

**The *lac* Operon:
Fluctuations, Growth and Evolution**

DAAN KIVIET

The *lac* Operon: Fluctuations, Growth and Evolution

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 vrijdag 19 november 2010 om 12:30 uur
door

Daniel Johannes KIVIET

doctorandus in de Scheikunde
geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. S. J. Tans

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. S. J. Tans,	Technische Universiteit Delft, promotor
Prof. dr. C. Dekker,	Technische Universiteit Delft
Prof. dr. U. Gerland,	Ludwig Maximilian Universiteit München
Prof. dr. ir. J. J. Heijnen,	Technische Universiteit Delft
Prof. dr. M. Lässig,	Universiteit van Keulen
Dr. H. J. E. Beaumont,	Technische Universiteit Delft
Dr. T. Shimizu,	AMOLF



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: Het *lac* Operon: Fluctuaties, Groei en Evolutie.

ISBN: 978-90-77209-42-4

A digital version of this thesis including links and additional color figures 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.

Cover by Pauline Min and Daan Kiviet. Typesetting with \LaTeX using the memoir class and the Adobe Utopia font. Digitally printed on 90 g/m² groundwoodfree offset paper by Ponsen & Looijen, The Netherlands.

© 2010 by D. J. Kiviet

Contents

1	Introduction	7
1.1	<i>Escherichia coli</i>	7
1.2	Causation: molecular processes underlying the consumption of lactose	9
1.3	Function: benefit of lactose consumption and its regulation	10
1.4	Phylogeny: evolution of <i>lac</i> regulation	12
1.5	Hysteresis: effect of fluctuations in protein level	13
1.6	Thesis outline	13
2	Empirical fitness landscapes reveal accessible evolutionary paths	15
2.1	Enzyme evolution	17
2.2	Evolution of molecular interactions	20
2.3	Outlook	22
3	Evolutionary potential of a duplicated repressor-operator pair	23
3.1	Divergence model	25
3.2	Results	28
3.3	Discussion	31
3.4	Materials and methods	35
3.5	Appendix	37
4	Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape	41
4.1	Introduction	42
4.2	Description of the system	44
4.3	Algorithm	46
4.4	Results	46
4.5	Discussion	51
5	Simple rules underlie an empirically determined genotype-phenotype landscape	53
5.1	Results and Discussion	56
5.2	Conclusion	62
5.3	Experimental	64
5.4	Appendix	65

6	Noise propagation in metabolic networks	67
6.1	Decoupling of the <i>lac</i> system	69
6.2	Single cell measurements of protein level and growth	72
6.3	Correlations between protein level and growth	79
6.4	Temporal correlations in cell lineages	83
6.5	Conclusion	88
6.6	Future work	94
6.7	Materials and methods	96
	Bibliography	109
	Summary	122
	Samenvatting	123
	Dankwoord	125
	List of Publications	126
	Curriculum Vitae	127



Introduction

A cell essentially consists of a lot of molecules bumping into each other. When you realize that the behavior of the cell is formed by these randomly moving molecules, one cannot help but wonder: How? There is no puppet-master manipulating these molecules, making sure that everything goes alright. One can imagine, however, that hidden processes or something like a complex computer code are at work. The thrill of cracking this code and reaching a better understanding of life, makes fundamental research into the workings of living cells so exciting. Apart from understanding how the interaction of dead molecules form living cells, another tantalizing question is how cells came to be? How did complex structures evolve and how do they continue to adapt? Fundamental research into the functioning and evolution of biomolecular systems aims to answer these fundamental questions of life. This thesis is part of this endeavor.

As most fundamental research, the work described here is performed on a model system, allowing us to stand on ‘gigantic shoulders’ consisting of the work of thousands of researchers before us. This thesis has one molecular system as its study object: the *lac* operon. This system consists of a set of proteins involved in the consumption of lactose. When trying to understand why the *lac* operon functions as it does, one can ask very different types of questions, ranging from how it physically functions, to what its use is and how it evolved. In the 1960s the Dutch Nobel laureate Niko Tinbergen defined a framework dividing such questions into four categories (see Box 1.1). In this chapter we will introduce different aspects of the *lac* system by means of these four categories. This serves to distinguish the two separate research questions that we address in this thesis, while at the same time clarifying their joint goal: understanding all aspects of a single biomolecular system. But before introducing our model system, we will briefly describe where one finds the *lac* system: inside *Escherichia coli* bacteria.

1.1 *Escherichia coli*

Escherichia coli is one of the most studied model organisms, especially in genetic and metabolic research. It is a Gram-negative bacterium that resides inside the intestines of mammals. Some *E. coli* strains are pathogenic, but most live in symbiosis with their host. It is highly adapted to life in the small intestines, but its habitat also includes the colon, feces, soil and water environments [1, 2]. It has a remark-

Box 1.1: Tinbergen's four questions.

Niko Tinbergen was a biologist that studied animal behavior patterns. He formulated four categories of questions that can be asked in order to explain any behavior. What is the evolutionary use of the behavior? How can the behavior be explained from its evolutionary history? How did the development of the organism affect the behavior? What are the physical processes that bring about the behavior? The figure below shows a scheme where these questions are organized by the different levels they act on: species vs individual and present vs historic. Interestingly, this scheme is not only useful for animal behavior, but also for biological phenomenon on the molecular level; biomolecular systems carry out a task just as animals perform a particular behavior. Molecular systems, however, do have a very different type of development than animals. Therefore, we have generalized this category to include all historic events that affect the functioning or behavior of a system. Typical questions that can be asked about the functioning of the *lac* operon are added in the figure as examples.

	SPECIES	INDIVIDUAL
PRESENT	<p>Function</p> <p>How do cells benefit from expression of the third lac protein: LacA?</p>	<p>Causation</p> <p>Which molecules can interact with lac repressor to induce expression of the other lac proteins?</p>
HISTORIC	<p>Phylogeny</p> <p>Did lac repressor-operator binding evolve to avoid cross-binding with other Lac family regulators?</p>	<p>Hysteresis</p> <p>Does lac regulation depend on exposure to lactose in the past?</p>

able ability to live in these different conditions, being able to grow both aerobically and anaerobically, and on a wide range of nutrients. *E. coli*'s ability to grow rapidly in many environments, but also to survive when conditions prevent growth, makes it very suitable for laboratory experiments. Intriguingly, practically all laboratory *E. coli* cells descend from a single 'founder' cell, which was obtained from a stool sample of an anonymous diphtheria patient in Palo Alto at 1922 [3].

E. coli cells are very small. Their average length is about 3 μm and their diameter about 1 μm . While their life takes place at a much smaller scale than that we humans experience (see Fig. 1.1), other physical rules apply. For example, viscous forces dominate their movement and inertia is completely irrelevant [4]. Also, the influx of metabolites is governed by diffusion, and cell movement does not matter for the rate of nutrient uptake [4]. But arguably the most important difference between the world of the single cell and ours, is that cells are more affected by the random heat motion of molecules. Processes in the macroscopic world generally consist of sufficient molecules such that their collective random action results in statistically deterministic behavior. Cells also consist of many molecules (life re-



Figure 1.1: Size of *E. coli* compared to human and molecule scale.

Different magnifications of *E. coli* cells drawn to scale on top of one euro a coin. Each subsequent image has a 100-fold magnification. Whereas two magnifications of a 100-fold are required to get from the macroscopic 'human' scale to the scale of bacteria (third drawing), a single subsequent 100-fold magnification brings us to the scale of molecules. Molecules drawn in the fourth image include DNA (long shaped) and proteins. The second image shows the southwest tip of England on the coin.

quires to be somewhat deterministic [5]), but many of its molecular systems consist of only a few particles and behave in a stochastic nondeterministic fashion.

A typical *E. coli* cell consists of two thirds water and one third solid material. The majority of the solid material (also called dry weight) consist of 4 elements: carbon (50%), oxygen (20%), nitrogen (14%) and hydrogen (8%) [6]. These elements mainly reside in macromolecules (96%) and only for a small fraction as metabolites and building blocks (see Table 1.1). More than half of the dry mass of a cell consists of protein. A fifth is present as RNA, which mainly consists of ribosomal RNA (rRNA) and to a lesser content of transfer RNA (tRNA) and messenger RNA (mRNA). Furthermore, a significant portion of mass is located in the cell wall (see Table 1.1).

Living organisms can be regarded as self-replicating entities. The evolutionary success of unicellular organisms such as *E. coli*, depend upon replicating at a high rate. As François Jacob put it: 'The dream of every cell is to become two cells'. The *E. coli* cell only needs some salts and sugar to accomplish this. Within an hour it can convert these molecules by hundreds of enzyme-catalyzed biochemical reactions into a faithful, living, copy of itself. The metabolic process that makes this possible, can be divided into two parts. First, catabolic reactions degrade the imported nutrients to precursor metabolites and energy. These reactions are also referred to as fueling reactions. Next, anabolic reactions convert this material and energy into the macromolecules that the cell consists of (see Table 1.1). When only a single sugar is offered to the cell, specific catabolic reactions become essential. In that case, not only all energy, but also all carbon atoms used for cell material are derived from the reactions that degrade this specific sugar. For the sugar lactose these reactions are carried out by the *lac* operon.

1.2 Causation: molecular processes underlying the consumption of lactose

An enormous amount of research has led to a detailed molecular understanding of *E. coli*'s system for the consumption of lactose [7]. The lactose sugar can serve as *E. coli*'s sole source of carbon and energy. To this end, the cell produces two proteins

that are specific for the catabolism of this sugar (see Fig. 1.2). The first protein imports lactose into the cell, which is separated from its environment by a wall consisting of two membranes. The outer membrane contains protein complexes (porin) through which lactose unselectively diffuses. The inner cell membrane is a more selective boundary and here the galactoside permease protein (LacY) imports lactose into the cell. Lactose that is imported into the cell is subsequently degraded into glucose and galactose by the second protein, β -galactosidase (LacZ). These monosaccharides are further catabolized by their own specific pathways.

Generally, LacY and LacZ are only present in the cell in very low amounts. But when lactose is the only (or most suitable) carbon source in the environment, synthesis is increased by as much as a 1000-fold (see Table 1.2). This is achieved by regulation of the expression of proteins from their respective genes (see Fig. 1.2). The genes that code for the *lac* proteins are clustered in a small region of *E. coli*'s DNA, called the *lac* operon. Expression of the *lac* enzymes is regulated by the *lac* repressor (LacI). When lactose is absent, the LacI protein binds to the *lac* operator, blocking expression of the *lac* enzymes by RNA polymerase. In the presence of lactose, the LacI repressor does not bind the *lac* operator, and RNA polymerase can express the *lac* genes. Together with LacY and LacZ a third protein, LacA, is expressed. This enzyme transfers acetyl groups to lactose and similar molecules, but is not essential for lactose metabolism.

1.3 Function: benefit of lactose consumption and its regulation

Why does *E. coli* consume lactose, and why does it regulate the expression of the proteins necessary for this? Convincing answers to these questions about the function of the *lac* operon can be formulated when one considers the environment where *E. coli* lives. One of *E. coli*'s important habitats are the small intestines of mammals [1, 2]. Here they may encounter lactose that has not yet been consumed by their host. Obviously, this energy rich sugar provides opportunities for cells that are able to consume it, hence the evolutionary advantage of having the *lac* operon, and the benefit of *lac* expression when lactose is present.

For long periods of time, however, the environment will not contain lactose. At

Substances	Fraction of dry weight	Molecular subgroups
Protein	55%	
RNA	20%	rRNA, tRNA, mRNA
Cell wall	15%	lipid, lipopolysaccharide, peptidoglycan
DNA	3%	
Glycogen	3%	
Others	4%	building blocks, metabolites, inorganic ions

Table 1.1: Molecular composition of a typical *E. coli* cell.

Groups of macromolecules and their fractional contribution to the dry weight of an *E. coli* cell. Note that the cell wall contains more substances than indicated, such as protein. Data from [6].

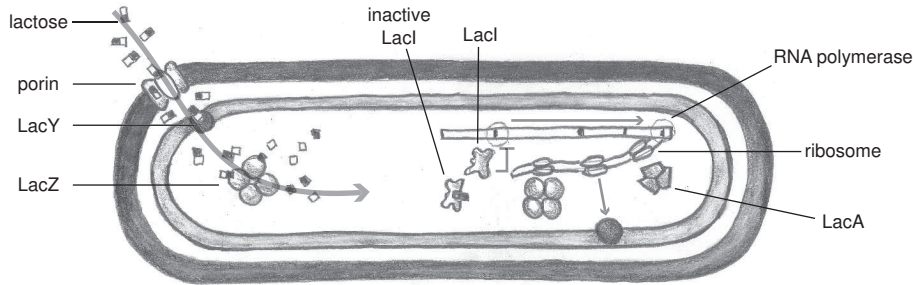


Figure 1.2: Schematic overview of metabolic and genetic features of the *lac* operon

The *lac* permease (LacY) selectively imports lactose into the cell. Subsequently, lactose is degraded by β -galactosidase (LacZ) resulting in substrates for central metabolism. Expression of the *lac* proteins is regulated by the *lac* repressor (LacI). When lactose is absent the LacI protein binds to the *lac* operator, blocking expression of the *lac* enzymes by RNA polymerase. Otherwise, lactose binds to LacI, allowing RNA polymerase to transcribe the *lac* genes into mRNA which is translated by ribosomes into protein.

such times, the *lac* proteins are useless and their production can be a large burden to the cell, leading to decreases in fitness [8]. Regulation of the *lac* operon by LacI allows cells to express *lac* proteins only when they are needed, such that the cost of protein production is always outweighed by the benefit from the action of the proteins.

However, environments are more complex than just having lactose or not having lactose. First of all, apart from lactose other sugars may be available in the environment. In such a case the benefit of expressing the *lac* proteins compared to not expressing them, depends on whether other sugars are consumed and used for growth. Optimally, those enzymes that provide the highest growth rate are expressed. It seems that regulation of the *lac* proteins are adapted to such consider-

Property	Typical Value
Copies of <i>lac</i> DNA / cell	~1.7
Linear length DNA <i>lac</i> operon	~3 μm , about the circumference of <i>E. coli</i>
Time to transcribe <i>lacZ</i>	~half a minute (80 bp/sec)
Time to translate LacZ	~half a minute (40 aa/sec)
Ribosomes / cell	~10.000
RNA polymerase / cell	~3.000
LacZ / cell when repressed	~3 tetramers
LacZ / cell when fully induced	~3.000 tetramers
Relative levels of LacZ, LacY and LacA	4:2:1
Protein diffusion time across cell	~0.1 sec, at diffusion constant $D = 10 \mu\text{m}^2/\text{sec}$
Metabolite diffusion time across cell	~0.001 sec, at diffusion constant $D = 1.0 \mu\text{m}^2/\text{sec}$

Table 1.2: Typical parameter values for an average *E. coli* cell

ations [9]. If for example the environment contains both glucose and lactose, the *lac* genes are repressed, resulting in the preferred metabolism of glucose on which it can grow faster.

