variabele omgevingen aan bod komen.

Hoofdstuk 2 geeft een overzicht van de bestaande literatuur over het genetisch-achtergrond-afhankelijk-effect van mutaties, epistasie, en de rol van epistasie in adaptieve evolutie. We hebben onderscheid gemaakt tussen intra- en inter-gen-epistasie, en we beschrijven hun rol in de evolutie, met de nadruk op het effect van epistasie op de vorm van fitness landschappen, de herkomst van robuustheid en modulariteit.

Hoofdstuk 3 beschrijft het genotype-fenotype landschap van het *lac* repressor-operator systeem in *Escherichia coli*. Twee aminozuren in het *lac* repressor-eiwit en vier baseparen in het operator-DNA bepalen het fenotype, de repressiewaarde. We vinden dat dit landschap negentien verschillende pieken heeft, dus de directe evolutionaire paden bevatten aanzienlijke verlagingen van repressiewaarden. In overeenstemming met met eerdere theoretische voorspellingen, vonden we reciproke epistasie in het minimum van de meest gunstige paden.

In hoofdstuk 4 laten we zien dat een genregulatie systeem in *E. coli* evolueert naar de voorspelde optimale regulatie respons. We hebben het synthetische systeem, gereguleerd door de *lac* repressor, blootgesteld aan twee verschillende omgevingen die een trade-off veroorzaken in de relatie tussen expressie van het systeem, het fenotype, en groei, fitness. Analyse van de gemeten trade-offs gaf ons de mogelijkheid om de optimale respons te voorspellen, en evolutie in de twee verschillende

omgevingen leidde, zoals voorspeld, naar de optimale respons. We vonden dat het evolutionaire pad naar de optimale response was belemmerd, waardoor adaptatie was vertraagd. Door mutaties kon deze belemmering weer worden opgeheven, waarna de adaptatie weer kon worden voortgezet naar de optimale *lac* repressor. Deze *lac* repressor had een omgekeerde allosterische respons op de ligand IPTG.

Hoofdstuk 5 laat zien dat evolutionaire belemmeringen, die tot nu toe vooral bestudeerd zijn in een constante omgeving, afhankelijk zijn van de omgeving. Hiervoor hebben we van *lac* repressor mutanten uit hoofdstuk 4 systematisch de tussenliggende mutanten gereconstrueerd tussen de wilde type en de omgekeerde *lac* repressor. We vonden dat één mutatie bepaalde hoe de andere mutaties interacteerden met de inducerende ligand IPTG. De overige mutaties veranderen het expressie niveau vooral in de afwezigheid van deze specifieke mutatie en in de aanwezigheid van IPTG, terwijl in de aanwezigheid van deze specifieke mutatie, de overige mutaties vooral de expressie veranderden in de afwezigheid van IPTG. In het algemeen zullen deze hogere-orde interacties tussen het genotype en de omgeving implicaties hebben op het verloop van de evolutie in variabele omgevingen, bijvoorbeeld op de evolutionaire dynamiek, dit zou onopgemerkt kunnen blijven in fylogenetische studies.

In hoofdstuk 6 is het genotype-fenotype landschap tussen een suboptimale variant en de optimale wilde type *lac* repressor in *E. coli* experimenteel gereconstrueerd. Twee aminozuren in de *lac* repressor en vier base paren in het operator-DNA bepalen de repressie, de mate van binding aan het DNA in de afwezigheid van IPTG. In de aanwezigheid van IPTG laat de *lac* repressor los van het DNA waardoor de expressie gemeten kan worden. We vinden dat de repressie en expressie landschappen beide meerdere pieken hebben. Doordat de landschappen negatief gecorreleerd zijn, zodat mutaties die de repressie verminderen, de expressie verbeteren en vice versa, hebben veel van de 720 paden in deze landschappen trade-offs. Counter-intuïtief kunnen deze paden met trade-offs de evolutie juist versnellen; door het fixeren van de mutaties in de ene omgeving, afgewisseld door het fixeren van mutaties in de andere omgeving, kunnen evolutionaire paden als het ware meanderen over de twee adaptieve landschappen, waardoor ze steeds voordelige mutaties kunnen fixeren in de ene omgeving, dan wel in de andere omgeving. Dit laat zien dat trade-offs, die bekend staan om het belemmeren van evolutie in verschillende omgevingen, adaptieve evolutie juist kunnen helpen het optimum te bereiken.

## Dankwoord

Het doorlopen van een promotie doe je niet alleen, net als Dante in zijn *Commedia* een gids nodig had om boven te komen, wil ik jou, Sander, hierbij bedanken dat je samen met mij het multipeaked pad van een PhD bent ingegaan. Ik ben op AMOLF gekomen om nieuwe dingen te leren, en te werken aan evolutie. Ik ben altijd van mijn mening geweest, en ben dat nog steeds, dat je het meest kan leren van mensen waar je het niet mee eens bent. Gelukkig waren we het regelmatig niet eens. Je hebt me geïnspireerd paden in te gaan waar ik zelf niet zou zijn gekomen, en ik kon je humor (ja, ook) erg waarderen. Je hebt het vermogen om ideeën van een nieuwe kant te bekijken, op te pakken, rond te draaien en met een nieuwe shiny kant naar voren weer neer te zetten, dat, naast je inzicht in goede vragen, en nog een boel dingen die ik niet allemaal ga opnoemen omdat je niet zo houdt van boekhoud-geschreven-regels, maakte je voor mij een goede begeleider, die mij uitdaagde verder te kijken dan het resultaat wat voor mijn neus lag. Sander, super bedankt!