Another way how the environment favors complex behavior in the regulation of the *lac* proteins, lies in the continuous scale between low and high amounts of lactose. When less lactose is present to consume, fewer *lac* proteins are required to degrade it at the same rate. Hence, at different lactose concentrations other *lac* expression levels are optimal with regard to the cost of protein production and benefit of protein action. The natural regulation function of the *lac* genes by LacI with regard to lactose concentration in the environment is very similar to the predicted optimal function [10, 11]. Hence, the form of regulatory function of *lac* might be explained from growth advantages and fitness.

1.4 Phylogeny: evolution of *lac* regulation

Due to the lack of historic sources, the study of evolutionary history of the *lac* operon is rather restricted. Although research into the evolution of *lac* is difficult, it is not impossible and at least three methods that look at present-day data can be used. Research has mainly focused on the *lac* repressor, LacI, being the key biological study subject of bacterial regulation.

One study method into evolutionary history is the comparison of current proteins with shared ancestry. A large group of regulatory proteins in *E. coli* and other organisms share a common ancestor with *lacI* [12, 13]. These proteins form the LacI/GalR family and are homologous in their structure, function and DNA sequence. Interestingly, there is also homology in the operators that these proteins bind [12, 14]. Comparisons of DNA sequence of the proteins showed that they diverged long ago [15], making it impossible to retrace subsequent mutations during their divergence process [16]. A large part of the sequence divergence is likely due to genetic drift, but early on some mutations probably caused changes in specificity. Conservation of the recognition domains in the proteins, indicate that for LacI and some other proteins only a few non-conserved amino acid locations serve to discriminate between different operators [12]. But the details of this divergence process are unknown.

Another approach in evolutionary biology is experimental evolution, where the process of evolution is studied in a controlled environment. When *E. coli* cells lacking the *lacZ* gene experience intense selection for growth on lactose, mutations in the *ebgA* gene consistently occur, restoring the ability to hydrolyze β -galactosides [17]. More recently, this approach has been applied on regulation of the *lac* operon. Cells were shown to rapidly adapt their expression levels to those that were predicted to be optimal for their environment [10]. Experimental evolution of the *lac* operon in fluctuating environments also indicated that its adaptation can be understood in terms of the tradeoff between underlying costs and benefits [18].

A third way in which the evolutionary history of the *lac* operon can be studied involves the creation of laboratory mutants. By replacing amino acids by other amino acids, the immediate mutational neighborhood of any protein can be char-

acterized. Such studies showed that more than half of the possible mutations in the *lac* repressor have negligible effect on function while the remaining are generally deleterious [19]. By selecting those amino acid residues that are key to protein function (in LacI's case DNA-binding specificity) and systematically characterizing relevant mutations, molecular fitness landscapes can be constructed (see chapter 2 and 3). Analysis of these, until recently unknown, landscapes allows the study of the step-by-step evolution of molecular functions. It also allows to address the question what molecular fitness landscapes look like (see chapter 4 and 5).

1.5 Hysteresis: effect of fluctuations in protein level

A well-known example of historic events affecting the function of the *lac* system is its 'all-or-none' behavior: when cells are exposed to a low amount of *lac* inducer, their response depends on whether they have been exposed to the inducer earlier [20]. This hysteric behavior can be explained from a negative feedback mechanism in which the presence of sufficient LacY proteins is necessary for the induction of the *lac* operon [20, 21]. As a result a population of cells may exhibit widely varying *lac* expression levels, which may have an adaptive function [21, 22].

Interestingly, homogeneous populations of cells that are exposed to identical environments also exhibit a wide variability in *lac* levels. Recently, this variability has been shown to be caused by strong fluctuations in gene expression [23, 24]. These fluctuations are likely due to the inherent stochastic nature of molecular processes in the cell. In growth conditions with low amounts of *lac* inducer these fluctuations may lead to bistability: a subpopulation with low and one with high *lac* expression [21]. Generally, *lac* protein production was shown to fluctuate dynamically, resulting in significant variability of *lac* levels at any growth condition [25].

How this noise in protein level affect the fitness and growth rate of cells has mainly been investigated theoretically. Such studies predict that noise in *lac* components alters the optimal regulation function of LacI and can explain the optimal number of LacI proteins in the cell [11, 26]. It has also been predicted that *lac* fluctuations reduce the mean growth rate when the average *lac* level is close to its optimal level [26]. Whether fluctuations in *lac* level dynamically propagate to growth rate in single cells is not known due to unknown dependencies between cellular metabolism and protein activity (investigated in chapter 6).

1.6 Thesis outline

This thesis is concerned with two distinct fundamental research questions that are both investigated using the *E. coli lac* system. In the first four chapters we investigate what the shape of biological fitness landscapes look like. Chapter 2 reviews recent progress in measurement of empirical fitness landscapes, and introduces the open questions in evolution that they may answer, such as why particular evolutionary paths are taken. In this chapter, we also introduce the concept of epistasis as a useful description of the local shape of fitness landscapes. In chapter 3 we describe existing *in vivo* measurements on *lac* repressor and operator mutants and

show how these can be used to construct a fitness landscape of *lac* regulation. Using computer simulations we simulate mutational pathways and reveal that new regulatory interactions can easily evolve. Chapter 4 deals with the local structure of the *lac* landscape. We determine that the landscape is multi-peaked and, consistent with earlier predictions, show the presence of reciprocal sign epistasis. We conclude our analysis of the *lac* landscape in chapter 5 with a more global analysis of its structure, focusing on which landscape features are important for evolution. This study reveals that the essential features of the *lac* landscape can be sufficiently captured by modeling the presence or absence of additivity between functional residues.

In chapter 6 we turn to another fundamental research question: how do random molecular fluctuations in the number proteins in a single cell propagate to its growth? Again, we use the *E. coli lac* system to investigate this question. But whereas the first part of this thesis consists of theoretical simulations of *lac* regulation, here we perform laboratory experiments on *E. coli* cells that require use of their *lac* enzymes for growth. By means of automated and highly sensitive fluorescence microscopy, we measure both fluctuations in *lac* level and in growth rate in individual growing cells. These experiments show that fluctuations in the growth rate of single cells can be linked to protein fluctuations, but also reveal a intricate dynamic interdependency between these two properties.

Empirical fitness landscapes reveal accessible evolutionary paths



When attempting to understand evolution, we traditionally rely on analysing evolutionary outcomes, despite the fact that unseen intermediates determine its course. A handful of recent studies has begun to explore these intermediate evolutionary forms, which can be reconstructed in the laboratory. With this first view on empirical evolutionary landscapes, we can now finally start asking why particular evolutionary paths are taken.

Evolutionary intermediates represented a central preoccupation for Darwin in his case for the theory of evolution. He remarked, for example: ‘...why, if species have descended from other species by insensibly fine gradations, do we not everywhere see innumerable transitional forms?’ [27]. Although Darwin developed a convincing rationale for their absence, he did realize that the lack of intermediates as proof leaves room for criticism. He noted, for instance: ‘If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down.’ [27]. Indeed, in their opposition to evolution, the proponents of ‘intelligent design’ have seized on our current ignorance of intermediates.

Building on earlier ideas [16, 28–30], an approach has recently been developed to explore the step-by-step evolution of molecular functions. The central innovation is that all molecular intermediates along multiple putative pathways are explicitly reconstructed. Together with a phenotypic characterization of each intermediate, one can determine whether paths towards a certain novel function are accessible by natural selection. Although others have reconstructed and characterized phylogenetically ancestral forms of proteins [29–32], here the focus is on fitness landscapes [33] in which multiple mutational trajectories can be compared. Fitness landscapes have been widely studied on a theoretical level (see refs [34–38] for example), but one can now obtain a glimpse of actual biological landscapes. This view finally allows us to ask which particular evolutionary paths are taken and why. In particular, to what extent do biomolecular properties constrain evolution? Does it matter in which order mutations occur? Are fitness landscapes rugged, with many local optima acting as evolutionary dead-ends, or are they smooth? Is neutral genetic drift essential for a new trait to emerge?

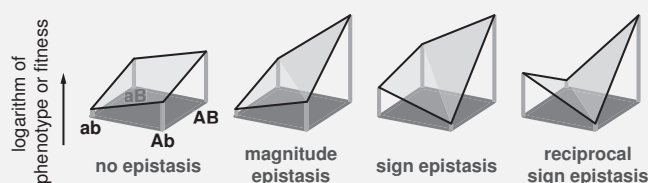
When examining the molecular underpinnings of the evolution of new traits, we distinguish two elementary cases. First, we discuss a single mutable component such as an enzyme. Second, we look at molecular interactions involving two or more mutable components, which is typical for regulatory evolution.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

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

Box 2.1: Epistasis and the accessibility of mutational paths.

Epistasis means that the phenotypic consequences of a mutation depend on the genetic background (genetic sequence) in which it occurs. In the figure below we distinguish four cases that illustrate paths composed of two mutations, from the initial sequence 'ab' towards the optimum at 'AB'. When there is no epistasis, mutation 'a' to 'A' yields the same fitness effect for different genetic backgrounds ('b' or 'B'), while for magnitude epistasis the fitness effect differs in magnitude, but not in sign. For sign epistasis, the sign of the fitness effect changes. Finally, such a change in sign of the fitness effect can occur for both mutations, which we here term reciprocal sign epistasis.

These distinctions are crucial in the context of selection. Mutations exhibiting magnitude epistasis or no epistasis are always favored (or disfavored), regardless of the genetic background in which they appear. In contrast, mutations exhibiting sign epistasis may be rejected by natural selection, even if they are eventually required to increase fitness. In other words, some paths to the optimum contain fitness decreases, while other paths are monotonically increasing. When all paths between two sequences contain fitness decreases, there are two or more distinct peaks. The presence of multiple peaks indicates reciprocal sign epistasis, and may cause severe frustration of evolution (Fig. 2.1b). Indeed, reciprocal sign epistasis is a necessary condition for multiple peaks, although it does not guarantee it: the two optima in the diagram may be connected by a fitness-increasing path involving mutations in a third site.



2.1 Enzyme evolution

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

Whether this picture is realistic was investigated for the adaptation of bacterial β -lactamase to the novel antibiotic cefotaxime [41]. The central step was to reconstruct and measure all likely intermediates, allowing a systematic study of all possible trajectories. The intermediate sequences can be easily identified, because

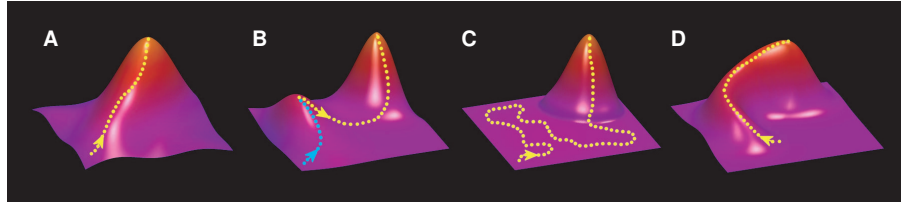


Figure 2.1: Schematic representations of fitness landscape features.

Fitness is shown as a function of sequence: the dotted lines are mutational paths to higher fitness.

(A) Single smooth peak. All direct paths to the top are increasing in fitness.

(B) Rugged landscape with multiple peaks. The light path has a fitness decrease that drastically lowers its evolutionary probability. Along the darker path selection leads in the wrong direction to an evolutionary trap [39].

(C) Neutral landscape. When neutral mutations are essential, evolutionary probabilities are low [37, 40].

(D) Detour landscape. The occurrence of paths where mutations are reverted [39] shows that sequence analysis may fail to show mutations that are essential to the evolutionary history.

the (five) mutations that control the cefotaxime resistance phenotype are known, resulting in $2^5 = 32$ possible mutants. The order in which the mutations are fixed can of course be different, giving rise to $5! = 120$ possible direct trajectories between the start and end sequences.

The trajectory analysis showed that the fitness landscape is not as simple as depicted in Fig. 2.1a. A majority of the pathways towards maximum cefotaxime resistance actually shows a dip in fitness (see light path in Fig. 2.1b), or contain selectively neutral steps (as in Fig. 2.1c), resulting in much smaller chances of being followed by natural selection [37, 40]. For 18 paths however, each step appeared to confer a resistance increase, making these trajectories accessible to Darwinian selection. The part of the fitness landscape mapped out in this manner therefore does appear to have a single peak, but one that contains depressions and plateaus on its slopes. We stress that such three-dimensional analogies, while useful for conveying basic characteristics, do not rigorously represent the many direct trajectories existing between two alleles. Also note that there may be additional paths that contain detours, involving other mutations that are eventually reverted [39] (Fig. 2.1d).

Interestingly, some mutations yielded either a resistance increase or decrease, depending on the preceding mutations. This phenomenon, called sign epistasis [38] (see Box 2.1), is both a necessary and sufficient condition for the fitness landscape to contain inaccessible paths to an optimum [38]. Some cases of sign epistasis could be understood in terms of competing molecular mechanisms. For instance, a first mutation in the wild-type enzyme increased the resistance by enhancing the catalytic rate, even though it also lowered the thermodynamic stability. This loss of stability was repaired by a second mutation, thereby further increasing the resistance. In contrast, when this ‘stabilizing’ mutation occurred first in the wild-type enzyme, the resistance was reduced. Such back and forth balancing be-

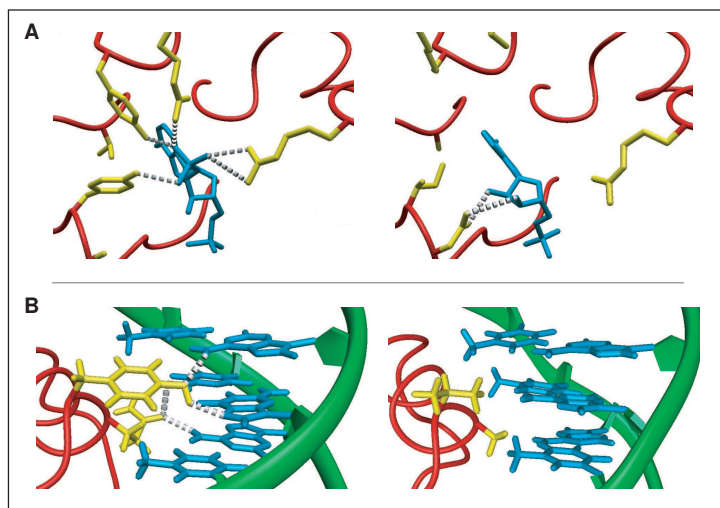


Figure 2.2: Molecular structures in different evolutionary forms.

(A) The left panel shows wild-type *E. coli* isocitrate dehydrogenase [44] (IDH), which is structurally similar to IMDH, with NADP as cofactor. The right panel shows an engineered IDH form with NAD as cofactor [45]. Main chains are shown in grey, cofactor in black, and hydrogen bonds as dashed lines.

(B) The left panel shows a wild-type *E. coli lac* repressor and operator [46]. The right panel shows a *lac* repressor and operator variant, with mutations mimicking the *gal* system [47]. Binding is tight and specific (despite the absence of hydrogen bonds): these variants bind wild-type partners poorly. DNA backbone and key bases are shown in dark grey, repressor chains in black, key repressor residues in grey, and hydrogen bonds as dashed lines. Figures prepared with MOLMOL [48].

tween structural and functional benefits might well be a more general evolutionary mechanism [42, 43].

In a second study [49], the connection between fitness landscape and underlying molecular properties has been explored for the evolution of isopropylmalate dehydrogenase (IMDH, Fig. 2.2a), an enzyme that is involved in the biosynthesis of leucine. As in the previous study, a set of mutational intermediates between different functions were characterized. Here the mutations changed the cofactor binding affinity of IMDH. *In vitro* measurements of enzyme activity did not show epistasis: each mutation gave a fixed catalytic improvement, which was independent of the order in which they occurred. Thus, the ‘enzyme activity’ landscape is single-peaked.

The story becomes more complete with the following elements. First, the study also considered evolutionary paths from the suboptimal cofactor NADP to the normal cofactor NAD [50]. Second, selection does not act directly on enzyme activity, but rather on the fitness of an organism. As fitness is typically nonlinear in enzyme activity, epistasis is introduced. Therefore, the IMDH mutants were also evaluated *in vivo*, providing a direct measurement of the fitness effect of a mutation. The

resulting fitness landscape was shown to contain a depression or valley, rendering the trajectories that pass through it selectively inaccessible. There is an intuitive rationale for a valley here: when the recognition of NADP is reduced, the fitness first decreases, before it rises again when NAD recognition is built up. Interestingly however, some trajectories also exist that avoid the valley by simultaneously increasing NAD, and decreasing NADP recognition. In the end, the genotype-fitness landscape has a single peak, but one that includes a depression on its slope.

2.2 Evolution of molecular interactions

The evolutionary puzzle becomes more complex at a higher level of cellular organization. In the web of regulatory interactions between ligands, proteins and DNA, the components are strongly interdependent, which might suggest that their evolution is severely constrained. The evolution of molecular recognition has recently been explored by two studies, which also used experimentally reconstructed intermediates. The first examined hormone detection by steroid receptors in the basal vertebrates (Fig. 2.3a) [51]. The second [39] looked at the adaptation of repressor-operator binding, in a large evolutionary landscape based on published mutation data for the *Escherichia coli lac* system [14] (Figs 2.2b and 2.3b). For both studies, the molecular interactions may be thought of as a key fitting a lock. The unifying question is: can a new lock and matching key be formed taking just one mutational step at a time? The adaptation of these components presents a dilemma: if the lock is modified first, the intermediate is not viable because the old key does not fit, and vice versa.

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

Surprisingly, these studies again show that new interactions can evolve in a step-by-step Darwinian fashion, despite the mismatching intermediates problem sketched above. In the hormone receptor case, this predicament is overcome by a molecular version of a master key: a putative ancestral ligand, 11-deoxycorticosterone, was found to activate all receptors (ancestral and present-day), allowing the mutational intermediates to remain functional even while the receptors diverged (Fig. 2.3a). The capability to synthesize aldosterone evolved later, finally providing a specific hormone that is recognized by just one of the two receptors. An existing receptor was thus recruited into a new role, as a binding partner to aldosterone, in a process that was termed ‘molecular exploitation’. Sign epistasis was again ob-

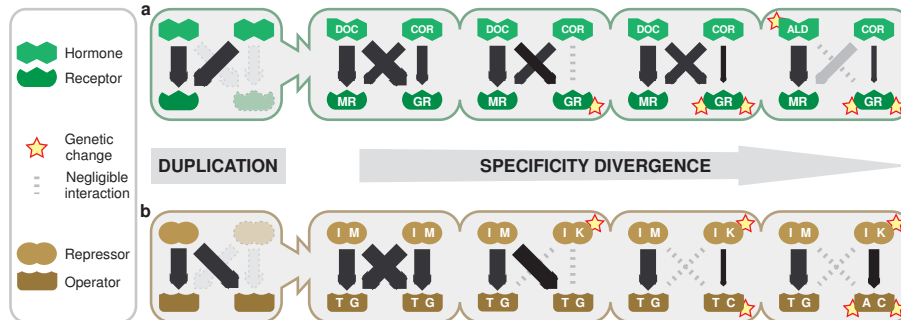


Figure 2.3: Evolution of molecular interactions based on reconstructed intermediates.

Arrow thickness denotes measured interaction strengths. DOC, 11-deoxycorticosterone; COR, cortisol; MR, mineralocorticoid receptor; GR, glucocorticoid receptor; ALD, aldosterone.

(A) Pathway towards independent steroid receptors after duplication, via intermediate receptors that remained sensitive to their ligands [51]. A changed mutation order produced a non-sensitive intermediate, making that path inaccessible. The grey arrow indicates that cortisol is absent in MR-expressing tissues.

(B) Pathway towards independent repressor-operator pairs following duplication, taking single-mutation steps without decreases in network performance. Many paths were compared in a landscape based on over 1,000 *lac* mutants [14], covering all substitutions on all key base pairs. For simplicity, the repressor dimer and two operator half-sites are not drawn.

served: an initial mutation drastically lowered the response to all substrates, but after another mutation, the same mutation improved cortisol response while decreasing the aldosterone response. Thus, just as in the β -lactamase and IMDH cases, at least one selectively accessible evolutionary pathway existed.

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

If the *lac* repressor-operator affinity landscape is rugged and multi-peaked, how can new recognition evolve in a step-by-step manner? The answer lies in the fact that selection does not act on a single interaction. Instead, multiple interactions in a network determine the regulation, and ultimately organismal fitness. In the *lac* case, deteriorations in one interaction were offset by improvements in another.

For example, initial mutations in one repressor duplicate were bad for binding to its designated operator, but good for relieving an undesired cross-interaction (Fig. 2.3b). These results substantiate the suggestion that network robustness [52] may promote evolvability [53, 54]. The observed compensations yielded a smoothed fitness landscape, making the new interactions selectively accessible. In fact, because compensation within biochemical networks is ubiquitously observed [55], we expect that evolution by network compensation constitutes a general mode of regulatory adaptation, molecular interdependence notwithstanding.

2.3 Outlook

The experimental reconstruction of evolutionary intermediates and putative pathways has provided an exciting first look at molecular adaptive landscapes. Although numerous paths appear to be selectively inaccessible, accessible pathways are generally also available. Importantly, various alternative types of fitness landscapes were not observed. The landscapes could have been so rugged and multi-peaked, that accessible paths to optima would not exist, thus requiring, for instance, two or more simultaneous mutations, larger genetic modifications through recombination, or periods of relaxed selection. We have put forward various mechanisms that can reduce landscape ruggedness and improve evolvability. These include the interplay between protein function and stability [41, 49], the exploitation of existing molecules into new roles [51], and compensation within biochemical networks [39].

That only a few paths are favored also implies that evolution might be more reproducible than is commonly perceived, or even be predictable. It is important to note that evolutionary speed and predictability are not determined only by molecular constraints, but also by population dynamics. Population dynamics still presents many open questions, in particular in the context of regulatory evolution and varying environments. The situation in which environmental fluctuations are fast relative to selection timescales has been explored in the repressor divergence study [39]. Recent theoretical considerations [56, 57] may provide promising approaches to address these questions more generally.

The molecular systems interrogated so far represent only a start, but one with great potential to spark further exploration. The analysis of intermediates is generally applicable, which makes finding new candidate systems not difficult. Mutational paths could also be revealed using the directed evolution methodology [58], in which randomly mutated pools are screened. A related approach is the experimental evolution [59] of cells in chemostats [60] or by serial dilution [10, 61]. The advantage of these methods is that more extensive and unbiased evolutionary changes can be explored, although they do not directly reveal why trajectories are chosen. Together, these developments may change the character of molecular evolution research from one that is primarily sequence-based to one that explicitly incorporates structure, function and fitness.

Evolutionary potential of a duplicated repressor-operator pair



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

Initially put forward by Stevens in 1951 [62] and later advocated by Ohno in his seminal work [63], gene duplication followed by functional divergence is now seen as a general mechanism for acquiring new functions [64]. Also, regulatory networks are thought to be shaped significantly by genetic duplication [65]. For instance, sequence analysis of transcription factor families points to various historical duplication events [15, 66]. However, very little is known about the subsequent mutational divergence pathways or about the corresponding stepwise phenotypical changes that are subject to selection. While these issues have not yet been explored experimentally, related generic aspects of mutational plasticity have been addressed theoretically [35, 67–70]. However, a central obstacle in studying mutational pathways through computer simulations remains the unknown relation between the sequence and binding affinity, for which, in general, a rather abstract mapping has to be assumed. To describe the formation of a new regulatory interaction after a duplication event, which is our current aim, such an abstract approach would be particularly speculative.

Here we reason that many characteristics of the adaptation of real protein-DNA contacts are hidden in the extensive body of mutational data that has been accumulated over many years (e.g., [7, 14, 19] for the *Escherichia coli lac* system). These measured repression values can be used as fitness landscapes, in which pathways can be explored by computing consecutive rounds of single base pair substitutions and selection. Here we develop this approach to study the divergence of duplicate repressors and their binding sites. More specifically, we focus on the creation of a new and unique protein-DNA recognition, starting from two identical repressors and two identical operators. We consider selective conditions that favor the evolution toward independent regulation. Interestingly, such regulatory divergence is inherently a coevolutionary process, where repressors and operators must be optimized in a coordinated fashion.

The mere presence of a selective pressure is clearly not a sufficient condition to achieve a new function. Rather, the evolutionary potential and limitations can be seen as governed by the shape of the actual fitness landscape and the evolutionary search within it. Studying these intrinsic limitations to divergence represents the overall aim of this work. Many open questions arise when considering the formation of a new protein-DNA interaction, which may be viewed as the construction of a new lock and uniquely matching key. For instance, how should the protein be modified step-by-step to recognize a new DNA-binding site that also does not yet exist, or vice versa? One would expect that complementary mutations need to occur in the protein and DNA-binding site. Does this mean that temporary losses in fitness must be endured when taking single-mutation steps? And, how many mutations must minimally accumulate before a noticeable new recognition is obtained on which selection can act? The latter is an important point: mutations conferring a selective advantage spread more readily through a population [40], resulting in a drastic increase of the divergence probability. These questions are addressed by exhaustively searching the landscape for optimal pathways, as well as by complementary population dynamics simulations.

Previously it has been shown that *lac* repressor mutants indeed exist that can bind exclusively to mutant *lac* operators [14]. Our simulations reveal that a du-

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

3.1 Divergence model

Selective pressure and the fitness landscape

We consider an ecological situation where natural selection would favor independent regulation of two genes X and Y. Regulation is not independent in the initial symmetric network with duplicated components (see Fig. 3.1): X and Y have two identical upstream binding sites (O_1 and O_2), which bind two identical repressors (R_1 and R_2) equally strongly. Such a situation will, for instance, arise upon duplication of a repressor that regulates two or more genes. Note that this selective pressure, of course, is not a general outcome of a repressor duplication. A duplication event may arise in the context of a different functional pressure, which could direct the evolution toward a different topological motif [71]. Most often, selective pressures for a new function will be absent, in which case silencing of one of the duplicates is the most probable outcome [64, 72]. However, the rare cases where a selective pressure is present are crucial to developing new functions.

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

network is denoted by the product of all optimization factors:

$$\text{Fitness} = \max(F_{O_1}) \left(\frac{F_{O_1 R_1}}{F_{O_1 R_2}} \right) \max(F_{O_2}) \left(\frac{F_{O_2 R_2}}{F_{O_2 R_1}} \right) \quad (3.1)$$

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

The fitness definition comes down to a minimum set of two demands for regulatory binding: each operator must bind a repressor tightly ($\max(F_{O_1})$ and $\max(F_{O_2})$ should be large) but also exclusively ($F_{O_1 R_1}/F_{O_1 R_2}$ and $F_{O_2 R_2}/F_{O_2 R_1}$ should be large). Prior to divergence the first demand is already met, but the latter is not. The challenge during divergence is therefore to improve binding exclusivity, while maintaining tight binding. Tight and exclusive binding is a core functionality of most regulatory systems, and most pairs of existing transcription factors must therefore score well on the employed fitness definition. Take for instance the LacI and RafR repressors, which regulate enzymes required for growth on lactose and raffinose, respectively. If operator binding would not be tight in the absence of lactose and raffinose, the wasteful expression of the downstream metabolic enzymes would lead to sub-optimal growth speeds [8, 10]. If RafR would also bind to the *lac* operator (and thus bind non-exclusively), the effect on growth speed would also be negative since the mere absence of raffinose would then lead to insufficient β -galactosidase for high lactose concentrations.

One therefore expects a conservative selective pressure that minimally includes binding tightness and exclusiveness, to keep the *lac* and *raf* regulation intact. Important here is that the *lac* and *raf* repressors are in fact related: their origin has been traced to duplication events from a common ancestor [15]. If a conservative pressure keeps their function intact now, it seems a good candidate for the initial divergence pressure as well. Full divergence to the current *lac* and *raf* systems clearly involves many additional developments after duplication. For instance, the divergence of ligand-binding properties [73] might have occurred prior to operator-binding divergence. While these considerations put additional constraints on the entire divergence process, they do not alter the particular operator-binding divergence studied here.