Daarnaast wil ik de groepsgenoten door de jaren bedanken. Ten eerste Daan, naast groepsgenoot hebben we ook een aantal jaren de office gedeeld, jij was mijn vraagbaak voor (rare/ ethische/wetenschappelijke) vragen, en je had altijd kwam een zinnig antwoord klaar- nog steeds! Frank, ook jij was 'van onschatbare waarde' (jaja, mooi cliché-edoch waar), ik heb nog steeds je overgeorven flesjes met FJP (niet FSM) ergens staan, daarnaast was ook jij mijn onuitputtelijke antwoordbaak voor al mijn vragen over de *lac* repressor in the begin van mijn promotietijd. Aileenio, I miss our chats! Mr. Manju, it was great to share the victors with you and your bacteria. Alexandre, miss your conciseness in the lab, great that we are still working together. Moshe, and students Nico, Ndika and Laurens, thanks a lot. Laurens, het denken in generalisten en specialisten is nooit meer weggegaan uit mijn hoofd. Jerien, de bijna-theoreet en mnr v. S.- course maatje, jij deed het coolste experiment gewoon op je stoel in onze office, en het is nog goed gelukt ook! Other group/ office members Bertus and Ienas, it was great to have nice semi-scientific chats. Phillipe N., I am grateful to have you as the other epistatic person! Sarah, the other bacteria girl in the group- 'we are living in a bacterial world, and we are the bacteria- girls' : ). Other students in the lab, Rick, Robert, Merlijn, Alexis, Rajiv, thanks for the atmosphere. Ali, thanks for your help with protein things, I enjoyed our discussions! And off course Vanda, I would like to semi-quote C. H. Waddington, 1952: 'the laborious work of classifying – and cloning of a large number of individuals involved was carried out by my technical assistant, Miss V.S., to whose care and devotion I should pay a tribute..' Thanks soo much! And for the newbees, Fatemeh, Noreen, and office mate Sergey, veel succes en plezier de komende jaren. Peter es war gut zusammen mit dir zu schreiben.

Verder ben ik dank verschuldigd aan E&I voor allerlei computer hulp, speciaal aan Marco Seynen, dank voor de import-victor (en de vele debuggingen). Ook dank aan de vele AMOLF- anderen, een kleine greep, Tarik, bedankt voor het bankschroeven van mijn bevroren epjes en de afstandbediening, Hinko voor het maken van DNA, Tatiana voor de vele bestel-dingen, Frans voor de vriendelijke woorden op de ietwat late (maar voor mij te vroege) ochtenden, en Bela en Mirjam voor het lezen en becommentariëren van mijn aanvraag.

I would also like to thank other AMOLF '(semi-)overlopers' that were here at my start at AMOLF (or nearly), Liedewij, Nienke, Laura, Juju, Tischi, Gert-Jan, Iza, office mates Thorsten, and Paigee, Niels, Meneer Rutger, Koos, Simon, Behnaz, Patrick, Siebe, Maaïke, Lukasz, Wiet wat later, thanks so much for all the nice time, in and outside AMOLF that made me feel at home. Further I would like to pay special thanks to the Pieter-Rein boys, Nils, Filipe, Andrew and Laurens, for all the good company, beers, art and music! Also Pierre- cigars, and Tomek- Balkan music, yes I start to appreciate it, a bit,- are well-thanked for the good company- and cynical comments. And José, my true swimming companion! As well as the other swimmers- Roland, dankjewel voor het leren zwemmen. Andrew, for the good company to yoga, Daan voor de karate, P. N. for some epistatic runs, the girls for the showdance- sorry for the kick, Jeanette, Eva was cool om je andere buurvrouw te zijn.. In addition, Corianne as the great PV/ rondleidingen/ en-nog-zulke dingen- mens en office maatje. Milena the other bacterial girl, thank god we could share some cloning annoyances in our 'complaints of the day' - and François the other lab bench mate, I still think your child looks like your brother. Others I would like to mention are Núria & Sophie, and Björn, and all the others that brightened up the lunch times :).

Also muchas gracias to the people I shared some PV time with, especially Rhoda, with whom I had the pleasure to start some science and society lectures, later gevolgd door Marina (estou c saudades de vc menina!), en Ymkje.

Daarnaast het nanosoil- groepje, wat begon als een enigszins wild plan om de woestijn te dedesertificeren(?), werd verder uitgevoerd, zo ken ik ondertussen het nanolab net zo goed als het biolab, en jullie ondertussen ook al enigszins. Timmo, Jochen, Ivana en Nils - Eens kijken hoe ver we het kunnen brengen!