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

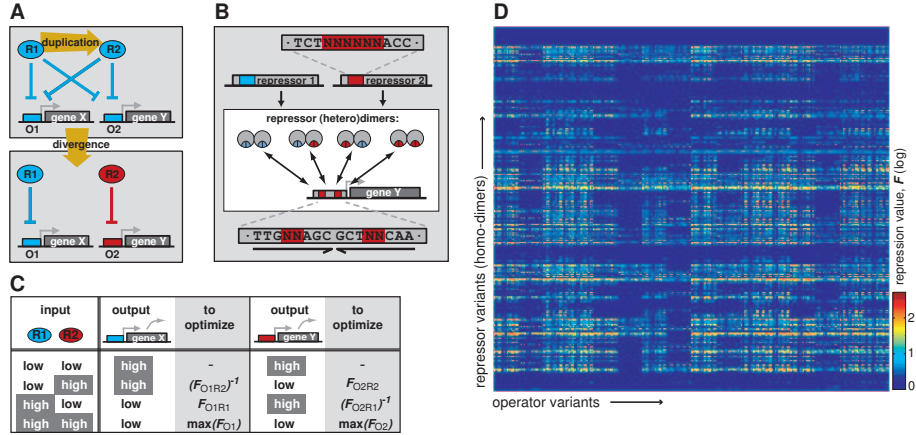


Figure 3.1: Divergence process, fitness criterion, and mutational dataset of repression values.

(A) Diagram of the studied divergence process: after a duplication event, a new regulatory interaction can be formed by mutating the two operators, O_1 and O_2 , and two repressors, R_1 and R_2 .

(B) Duplication and divergence yields heterodimers, which can all bind to the operator. The (initially symmetric) operators and repressors are based on the *lac* sequence, as indicated. Base pairs that are key to altering specificity (colored red and blue) can be mutated to arbitrary sequence.

(C) The selective pressure for independent regulation follows from four input conditions that contribute to the total fitness. When, e.g., R_1 is high and R_2 low, this implies that X should be low and Y high. Out of all interaction parameters of the network, in this case only $F_{O_1R_1}$ and $(F_{O_2R_1})^{-1}$ are relevant and need to be optimized. When R_1 and R_2 are high, both X and Y should be low, regardless of which repressor-dimer causes repression. Therefore $\max(F_{O_1})$ (the strongest interaction with O_1 by either homodimers of R_1 or R_2 or by the heterodimer of R_1 and R_2) and $\max(F_{O_2})$ need to be optimized. When both R_1 and R_2 are low, no parameters need to be optimized.

(D) Resulting repression value landscape, showing repression values based on actual measurements of mutants.

Mutation data and pathway simulations

In our simulations, the strength of a mutant repressor-operator interaction (as expressed by the repression value F), is assigned using data from mutational analysis [14]. In these experiments, repression values have been determined *in vivo* from the repressed and unrepressed expression levels of a *lacZ* gene, controlled by a mutant *lac* operator and mutant *lac* repressor (see section 3.4). Obviously not all possible *lac* mutants have been constructed. Therefore, a potentially significant limitation of our simulations is the restricted number of base pairs that can be mutated *in silico* and linked to experimental data. At the same time however, while the tightness of DNA binding is the result of the integral protein architecture, surprisingly few base pairs (ten in total) have been found to be important for

altering binding specificity [14] (see Fig. 3.1B). Focusing on these key base pairs is therefore reasonable for the minimal paths that we are interested in here. Using measurements on 1,286 mutants, repression values of all variants in these key base pairs could convincingly be determined [7, 14, 74]. These variants thus include all multiple mutants in both the repressor and the operator. Repression values of heterodimers and asymmetric operators are calculated using an additive contribution of the repressor monomers to the dimer-DNA binding [75] (see section 3.4). In total, about $1 \cdot 10^7$ possible repressor-operator combinations are obtained (see Fig. 3.1D for the homodimer variants).

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

The aim of the simulation method (see section 3.4 for details) has been to reveal the intrinsic possibilities for the divergence of repressor-operator binding, given the measured data and the constraints of single base pair substitutions and no fitness decreases. For this purpose, we search the landscape for optimal paths and study what their limitations and potential are. To trace these optimal paths, all mutants with a single base pair substitution with respect to their parents are evaluated based on the fitness described above, and the best performers are selected for the next round. The number of selected mutants L is varied to assess its effect. We also question whether these optimal paths are not just rare cases, by comparing them with pathways generated by a different simulation method, where a random mutation is accepted with a probability that depends on its associated fitness increase [76] (see page 37 of section 3.5).

3.2 Results

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

Most optimal paths are rather short: 70% require just five to nine mutations for

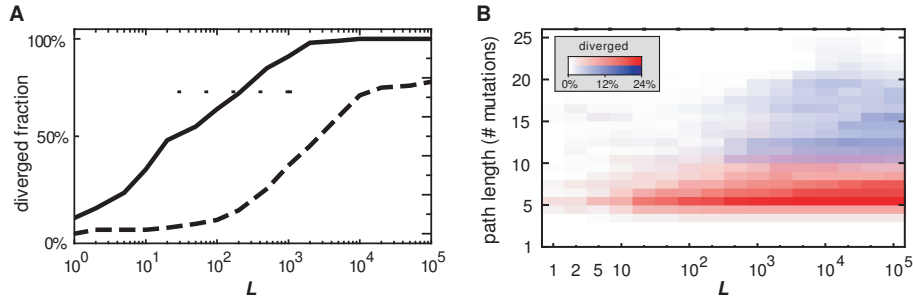


Figure 3.2: Divergence success ratio and path length distributions.

(A) Fraction of starting sequences (numbering 132 in total) that successfully diverge, as a function of the number of networks carried to the next round (L). Dashed line, idem, but with the additional requirement of continued tight binding ($F \geq 100$) for both repressors.

(B) Distribution of path lengths until divergence. Red color map, optimal co-divergence pathways. Blue color map, pathways with the additional requirement of $F \geq 100$ for both repressors. Note that a vertical summation of the color maps yields the lines in (A).

$L = 20$ (Fig. 3.2B). The systems almost exclusively find the nearest diverged state in sequence space (Fig. 3.3B) and do so without taking any detours (Fig. 3.3A). Notably, despite the fact that the starting points lie in very different areas of the sequence space, a generic sequence of network changes is generally observed (see Fig. 3.4 for an example). First of all, one repressor-operator combination remains unchanged, except at the very end, as the other diverges away. This is an example of asymmetric divergence due to positive selection, as has also been found in phylogenetic analysis of duplicate genes in eukaryotes [77]. A striking general feature of the pathways is an early reduction in the binding strength of the diverging repressor, brought about by a single base pair substitution (Fig. 3.4B, red trace). Such a mutation would be unfavorable for a single repressor-operator pair, but here it can be fitness neutral, partly because the unchanged duplicate repressor ensures a continued repression. At this specific point the diverging repressor is freed from functional constraints and therefore most vulnerable to degenerative mutations resulting in silencing of the gene. The probability of silencing is reduced however, because already at the second mutation and onward, new and unique protein-DNA recognition can be built up. At the sequence level, this phase is characterized by transient asymmetries. The operator must go through non-palindromic sequences because it can only receive one mutation at a time. Heterodimers are the best binders in this phase because of their ability to mirror the non-palindromic operator sequences. Eventually all successful trajectories recover palindromic operators, even as the selective pressure does not explicitly specify this. With all dimer varieties present, a homodimer is available and now binds most tightly to the palindromic operator.

In order to obtain a better insight in the essential ingredients for the observed evolvability, various additional simulations were performed. For instance, we were triggered by the recurrent early knockout of one of the repressors, which is one of

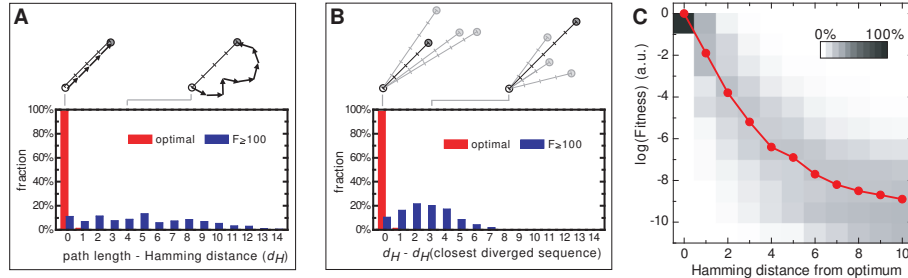


Figure 3.3: Analysis of pathway detours and local environment of fitness optima.

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

the most noticeable features of the mutational pathways. To test for the importance of this step, both repressor-operator pairs were required to maintain a significant repression ($F_{O_1R_1} > 100$ and $F_{O_1R_2} > 100$). Divergence is indeed significantly frustrated by these conditions (Fig. 3.2A, hatched line). The amount of selected mutants needs to be two orders of magnitude larger ($L > 1,000$) for half of the starting sequences to diverge. The saturation of the diverged fraction for very high L , where prolonged neutral drift is allowed, indicates that for 22% of the starting sequences no pathways exist. Moreover, in contrast to the optimal paths, the nearest diverged state in the landscape is generally not found, and the paths contain significant detours (Fig. 3.3). The same is seen from the increased path length: 70% of the paths take 11-21 mutations (Fig. 3.2B). These paths lack a recurring mutation pattern as observed for the optimal paths and instead show a large variation in the sequence of events. Both repressors and operators are significantly mutated, and the fitness increases slowly or is neutral over multiple rounds (see Fig. 3.5 for an example).

Another defining feature of duplicated transcription factors is the heterodimerization of transcription factor monomers. It is not a priori evident whether this constraint on the network topology either promotes or hampers divergence. To assess its effect, simulations were performed where heterodimers are not able to form (data not shown). The results indicated a surprisingly limited effect on the divergence. The paths do initially show a slower fitness increase, but the path length does not appear much affected, nor the success rate of divergence. The other simulation variations we conducted (with unequally weighted factors in the fitness definition), did not qualitatively alter the main divergence features, such as substantial divergence success without fitness decreases, short paths, and an early repression dip, indicating the robustness of our results.

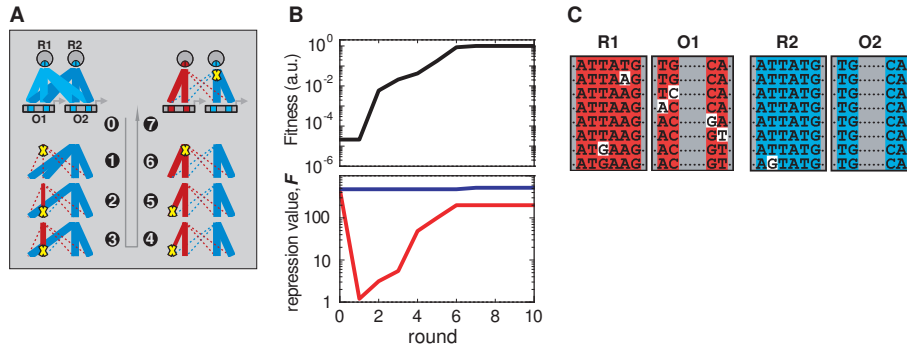


Figure 3.4: Typical divergence pathway: network changes, fitness, and sequence.

(A) Evolving interaction network, where line thickness denotes binding strength between repressor monomer and operator-half. Dotted lines denote negligible repression. Yellow crosses indicate repressor and operator mutations, which are positioned at the top and bottom of the interaction lines respectively.

(B) Fitness trajectory (black) and corresponding repression of each repressor on its operator (red and blue). Fitness is normalized to the maximum value ($\sim 1 \cdot 10^{10}$).

(C) Sequences for each round. Mutated positions are colored white.

3.3 Discussion

Duplication and coevolutionary divergence

We obtain a first view on a fitness landscape for regulatory divergence that is based on actual measured data. We show that the landscape allows evolutionary paths toward independent repressor-operator interactions, exhibiting a step-by-step increasing fitness, starting as early as the first or second mutation. Since the possibility of following such paths critically depends on molecular properties, the use of empirical data is essential for such claims. One could also have imagined fitness landscapes where paths to diverged networks do not exist, or where they are very long, involving large detours. Our results contrast with the notion that a number of neutral or even deleterious mutations have to accumulate before a new function can develop (see for a discussion e.g., [78]). Having beneficial mutations available early on is important, since it greatly enhances divergence probabilities [40]. A lack of early selection would result in much higher probabilities of silencing of one of the duplicates by the accumulation of mutations [64, 72].

While the presented systematic search for optimal pathways is useful in revealing necessary conditions for divergence, one may wonder whether paths are not very different in a probability-based fixation process that typifies natural evolution. However, we found that population genetics simulations reveal the same pathway characteristics: a significant fraction of paths are successful with monotonous fitness increases, one repression dip early on, and few neutral mutations are present (see Fig. 3.6 and page 37 of section 3.5).

The coevolutionary search for a new and independent recognition, which is

one can readily imagine fitness landscapes where the optima are very difficult to reach. Upon first inspection, the measured landscape we consider indeed contains many potential frustration sources: over 99% of all optima in the landscape are in fact below our divergence criterion. Such local optima represent traps in which the system gets permanently stuck once it encounters one. However, the results show that the system is still guided in the right direction to (near) global optima, which indicates that the fitness landscape contains funnel-like features. Moreover, the optimal paths contain negligible detours (Fig. 3.3A) and lead to the nearest optimum (Fig. 3.3B), showing that the funneling is efficient and not constrained by ruggedness. A funnel-like organization of the landscape is also supported by the monotonous fitness increases of the probabilistic pathways (Fig. 3.6C), as well as by the smooth fitness decrease when stepping away from a global optimum (Fig. 3.3C).

The underlying causes for funnels in the fitness landscape may be found at two levels. The first level is that of a single repressor-operator interaction. The surface smoothness that is needed for the funnels may be partly understood from the reported additive contributions of the *lac* amino acids to the binding energy. In mathematical models, additive interactions have been shown to yield smoother fitness surfaces because they can be optimized independently [35].

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

The evolutionary fate of redundant genes has previously been studied primarily using sequence analysis [64, 82]. By using a different dataset and approach, our simulations strengthen recent evidence for a more rapid fixation of mutations in redundant genes [82] (termed "accelerated evolution"). Our analysis enables a next step in our understanding of this important process: It provides a mechanistic rationale for why such a rapid divergence can indeed occur, in terms of minimal selective conditions bacteria must experience, in combination with independently measured plasticity of protein-DNA interactions. Furthermore it yields a quantitative prediction for the minimum number of essential mutations to achieve divergence.

Suggested experiments

Our results show divergence to be possible with monotonic increasing fitness, which hints at the feasibility of monitoring similar processes in experiments. It has recently been shown that the serial dilution assay, as pioneered by Lenski and coworkers [61], can be employed to adapt bacterial strains to a new condition within weeks [10, 83]. Similarly, one could attempt to evolve a duplicate *lac* repressor/operator copy towards the independent regulation of a second operon. However, this more complex assay does require key modifications: (1) growth and selection of the mutants should occur in alternating media, in analogy to our discussion of multiple input conditions, and (2) a starting network must be engineered that satisfies the conditions for DNA-binding divergence: a duplicate repressor/operator and a selective pressure for tight and independent binding.

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

Other network growth scenarios

For biological regulatory networks to grow, not only new components are required, but also new and independent interactions. Next to the coevolutionary duplication-divergence scenario for network growth, alternative models for the creation of new regulatory interactions have been proposed. In the first alternative, a new operator must emerge upstream of the regulated gene in an effectively random DNA sequence [84]. This scenario has mainly been considered for eukaryotes with large upstream regulatory regions and short binding sites. For longer operators in prokaryotes, this scenario requires many neutral mutations before improvements can be selected for (see page 39 of section 3.5), which represents a major evolutionary obstacle.

Another possible source for new regulatory interactions is lateral gene transfer, which is thought to be the source of many paralogs found in prokaryotes [85]. In this scenario divergence would occur while two genes each reside in different organismal lineages (essentially being orthologs at that stage) and each experiencing different selective constraints. Lateral gene transfer unites the diverged genes, re-

sulting in immediate contributions to fitness by both homologous genes. Although examples of this scenario have been found for enzymes [86], transcription factor-operator interactions are a special case, as there is no obvious internal or external selection pressure for their interaction to diverge by itself. Our results illustrate the feasibility of coevolutionary divergence of two transcription factors within a single organismal lineage. These findings are supported by the lack of evidence for horizontal transfer of the *lac* system in *E. coli* [87]. However, this is not to say that lateral gene transfer and duplication-divergence are mutually exclusive. Summarizing, the coevolutionary divergence studied here differs from alternative models of network growth by providing both a high probability of selective advantageous point mutations and a rationale for a divergence pressure.

Finally, it is of interest to consider different selective pressures within the same duplication scenario. While the pressure for independent regulation seems to be a dominant one, as evidenced by the many independent transcription factors that are paralogs, duplications also have yielded other network motifs. An interesting example is the UxuR/ExuR pair of repressors. Like the case studied in the present work, they have originated by duplication and share two operators. However, they seem to have diverged under a different selective pressure, since their cross interaction was not eliminated, but instead has been retained, forming a so-called *bi-fan* motif [71].

This work describes how regulatory network connections can be formed and broken after a duplication event. Our quantitative approach takes the selective conditions and molecular adaptability explicitly into account, and opens up a new angle on the duplication-divergence question that is complementary to existing approaches. Evolution of network connections is treated more abstractly in numerical studies of biological network growth, which have recently received much attention [69, 88, 89]. The use of experimental data will help to perform such studies on a more realistic footing. Finally, the promising new field of experimental network engineering [90–92] and evolution (see e.g., [93]) will also benefit from the quantification of network adaptability.

3.4 Materials and methods

Mutational dataset

In this work we used an extensive dataset of binding affinities of *lac* repressor and operator mutants, obtained by B. Müller-Hill and coworkers. In these experiments, repression values $F_{O_i R_j}$ have been determined in vivo as the ratio of the unrepresed and repressed expression levels of a β -galactosidase (*lacZ*) reporter gene, controlled by a mutant *lac* operator O_i and mutant *lac* repressor R_j . This was done using the standard assay by Miller [94]. Since the β -galactosidase synthesis is proportional to the fraction of free operator (see e.g., [95]), we find for the repression value $F_{O_i R_j} = 1 + [R_j]/K_D$, where K_D is the equilibrium dissociation constant and $[R_j]$ is the concentration of active repressor R_j . The dataset contains repression values of base pair substitutions leading to changes in amino acid residues 1 and 2 of the recognition helix of the *lac* repressor (Y17 and Q18) and base pairs 4 and 5 of the symmetric *lac*

operator [96]. These residues and base pairs were found to be most important for altering repressor operator-binding affinities [14]. The dataset covers a considerable fraction of all possible substitutions involving a homodimeric repressor and a symmetric operator (1,286 out of a total of 6,400). Part of this raw data is published in Lehming *et al.* [14]; the full dataset is found in [74]. The contributions of the two repressor amino acids to the repression value were found to be additive. With this knowledge, repression values could convincingly be assigned to all mutants, including those that were not measured [14]. In the present study we use these assigned repression values, all of which are given in [14]. Moreover, we extend the dataset to include heterodimeric repressors and non-palindromic operators (see below), to obtain the complete mapping between sequence and repression values for all possible mutants ($1 \cdot 10^7$) in the key repressor residues and operator base pairs.

Repression values of heterodimers and non-palindromic operators

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

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

With this equation, repression values involving non-palindromic operators are also automatically taken into account: each dimer-operator interaction is built up additively [75] from two interactions between a monomer and an operator-half. In this derivation the dimerization free energy was assumed to be fixed, since it does not directly affect the specificity by which the repressors recognize their operators. The heterodimer repression value then becomes independent of the dimerization energy.

Optimal pathway simulations

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

In order to determine the optimal mutational pathways in the fitness landscape, an evolutionary algorithm was employed. Beginning with one of the starting sequences, each round we generated all mutants that differ by one base pair (60 in total). Of each mutant network, the strength of all eight possible interactions was determined (see Fig. 3.1B where four possible interactions are schematically shown between the repressor dimers and one of the two operators). Interactions between repressor homodimers and palindromic operators were directly assigned from the published repression values [14]. Other interactions were calculated from the measured data as described above. Next, we selected the best L networks to the next round based on a fitness parameter that is directly calculated from the interaction strengths (see equation 3.1). The next round started by again generating all single base pair mutants of the L selected networks. The effect of L was assessed by varying it between 1 and 105. Decreasing fitness steps were not allowed, and in case of equal fitness, parents were ranked above their offspring. These rules make divergence harder because they constrain the space that can be explored. The evolutionary cycle was repeated until the fitness could not be further improved. Pathways were considered to be successful when the fitness came within a factor 10 of the highest fitness in the landscape.

3.5 Appendix

Simulation of mutational pathways incorporating probabilistic population dynamics

Here results are presented of a second simulation method, where mutations are fixed with a probability that is based on the associated fitness increase (see methods below). Compared to the optimal pathway simulations, this probabilistic approach does not search the landscape as systematically, but it is arguably closer to natural evolution, in that the fixation chance of mutations with no or lower fitness increases is more well-defined [40].

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

Our probabilistic model allows us to vary the amount of drift present in our pathways by varying both the effective size of a population (N) and the growth advantage that diverged networks have over undiverged networks (s_{\max}). Taking a conservative growth advantage of 5% [10] we simulated probabilistic pathways for population sizes ranging from 10^3 to 10^8 . At high population sizes pathway char-

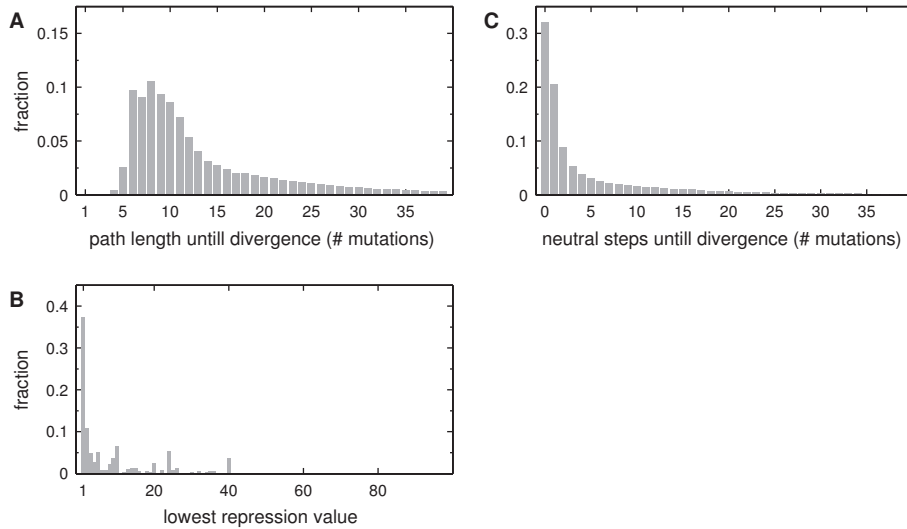


Figure 3.6: Qualitative features of successfully diverging paths in the probabilistic pathway simulations.

Simulations were performed with a 5% growth advantage of a diverged network over the initial duplicate network, and a population size of 10^5 . Of all traced paths, 17% successfully diverged, despite the strict requirements that promote trapping in local optima (fitness cannot decrease). Relaxing these conditions would lead to larger divergence probabilities.

(A) Histogram showing the number of base mutations until divergence for the successful pathways.

(B) Histogram showing the lowest repression values of each repressor on its operator during the successful divergence pathways.

(C) Histogram showing the number of neutral mutations that occur until the pathways successfully diverged.

acteristics remain very similar. Only at population sizes below 10^4 we start to see a strong effect of genetic drift: more neutral mutations, and hence longer pathways. However, the fraction of diverged pathways and the reduction in repression of one of the repressors does not significantly change.

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

Probabilistic pathway simulations

To model the effect of genetic drift in the evolutionary pathways, probabilistic simulations were performed. Thousand such pathways were traced for each of the starting sequences. In each simulation step 60 different single base pair substitutions are possible (the 20 base pairs can each mutate to 3 different bases), which were assumed to occur with equal probability. The fixation probability of each specific mutation depends on its associated fitness increase, and was calculated with a standard and simple population genetics model [40, 76] (and see below). In each simulation step, the fixation probabilities of all 60 possible single base pair substitutions were calculated, and one substitution was randomly chosen according to its share in the total probability. Each path was continued until a (possibly local) optimum was reached, so that the fitness could not be further improved. The purpose of this Monte Carlo-like scheme was to check whether biased random walks show similar features as the ones generated by our optimal pathway simulations.

Mutations that decrease the network fitness were assumed not to be able to fix, while those that keep the fitness constant have a fixation probability of $1/N$, where N is the effective population size. Mutations that do increase the fitness have a fixation chance of $2\Delta s$, where Δs is the selective advantage that this fitter mutant has over its parent. In our simulations we let an increase in the fitness parameter by a factor of 10 correspond to a Δs of 1%. In this way, a successfully diverged network has a selective advantage of 5% (s_{\max}) over the initial duplicated state (fitness rises from 10^{-6} to 10^{-1} , see main text), which matches typical experimentally observed growth advantages [8, 10].

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

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

Here we consider an alternative mechanism for creating a new regulatory interaction, where a new operator must emerge in an effectively random DNA sequence. In a typical prokaryotic case, like e.g. the *lac* system, a 20 base pair operator has to emerge somewhere in a roughly 100 base pair region in order to effectively block RNA polymerase binding. Then there are 10 expected prior matching base pairs for the best binding site within the promoter region. However, experimental data suggests that at least 15 base pairs need to match before appreciable binding is achieved [97, 98], from which point further mutations can be positively selected for. This means that more than 5 base pairs need to be optimized without selection, while the coevolutionary pathways can be selected for almost immediately.

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 multi-peaked, harboring in total 19 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.

4.1 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 [99]. Ever since Wright [33] introduced the metaphor of an adaptive landscape, its shape has been hotly debated, but nonetheless essentially remained unknown due to insufficient empirical data [99] [38, 49, 100–107].

The architecture of adaptive landscapes is tightly related to the notion of epistasis (see Fig. 4.1) [33, 108]. 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 (see Fig. 4.1A, top). This results in adaptive landscapes that are smooth and single peaked (see Fig. 4.1B, left). In the case of sign epistasis, the sign of the fitness or phenotypic effect of a mutation does depend on the genetic background (see Fig. 4.1A, 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 (see Fig. 4.1A, bottom right). It has been suggested that reciprocal sign epistatic interactions play a central role to generate adaptive landscapes with multiple distinct peaks (See Fig. 4.1B) [105]. The occurrence of multiple peaks can give rise

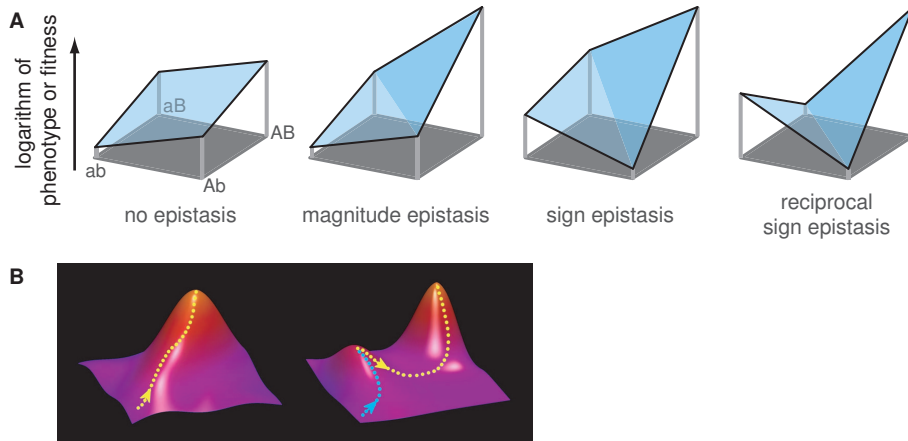


Figure 4.1: Relationship between epistasis and landscape ruggedness.

(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. As a disclaimer, note that several features of these surfaces do not correspond to 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.

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 [41, 49, 105, 106]. 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 [14], we have previously reconstructed this landscape to investigate the divergence between two repressor-operator pairs [39]. 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 [41, 49], 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.

4.2 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 [109] discovered it. Transcription regulation of this operon by binding of the *lac* repressor (LacI) to its operator regions in the *lac* promoter (see Fig. 4.2) is understood in great detail and continues to be of great value in the study of gene regulation [14, 110–119]. LacI is a prototypic member of the large GalR-LacI family of prokaryotic transcription factors, a group that has more than 1000 members [12, 13]. 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 [13].

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 [117], 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 [96], 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 (see Fig. 4.2A). 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 [14]. Repression values for repressor-operator pair variants were determined as the ratio of repressed and unrepressed expression of a downstream β -galactosidase (*lacZ*) reporter gene, as measured via a standard Miller assay [94]. 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) [14, 74, 94]. 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.