Hiernaast wil ik graag Cj bedanken voor onze reis, die ons bracht naar verre landen, en dichtbij oorden, die soms nog verder weg leken. We hebben een lange weg afgelegd, veel geleerd, dankjewel voor jouw steun in onrustige tijden.

Ook Anoeek, Linda, Jo, wat hebben we een hoop verhalen beleefd, en opgeschreven, het was leuk om ieder's weegs te volgen- en dan toch weer op het

zelfde punt uit te komen, fysiek – ik ben blij dat jullie weer in de buurt zijn!- maar ook geestelijk...

En natuurlijk ben ik grooten-dank verschuldigd aan Joost, voor de cover!

Asaf, Ran and Dewi, thanks for the end-encouragements. Ran it was great to see that we both looked Chinese when we were working hard in the summer.

Q! Ook wij hebben al een heel pad afgelegd zoals je al zei, ik vind het fijn dat we al zo lang parallel lopen, en vind en vond het superfijn om zoveel met je te mogen delen. Ik ben benieuwd hoe onze toekomst gaat uitpakken!

SKitty, ik vond het erg fijn om met je te discussiëren over vrouwen in de wetenschap, en communicatie dingen, nou ja, discussie, meestal liepen we ons op te winden over dezelfde dingen - binnenkort kom ik naar die Sweiz, want dan is het jouw tijd!

Verder mijn trouwe vrienden Ansa, Kim, mijn paranimfen, en Roel, we kennen elkaar al lang en hebben al veel samen meegemaakt. Ik dank jullie voor jullie steun, enthousiasme, en fijne random dingen. Graag weer fijn samen eten. Jeff, we miss you-come back, Canada is koud en ver - maar komende zomer komen we jouw kant op. Op meer toekomst samen!

And no PhD without music, so especially for all the music suppliers: MfM9gQkfwyg.

Pap en mam bedankt veur mien leave, mer auch veur euge steun in alles wat ich deeg or doa, al leek veul van wat ich dooch neet altied dur veur de hand liggende of mekkelikste weeg. Pap, mam en Ralph zonder euch woar ich neet wo ich noe bin, en woar ich neet weh ich noe bin. Danke.



### List of publications related to this thesis

#### Chapter 2:

Revealing evolutionary pathways by fitness landscape reconstruction  
Kogenaru M, de Vos MG, Tans SJ.  
Crit Rev Biochem Mol Biol. 44(4):169-74 (2009)

#### Chapter 3:

Multiple peaks and reciprocal sign epistasis in an empirically determined  
genotype-phenotype landscape  
Dawid A, Kiviet DJ, Kogenaru M, de Vos M, Tans SJ.  
Chaos. 20(2):026105 (2010)

#### Chapter 4:

Tradeoffs and optimality in the evolution of gene regulation  
Poelwijk FJ, de Vos MG, Tans SJ.  
Cell. 146(3):462-70 (2011)

#### Chapter 5:

Environmental dependence of genetic epistasis in the *E. coli lac* repressor  
de Vos MGJ, Poelwijk FJ, Tans SJ.  
*Manuscript in preparation*

#### Chapter 6:

Crossing multi-peaked landscapes in variable environments  
de Vos MGJ, Dawid A, Tans SJ.  
*Manuscript in preparation*

Optimality and evolution of transcriptionally regulated gene expression.  
Poelwijk FJ, Heyning PD, de Vos MG, Kiviet DJ, Tans SJ.  
BMC Syst Biol. 16(5):128 (2011)



## Curriculum Vitae

Marjon G. J. de Vos was born on 8 June 1981 in Heerlen, The Netherlands. She finished her Secondary Education at Sintermeerten college in Heerlen in 1999, and started her studies in Biology, with a focus on molecular biology and microbiology in the same year. During her MSc she participated in three research projects. In one project she contributed to the elucidation of the five-carbon metabolism of the hyperthermophilic archaeon *Sulfolobus solfataricus* at the department of biological and chemical engineering of The University of Sheffield, under supervision of dr. A.P.L. Snijders and Prof. P. C. Wright of The University of Sheffield, and Prof. J. van der Oost and dr. S. Brouns of Wageningen University. In the laboratory of Molecular Biology of Wageningen University she worked under supervision of Dr. H. Franssen and Prof. T. Bisseling on nodule formation in leguminous plants. Further, she did an internship at the Instituto de Microbiologia Prof. Paulo de Góes at the Universidade Federal do Rio de Janeiro in which she investigated the differences in nitrogen fixing microbial populations under different fertilization regimes in the rhizosphere of Sorghum. The work was supervised by Prof. L. Seldin from the UFRJ and Prof. H. Smidt from Wageningen University. Directly after her studies in 2006 she worked for four months on a research project at the the Food and Biobased Research division of Wageningen University and Research center under supervision of J. Springer, Dr. H. Mooibroek, Dr. A. van Amerongen. At the start of 2007 she started her PhD at FOM institute AMOLF.