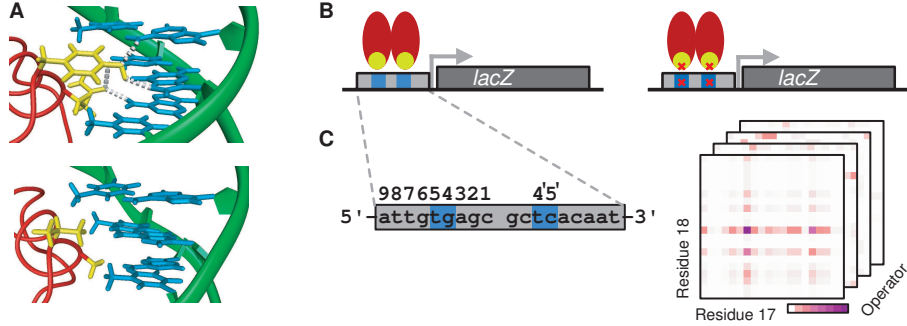


Figure 4.2: Description of the studied system.

(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.

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 [14, 74]. 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 [39, 75]. Briefly, extrapolated repression values are calculated according to equation 3.2,

$$F_{0_1 0_2} = 1 + \sqrt{(F_{0_1} - 1)(F_{0_2} - 1)}$$

For a palindromic operator, $F_{0_1 0_2} = F_{0_1} = F_{0_2}$, where $F_{0_{(1/2)}}$ is the product of two factors (one for each of the key residues) taken from Table II of Ref. [14]. F_{0_1} and F_{0_2} 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 (see Fig. 4.2).

4.3 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.

4.4 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 106 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 [117], see Fig. 4.2). 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. 4.2), 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 4.1 lists the genotypes and associated repression values of the 19 peaks of the landscape.

In order 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 4.1.

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 4.1). These two peaks have repression values of 200 and 325, respectively. The peak tgcaSQ is the optimum that is closest to the wild-type 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 4.3 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

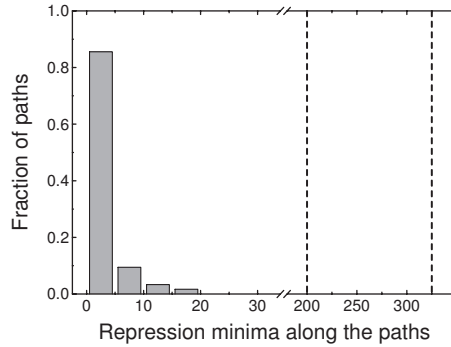
Peak rank	Genotype		Repression level
	Operator	Repressor	
1	tg··ca	SM	520
2	tt··aa	HM	500
3	tg··ca	SQ	325
4	gt··ac	KS; KT; KM*	300
5	aa··tt	IS	300
6	aa··tt	SS	225
7	ta··ta	SG; IG; PG**	220
8	tg··ac/gt··ca	KM	219
9	ac··gt	PK	200
10	ac··gt	MK	200
11	aa··ac/gt··tt	SS	184
12	at··at	VM	160
13	ag··ac/gt··ct	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

Table 4.1: Genotypes and repression values for the 19 independent peaks in the phenotype landscape of *lac* operator-repressor pair variants.

Non-palindromic operators are indicated by their two equivalent reverse-complement sequences. *: S, T, and M are connected, so 1 optimum. **: S, I, and P are connected, so 1 optimum.

Figure 4.3: 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).



system has to overcome a loss of at least tenfold in repression.

A number of typical paths are illustrated in Figure 4.4, 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. 4.4B, 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. 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 most likely path between optima acgtPK and tgcaSQ is path 3 depicted in Figure 4.4, 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

Multiple peaks and reciprocal sign epistasis in an empirical landscape

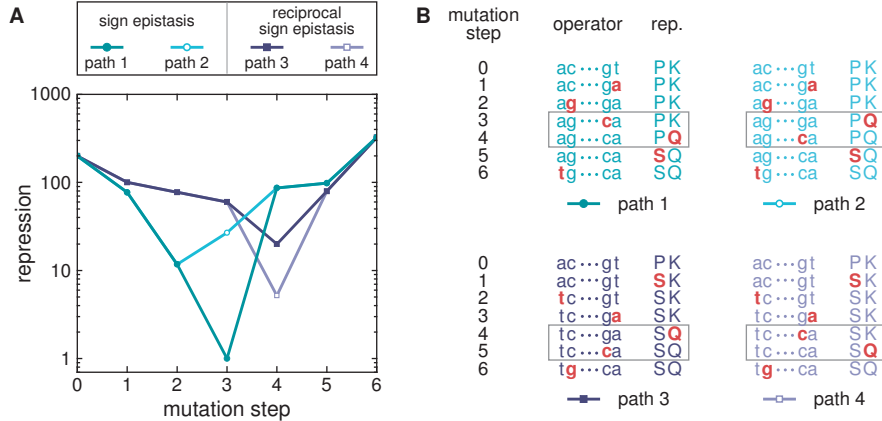


Figure 4.4: Examples of direct evolutionary paths between peaks acgtPK and tgcaSQ. (A) Repression values of the intermediate mutants along each path. (B) Intermediate sequences along the mutational paths. Mutations are shown in bold red. Steps exhibiting epistatic interactions are outlined.

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 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 (see Fig. 4.4B, 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 (see Fig. 4.4A).

Reciprocal sign epistasis occurs when two mutations are individually deleterious but jointly result in a positive effect (see Fig. 4.1A). Such a situation captures, at the level of individual mutation steps, the constraints created by a multi-peaked

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 interaction stands at key locations of a multi-peaked 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 [120].

From Table 4.1, 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), there 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.

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. [121] 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 Figure 4.4, 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 Figure 4.4, 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 4.1). 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 4.1). 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.

4.5 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 [49] or at the genotype-fitness level [41], thereby showing that paths can be selectively inaccessible (see Fig. 4.1B). Nevertheless, some paths to the global optimum remained selectively accessible, indicating that the landscapes were single peaked [41, 49]. 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 princi-

ple 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 [105]. Our study shows that such a master key does not exist for the *lac* repressor-operator system.

A multi-peaked 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 multi-peaked landscapes, which can also be theoretically addressed in a more rigorous manner [120], is supported by our analysis: as predicted, such interactions appeared present along paths exhibiting the highest minimum.

It has frequently been recognized that a multi-peaked 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 [15], 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 [39]. 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 [122].

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.

Simple rules underlie an empirically determined genotype-phenotype landscape



Deciphering the architecture of genetic interactions is central to understanding the constraints of evolution. Most studies have concentrated on two-way genetic interactions, or epistasis, even though higher-order genetic interactions may give rise to severe constraints such as sub-optimal fitness peaks. It thus remains unclear whether evolution depends on specific and detailed information enclosed within genotype-phenotype landscapes or rather on more generic landscape properties. Here we addressed this question for the recognition between the E. coli lac repressor and its operator, using systematic experimental data on the repression achieved by higher-order mutants. To perturb existing genetic interactions in an unbiased manner, we randomly permuted the repression values between certain genotypes, and then monitored the effect on evolving a new recognition by tracking the fate of mutational trajectories within the landscape. The analysis showed that the success rate of evolving a new recognition was strongly increased by permutations between repressor genotypes, in accordance with the known absence of genetic interactions between repressor residues. However, the success rate was almost unchanged by other permutations, such as those between operator genotypes. This observed robustness against randomization suggested that the evolutionary constraints may be similar to a random landscape that accounts for the absence of genetic interactions between repressor residues. We indeed observed that such random landscapes displayed strikingly similar evolutionary dynamics. This study suggests that genetic constraints in biological systems are captured by generic rules for the underlying genetic interactions.

In living systems, the properties of one component typically affect the functioning of many other components and of the system as a whole. Complex interdependencies between functional parts are observed throughout all levels of biological organization, from the nucleic acid interactions that give rise to secondary and tertiary structure within RNA molecules [123], to the specific recognition between proteins that underlie intricate regulatory responses [124], and the functional relations in the activity of different metabolic modules [125]. This integrated architecture of living systems is believed to have important evolutionary consequences, as it can lead to higher-order interactions between multiple genetic changes (also termed genetic interactions): when challenged with a new environment, a genetic change in one locus may be beneficial for the system as a whole, but conditional on corresponding genetic changes in many other loci.

The evolutionary consequences of genetic interactions at the most elementary level, namely between two genetic changes, have since long been studied theoretically [38, 108, 126], and more recently also experimentally [39, 41, 49]. An absence of interactions means that the effect of a mutation in one element does not depend on the occurrence of the other. Such additively contributing mutations have been observed for IMDH catalytic activity [49]. A presence of genetic interactions has been seen in for instance the gene β -lactamase, where a mutation in one residue was found to lead to a large increase in resistance to a new antibiotic, but only if another residue was substituted first [41]. Otherwise the mutation produced a decrease in resistance. These interactions between mutations constrain evolution, in the sense that they limit the number of evolutionary trajectories that are selectively accessible when taking single-mutation steps.

However, it is much less clear whether higher order genetic interactions — involving more than two genetic changes — constrains the evolution of biological systems. For instance, the presence of ‘ridges’ in the genotype-fitness landscape may lead evolution away from or towards the global optimum, while sub-optimal peaks can lead to entrapment and evolutionary stasis. In the latter case it is not merely the presence of sub-optima that is determining, but rather their precise position: far away from the evolutionary path they will be irrelevant, but nearby they can act as attractor. Both these features are properties of the larger landscape and necessarily involve interactions between multiple mutations. For instance, while two mutations that are jointly beneficial but separately deleterious (reciprocal sign epistasis) may seem to offer two peaks, this will only be true if the other mutations that are possible do not bridge these peaks in a manner that reduces them to one peak. Indeed, we have shown theoretically that the necessary conditions for multiple peaks cannot be defined in terms of local two-way genetic interactions [120]. Thus, the outcome of evolution may be decided by landscape features that involve more complex higher order genetic interactions. Whether this indeed the case for actual biological systems is an urgent question given the recent advances in developing quantitative evolutionary approaches [39, 41, 49], as it will determine whether evolutionary dynamics can be captured by a limited number of parameters or instead critically depends on intricate details of the relationship between genotype, phenotype, and fitness.

Here we address this issue using the *Escherichia coli lac* operon as a model sys-

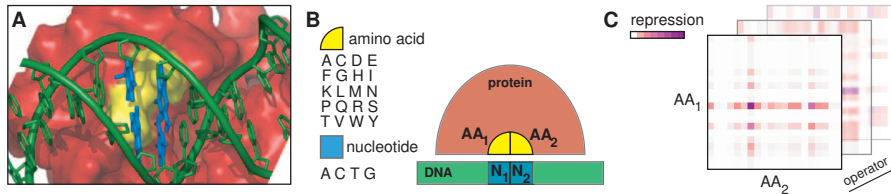


Figure 5.1: *In vivo* measured repression landscape

(A) Molecular structure of the *E. coli lac* repressor interacting with its operator [46]. The molecular elements responsible for binding specificity are highlighted. Image created with PyMOL [127].

(B) Diagram of the binding interaction, indicating the two amino acids in the *lac* protein and the two nucleotides in the operator on the DNA. Each amino acid residue can take twenty forms, and each nucleotide four.

(C) Representation of the *in vivo* measured repression landscape. By constructing mutant forms with combinations of all four components, Lehming et al. obtained a complete landscape [14]. Note that in reality the landscape is multidimensional and this projection provides a limited representation.

tem. Repression of the *E. coli lac* operon by the transcription factor LacI is one of the best studied biomolecular functions [7, 128]. In the absence of lactose, LacI binding to the *lac* promoter blocks expression of the *lac* operon, thus alleviating the cell from spurious and burdensome [10] production of the proteins that allow lactose consumption. Lac repressor-operator binding is well-suited for evolutionary studies, as the effects of mutations in the system has been studied extensively [19, 117, 129, 130]. While many base pairs affect binding, only six in the repressor and four in the palindromic operator contribute to the specificity of the interaction (Fig. 5.1A and B). These key base pairs have been systematically mutated, and their effects on repression have been quantified (Fig. 5.1C). We have previously used this data to map the repression values for all combinations of these ten base pairs, by making use of the observed simplifying additivity between the two residues and between the two monomers that bind as a dimer in these experiments [14, 39]. This genotype-phenotype landscape exhibits reciprocal sign epistasis and multiple peaks [131], and has been used to describe the evolution of new repressor-operator interactions [39].

In order to study genetic interactions within the *lac* repression landscape in an unbiased manner, we follow a statistical approach. The central feature of the approach is that existing interactions are altered by shuffling the repression values within the landscape between genotypes. Next, the consequences on the success of evolving a new repressor-operator interaction are quantified by simulating evolutionary trajectories from many different starting points within the landscape and monitoring the rate of success. In this way, we identify these genetic interactions that affect the potential to evolve, rather than attempt to quantify specific interactions or features throughout the landscape, which may not be relevant to this evolutionary transition. Secondly, the process of shuffling retains the repression

values based on experimental results, and thus preserves the distribution of individual repression values.

Using this approach, we found that additivity between the amino acids in the *lac* recognition helix is critical to the success of evolving new repressor-operator interactions. Without it, the fraction of trajectories that become trapped in sub-optimal peaks increases significantly. Other tested interactions, such as the one between the operator base pairs, were found to have only a marginal effect on the success rate. This all-or-nothing outcome suggests that a set of simple rules govern the potential to evolve a new repressor operator interaction: additivity between repressor residues, without any other specific interactions. To test this idea, we defined new landscapes based on only these rules and randomly generated numbers for the repression values. The outcome of the same evolutionary process appeared to be the same as for the empirically determined landscapes, thus confirming that these simple rules are sufficient and that the empirical landscape does not contain other hidden specific interactions that are critical to the evolutionary outcome. These results open the way for simulation of network evolution on a realistic footing and is a first step in bringing together the world of theoretical and experimental landscapes.

5.1 Results and Discussion

Specificity in *lac* repression landscape

Evolution of new and specific regulatory interactions is an essential part of adaptation. In the case of the LacI protein, historic occurrences of divergence are evident from the large family of regulator proteins it belongs to, all originating from a common ancestor [15]. Members of this protein family diverged a long time ago, but still share DNA-sequence similarities and conserved protein domains for sensing, oligomerization and binding [73]. While the general function of the binding domain remained the same (i.e. binding to the DNA), most proteins in the LacI family have diverged to recognize a different sequence on the DNA. *E. coli* contains at least ten other proteins from this family, including LacI, GalR, EbgR, MalI and CytR [12, 13]. Interestingly, when the key residues of these proteins and the key base pairs of their operator are transplanted into the *lac* system, strong binding is retained [14]. Importantly, these mutant *lac* repressors do not bind strongly to the original *lac* operator, nor does the original *lac* repressor to their operator. This suggests that regulatory proteins of the *lac* family diverged under an evolutionary pressure to limit cross-talk in repressor-operator binding [14, 39]. By considering a penalty for such cross interactions, the evolution of new and specific interactions in the mutational repression landscape can be simulated [39].

Evolutionary trajectories in wild-type landscape

We quantified the capacity to evolve a new *lac* interaction by computing evolutionary trajectories in the genotype-phenotype landscape based on measured repression values, as described previously [39]. Briefly, the trajectories start at one of

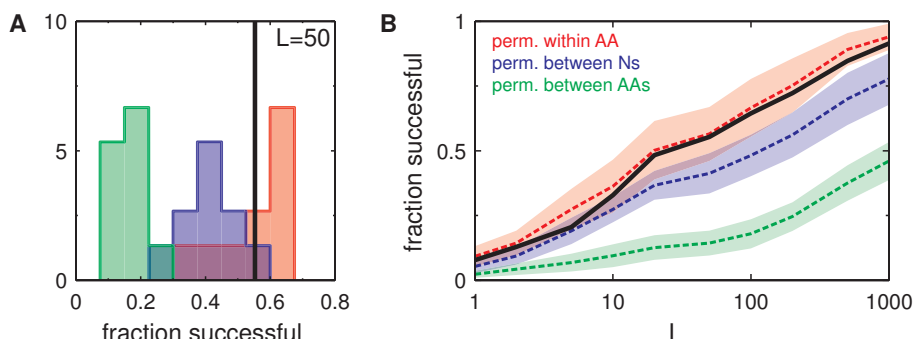


Figure 5.2: Evolution on wild-type and permutated landscapes

(A) Fraction of successful divergence paths in simulations for natural (bold vertical line) and permutated landscapes (histograms). Data is shown for simulations where 50 mutant networks are carried to the next round (L). Permutations shown are for repression contributions within amino acid residues (red), between amino acid residues (green) and between base pair residues (blue).

(B) Average fraction of successful divergence paths as a function of L . Colored areas around the average line indicate standard deviation of the data. Colors are as in panel A.

the 132 repressor-operator combinations throughout the landscape that provide a repression value of over 100, with a population of L identical members. This population is expanded with all single point mutation neighbors, which have either a base-pair substitution in the operator or in the repressor. Their performance in developing a new interaction is scored by a measure that quantifies both their repression value and a penalty for binding to the original operator or repressor. The best L variants are taken to the next round, which completes the cycle. This procedure is continued until the trajectories are trapped on a (sub)optimum. To compare the success rate for evolving in different (shuffled) landscapes, we use an arbitrary but consistent cutoff for their performance (a factor of 10 within the highest possible performance). As a control, we also simulated trajectories using a probability based method [40, 76], in which a single variant is followed and the subsequent point mutations is randomly chosen based on the increase in performance. The two methods were found to provide similar conclusions.

The success of evolving a new *lac* interaction based on the measured landscape and using the above described method, is given in Fig. 5.2. As reported previously [39], for a population of $L=50$, a fraction of about 0.55 of all starting points led to a new *lac* interaction (Fig. 5.2A). Towards lower L (Fig. 5.2B), this success rate decreases as expected because the reduced genetic diversity in the smaller population means that fewer routes can be explored, leading to a higher chance to become trapped on a suboptimum. At $L=1$, where the procedure corresponds to a steepest-ascent hill climber [132], the fraction reduces to about 0.08. Towards large L the successful fraction approaches unity, indicating that with unlimited diversity in the population successful pathways always exist. These results for the original landscape will serve as a reference for the success rates of permutated landscapes.

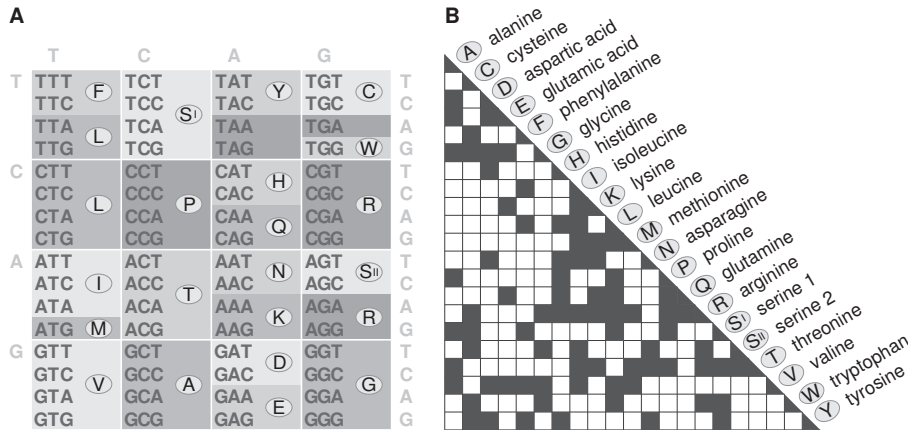


Figure 5.3: The genetic code and the amino acid point mutation connectivity

(A) The amino acid codon translation table indicates which triplet codon codes for which of the twenty amino acids. Three codons do not code for amino acids (TAA, TAG and TGA), but these 'stop codons' signal a termination of translation resulting in the end of a protein. The codons for all amino acids are intra-connected, except for serine, which has two separate sets of intra-connected codons, indicated by S_I and S_{II}.

(B) Due to the triplet code, single base substitutions can only alter amino acids to some a subset of all other amino acids. The resulting point mutation connectivity is shown here as a connectivity matrix.

Interactions at one residue position

Amino acids are encoded in the DNA by means of triplet codons (three subsequent nucleotides). As a result, an encoded amino acid cannot mutate into all of the other 19 amino acids by means of a single base pair substitution. Amino acids thus form a connected network, in which some pairs are directly connected while others are not (see Fig. 5.3). On average each amino acid is connected to eight other amino acids. The multiple redundant triplet codes of each amino acid are typically all connected, except serine (S), which is encoded by two groups of triplet codes (S_I and S_{II}) that are not connected by single base pair substitution.

The effect of the connectivity network on evolution depends on how each amino acid contributes to phenotype or fitness. For example, at a certain position in a protein, valine (V) may provide a higher contribution to fitness than tyrosine (Y), and thus be favored by selection. But these amino acids are not connected, and the Y to V change thus requires passage through an intermediate amino acid (e.g. aspartic acid (D) or phenylalanine (F)). If the contribution of these intermediates were *lower* than Y, then the trajectory would not be selective accessible by single base pair substitutions. In contrast, if their contribution were *higher* than Y (and lower than V), then the path would be accessible. Thus, such possible correlations between the contributions of amino acids to repression at one residue position may critically affect evolutionary trajectories. Note that this would not be the case if all

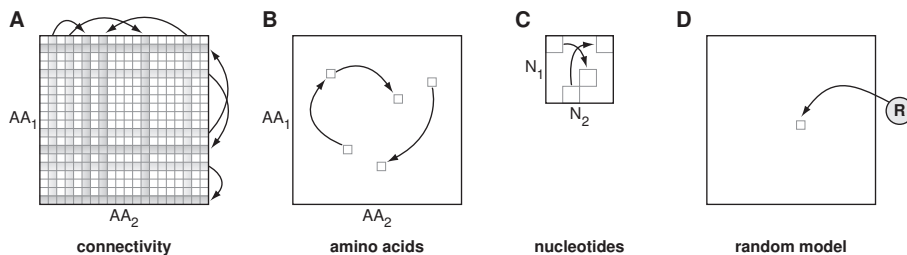


Figure 5.4: Landscape permutation and random landscape construction

(A) Due to the amino acid triplet code, a single mutational step cannot change an amino acid into all other possible amino acids. Correlations in the binding landscape due to this limited connectivity between amino acids were randomized by permutation of their data values. Note that such permutations would have no effect in the operator, as any nucleotide can mutate to any other.

(B) Correlations due to additivity between the two amino acids were randomized by permutation of the data values in the amino acid plane.

(C) Correlations due to additivity between the two nucleotides were randomized by permutation of the data values in the nucleotide plane.

(D) Random landscapes were created with a simple statistical model. Each data value was randomly created based on the rules described in the main text.

pairs of amino acids would be directly connected, as the optimal amino acid would always be directly selectively accessible.

To investigate whether the evolution of *lac* recognition is affected by correlations between amino acids in their contribution to repression, we randomly shuffled the values of these contributions between amino acids (Fig. 5.4A). This is straightforward because for the *lac* repressor, the contribution of each of the two amino acids is independent. For instance, consider the first key residue (AA₁) that is a tyrosine for wild-type (*WT*) LacI. We can take the value of its contribution to binding the *WT* operator, which has an alanine as the first key base pair (N₁) and a guanine as the second (N₂), and attribute it instead to the repression contribution that valine would provide as the first key residue to the same operator. The same tyrosine to valine attribution is then performed for the values of all operators. By permuting all amino acid contributions randomly a different repression landscape is obtained, which can be tested for the success rate of evolving a new interaction. Multiple permuted landscapes are obtained, corresponding to different random realizations of the permutations.

The success rates of evolving a new interaction after these permutations are indicated by the red data in Fig. 5.2A for a population of 50 ($L=50$). Some permuted landscapes showed a higher, and some a lower success rate, with their average (0.56 ± 0.10) not significantly different from the wild-type success rate (0.55). This correspondence between wild-type and the average of permuted landscapes was consistent over the full range of population sizes L (Fig. 5.2B). These results indicate that correlations between the amino acids in their contribution to repression do exist, and do affect the success rate. However, in the wild-type landscape, these

correlations cannot be distinguished from the correlations resulting from randomly attributed repression contributions, and are not biased to result in a higher capability to evolve a new interaction.

Interactions between two residue positions

A central conclusion from the mutant analysis of Müller-Hill and coworkers, is that the two residues that are key to specificity interact independently with the operator [14], such that their contribution to repression is additive. This independence means that there are no genetic interactions between the two key residue positions, which is also visualized in Fig. 5.1C by the horizontal and vertical lines at certain amino acids that provide a high contribution to repression irrespective of the amino acid at the other residue position. In order to address the importance of this independence, we randomly permuted the values for the amino acid contributions within the AA_1 - AA_2 plane (Fig. 5.4B). This process of reshuffling breaks up the horizontal lines and thus introduces genetic interactions between AA_1 and AA_2 .

In contrast to the previous section, here we observed significantly altered success rates (green data in Fig. 5.2). In the simulations with a population of 50 ($L=50$), on average a fraction of 0.14 of the starting sequences successfully evolve a new interaction, compared to 0.55 for the wild-type landscape. This lower success rate was observed over all values of L (Fig. 5.2B). Hence, the independent interaction of the two amino acid residues with the two base pairs in the operator significantly enhances the capability to evolve new interactions. These results are consistent with theoretical studies, which have shown that additivity between contributions to fitness facilitates evolution [35]. Components that do not contribute additively to fitness produce epistasis, landscape ruggedness, and the entrapment of evolutionary trajectories in local optima.

Interactions between operator base pair positions

Two adjacent base pairs in each operator half, at position 4 and 5, are responsible for the specificity of binding one repressor monomer (Fig. 5.1A and B). In order to investigate the role of interactions between these two key base pairs, it is insightful to consider a projection of the landscape data that is different than in Fig. 5.1C, where the repression values for a specific operator variant were displayed within the AA_1 - AA_2 plane. Here we took a specific repressor variant, and considered its repression values within the N_1 - N_2 plane. Next, we randomly permuted the repression values within this N_1 - N_2 plane (Fig. 5.4C) in a similar fashion as for the AA_1 - AA_2 plane in the previous section.

Analysis of the corresponding evolutionary trajectories showed that up to $L=10$, the success rate for evolving a new interaction was statistically identical as in the wild-type landscape (Fig. 5.2, blue data). Beyond $L=10$ the success rate started to deviate slightly to lower values. These results indicate that the key base pairs in the *lac* operator interact nearly as strongly as expected for a random landscape. The

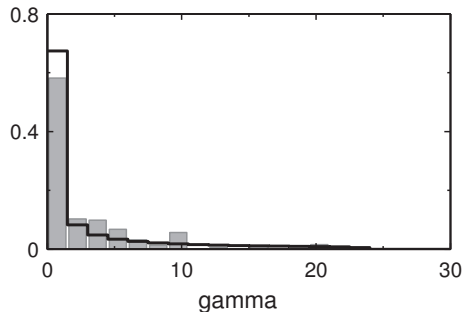


Figure 5.5: Fit of gamma as found in empirical binding landscape

The distribution of gamma (amino acid contributions to repression value) in the natural landscape (grey histogram) is captured with a phenomenological function (black solid line, see main text for details).

small reduction of the success rate could be explained by small degree of additivity between the two key base pairs.

Evolution in randomly generated landscapes

The analysis so far indicates that the lack of genetic interactions between the two residue positions is central to the success of evolving a new interaction. When interactions are introduced randomly by permutations, the rate of success dramatically decreases. These results are in line with the notion that additivity between contributing components facilitates their joint evolutionary optimization [35]. Perhaps more notable is the observation that random permutations in some cases do not substantially affect the success rate. It suggests that while (strong) genetic interactions exist, they are not *specific*. These observations lead us to speculate that the natural *lac* repression landscape may be similar to a random landscape, in terms of its capacity to allow the evolution of new recognitions. To test this idea, we constructed fully randomly generated landscapes, using only the simple rule that the two repressor residues contribute additively to repression, and quantified the success rate for evolving a new interaction in the same manner as before.

In practice, we assigned repression values using a random number generator to genotypes with the same number of base pairs representing the repressor and the operator, and following the additivity rule for residue contributions (see Experimental section on page 64). For a clear comparison with the empirical landscapes, the magnitudes of the randomly generated repression values exhibited the same spectrum as the measured repression values (Fig. 5.5). The success rate of evolving a new recognition was found to be 0.55 ± 0.14 ($L=50$, Fig. 5.6A), which is remarkably similar to the value observed for the measured wild-type landscape (bold vertical line in Fig. 5.6A). Also the dependence of the success rate on L is remarkably similar (Fig. 5.6B). Note that this comparison did not involve any free parameter, or any fitting. As a control, we also investigated the effect of the previously performed permutations (Fig. 5.4). Overall, these data were also similar as observed for the empirical landscape, with a large reduction in success rate when permuting between amino acids (to 0.25 on average, vs 0.14 for the empirical landscape at $L=50$, Fig. 5.7B), and negligible changes when permuting within one residue position. Only the permutations of the values for the operator base pairs showed some

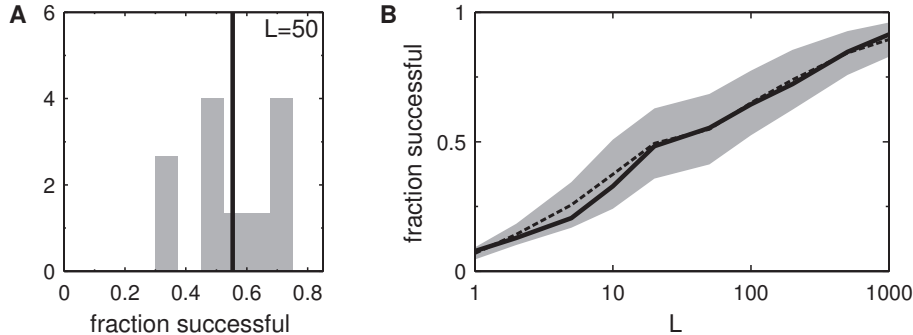


Figure 5.6: Evolution on random landscapes

(A) Fraction of successful divergence paths in simulations for natural (bold vertical line) and random landscapes (grey histogram). Data is shown for simulations where 50 mutant networks are carried to the next round (L).

(B) Average fraction of successful divergence paths as a function of L for natural landscape (bold line) and random landscapes (dotted grey line). Colored area around the average line indicates standard deviation of the data.

difference, with the empirical landscape showing a small shift, while the random landscape did not show a shift.

5.2 Conclusion

Recent experimental investigations of genotype-phenotype relations have provided a novel view on the genetic constraints that limit evolutionary processes. For example, these empirically determined adaptive landscapes have revealed the presence of sign epistasis between two genetic changes [41, 49], and that this results in a preferred sequential order for the fixation of mutations and thus limits the number of adaptive trajectories. Such understanding of how simple landscape features give rise to evolutionary constraints is important given the intractable high-dimensional structure of genotype space. However, it remains unclear whether evolutionary constraints can only be understood by detailing the full genotype-phenotype relation with all its higher order genetic interactions, or instead is determined by more a limited level of information.

Here we sought to address this question with a statistical approach, using the *E. coli lac* operon as a model system. In particular we considered how the *lac* repressor and operator can acquire a novel and specific binding, a key transition that underlies the current *lac* family of transcriptional regulator. Binding specificity between the *lac* repressor and operator is predominantly determined by two repressor residues and four operator base pairs, which gives rise to a genotype-phenotype relation that spans $4^{10} \sim 10^6$ sequences with corresponding binding strength. The central finding of this study is that the success rate to achieve a new and specific recognition is determined by just a few simple rules that contrast with the large possible information content of the full genotype-phenotype landscape: repres-

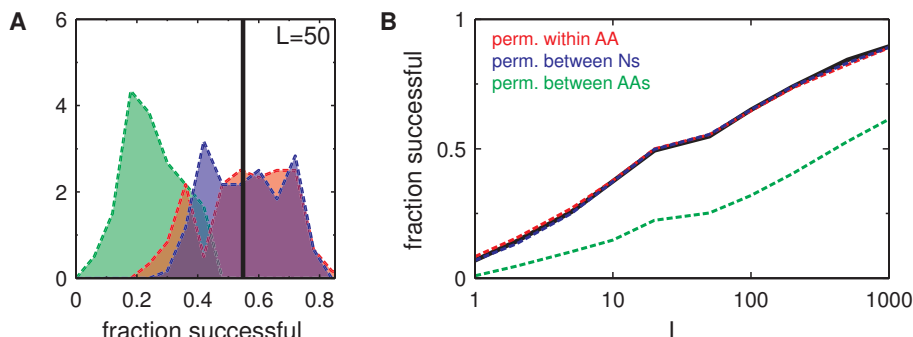


Figure 5.7: Evolution on permuted random landscapes

(A) Fraction of successful divergence paths in simulations for average of the random landscapes (bold vertical line) and permuted random landscapes (histograms). Data is shown for simulations where 50 mutant networks are carried to the next round (L). Permutations shown are for repression contributions within amino acid residues (red), between amino acid residues (green) and between base pair residues (blue).

(B) Average fraction of successful divergence paths as a function of L . Colored areas around the average line indicate standard deviation of the data. Colors are as in panel A.

sion values are drawn randomly from an exponential distribution, while assuring full additivity between repressor residues in their contribution to repression. This shows that there is no specific information beyond randomness and additivity within the large genotype-phenotype relation that is relevant for the evolutionary trajectories of interest here.

The sufficiency of simple rules does not mean that higher order genetic interactions do not exist. Rather, it shows that the higher order genetic interactions that exist in random landscapes are similar to natural ones, in their effect on evolving new and specific interactions.

This work touches upon the question whether biological systems have evolved specific genetic interactions within their architecture that reduce evolutionary constraints. Our analysis does not show evidence for such relations. It is intuitive to understand that random landscapes are multi-peaked, which in fact does indicate higher-order genetic interactions. However, random landscape do not contain specific features that would affect constraint, such as ridges or a particular ordering of peaks with respect to each other, such as for instance sub-optima that are separated from the global optima. Thus, the empirical landscape also does not contain such specific features that affect the evolution of new and specific binding, except for the additivity between repressor residues.

The empirical landscape did not indicate a level of epistasis that is stronger than observed in random landscapes, as this would have led to an increased success rate upon permutation. While one might not intuitively assume such high levels of epistasis in biological systems, it remains a possibility. Here a comparison with man-made systems is insightful, as they can exhibit highly conditional features that can be viewed as constraint. For instance, a digital lock-key recognition is highly

epistatic if one mismatch leads to a complete lack of recognition.

This work provides a step towards understanding the roles of higher-order genetic interaction in evolutionary processes and the architecture of biological systems. It will be intriguing to discover whether other biological functions are governed by specific interactions in their evolutionary history, or whether instead they can similarly be reduced to simple rules, and finally arrive at a more complete understanding of the generic nature of genetic constraints in biological systems. The approach developed here is general and can be applied to analyze other landscapes when they become available.

5.3 Experimental

Mutational dataset and simulations

In this work we use a dataset of repression values of mutant versions of the ideal *lac* operator in combination with mutant LacI repressor proteins, as obtained by Müller-Hill and coworkers [14]. Repression values were determined *in vivo* as the ratio of repressed and unrepressed expression of a downstream β -galactosidase (*lacZ*) reporter gene, as measured via a standard Miller assay [94]. Data values were obtained for all possible combinations of two amino acids in the repressor protein and two base pairs in the operator. Details of this dataset and discussion of its relevance can be found elsewhere [39, 131].

For each landscape (empirical, permuted or random), simulations start with the duplication of one of the repressor-operator pairs having a repression value of 100 or higher. Subsequently, single base pair substitutions are applied and accepted based on their effect on fitness. The fitness of the system is based on the strength of repression, while considering a penalty for cross-interactions (see [39]). We let simulations proceed until they are stuck on an optimum, after which we consider evolution of a new and specific interaction to be successful when the fitness came within a factor 10 of the highest fitness in the penalty landscape. We perform two types of simulations which differ in how mutations are selected and which have both been described in detail previously [39].

Landscape permutations

We create permuted versions of the empirical landscape by first creating an empty 'permuted' landscape. Next we generate random sequences, which are used to copy data values from the empirical landscape into a new location of the permuted landscape. For the 'connectivity' permutation two random sequences of the numbers one to twenty are generated (one for each of the amino acid residues). These sequences indicate the new location of each repression value in the permuted landscape, with respect to an alphabetically ordering of the twenty amino acids. For example, when the first numbers in the two sequences are seven and twelve, the empirical repression values of genotypes containing alanines (A) in both amino acid residues are assigned to genotypes containing histidine (H) in the first amino acid residue and asparagines (N) in the second amino acid residue. This

procedure is performed for each repression value in the empirical landscape, resulting in the previously empty permuted landscape to be completely filled. We arbitrarily chose an alphabetical ordering, but any different ordering would result in the same type of random permutation.

For the ‘amino acid’ permutation a random sequence of the numbers one to 400 is generated. This sequence indicates the new location of each repression value in the arbitrarily numbered amino acid plane (see Fig. 5.4B). Similarly, for the ‘nucleotide’ permutation a random sequence of the numbers one to sixteen is used. Note that the landscape modifications as described above, only move data values, hence the spectrum of repression values in the permuted landscapes stays the same. The shape of the landscapes, however, is altered. This could result in removal or addition of peaks.

Each type of permutation can produce a huge number of possible landscapes. For the connectivity permutation more than 10^{36} ($= 20! * 20!$) different outcomes are possible. The amino acid and nucleotide permutation can result in $400!$ and $16!$ different landscapes, respectively. We sampled a random fraction of the possible permutations: ten landscapes for each type of permutation.

Model for random landscapes

We create random repression landscapes by first creating an empty landscape of the same size and structure as the empirical landscape. Next we assign a repression value to each genotype using a random number generator. We introduce additivity between the two amino acid residues by defining the repression value as the product of two random numbers, where each represents the contribution of one of the amino acids. In order to produce landscapes with a similar spectrum of repression values as the empirical landscape, the random values for the amino acid contributions were taken from a distribution mimicking the empirical data (see Fig. 5.5). We used the following phenomenological function to randomly produce binding contributions of each amino acid:

$$F = F_{\min} + F_{\max}e^{-\text{skew} * u}$$

, where F_{\min} is set to 0.02 and gives a lower bound to the binding contribution, where F_{\max} is set to 23.5 and gives a higher bound to the binding contribution, skew is set to 9.5, and u is random uniform number between 0 and 1. For each random repression landscape 640 random numbers have to be generated: each amino acid residue requires a value for each of the possible operators, and for each of the possible amino acids ($= 2 * 16 * 20$). We constructed and analyzed ten random landscapes.

5.4 Appendix

Probability-based simulations

As a control, we simulated trajectories using a probability-based fixation process that typifies natural evolution [40, 76]. This process assigns a probability to all sin-

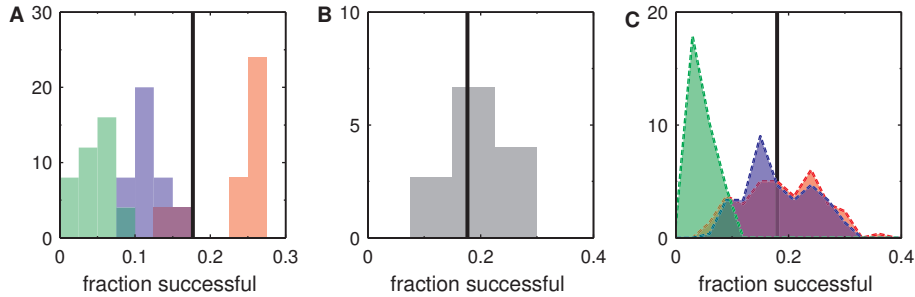


Figure 5.8: Probabilistic pathway simulations

(A) Fraction of successful divergence paths in probabilistic simulations for natural (bold vertical line) and permutated landscapes (histograms). Permutations shown are for repression contributions within amino acid residues (red), between amino acid residues (green) and between base pair residues (blue).

(B) Fraction of successful divergence paths in probabilistic pathway simulations for natural (bold vertical line) and random landscapes (grey histogram).

(C) Fraction of successful divergence paths in probabilistic pathway simulations for average of the random landscapes (bold vertical line) and permutated random landscapes (histograms). Colors are as in panel A.

gle mutation neighbors, based on their fitness increase with respect to the current genotype. During each step in the simulation one of the possible mutations is randomly chosen based on their probability, until no fitness increases are possible. From all starting sequences a thousand of such runs are performed, and the fraction of runs that successfully evolve a new interaction, indicates the success rate of evolution on that particular landscape. The results of these simulations (see Fig. 5.8) were similar to those obtained with the method in the main text.

Peaks in *lac* repression landscape depend on amino acid connectivity

Our simulations showed that repression values can be randomly distributed over the different amino acids without significant effect on the capacity to evolve new interactions in the landscape. Nonetheless, the limited connectivity between amino acids is important for evolution through its effect on the structure of evolutionary landscapes. The effect of limited amino acid connectivity can be illustrated by the number of distinct peaks within the *lac* repression landscape. These maxima can be identified using an algorithm that determines whether points in the landscape with higher repression are accessible without drops in repression [131]. The natural *lac* repression landscape contains 15 independent peaks with palindromic operator sequences (19 including non-palindromic operators). But when all amino acids are considered to be directly connected, the number of independent peaks reduces to 7. This result shows that the topology of the connectivity network (Fig. 5.3B) has a significant effect on the ruggedness of the *lac* repression landscape.

Noise propagation in metabolic networks



The composition of bacterial cells fluctuate dramatically due to molecular stochasticity. How such fluctuations affect basic cell functions leading to growth, is largely unknown. Here we investigate how random fluctuations in enzymes at the beginning of an essential metabolic pathway, propagate to growth rate in individual cells. By decoupling the regulation and the metabolic function of the lac proteins in E. coli, we measure how enzyme level and growth correlate over time at different metabolic states. The fluctuations in both protein level and growth rate are large and their duration are linked to the cell's division time. When lac levels are artificially lowered, enzyme fluctuations propagate through the metabolic network and transmit to growth with little delay. At non-limiting conditions, lac fluctuations are buffered or fluctuations in other components dominate. Interestingly, our dynamic measurements in these conditions reveal that single cell fluctuations in growth rate do cause the lac level to fluctuate with a delay of tens of minutes. As this effect is also observed when the lac proteins do not support growth, it is likely a cell-wide phenomenon acting on all genes. These results reveal a largely overlooked interdependency of fluctuations in single cell growth rate and protein levels, which is important for the understanding and modeling of biological networks and complete cells.

Biological cells are amazing factories that, despite their small size, perform all tasks required for life. They are built from different types of molecules, many at very low numbers, whose interactions with each other and with their environment shape the cell. As these molecular processes are inherently stochastic, the characteristics of individual cells are not deterministic, and vary randomly over time. Only recently it has been recognized that these variations are quite large: protein levels generally vary over 10%, and often much more than that [23, 133, 134]. Importantly, these fluctuations were also shown to persist over long timescales, lasting over a cell cycle [21]. This is much longer than for example the equilibration times of the metabolites in a metabolic network [135]. The realization that all protein levels fluctuate significantly, prompts the question what physiological consequences this has for living cells.

The most noisy proteins in the cell are those at low expression levels, such as regulatory proteins [136, 137]. Many studies have linked the noise in these regulators to dramatically different cellular behaviors. In *B. subtilis*, for example, variation in the concentration of either ComK or ComS trigger some cells to become competent for DNA uptake and others not [138]. In *E. coli*, the all-or-none expression of the *lac* operon when induced with the artificial inducers IPTG or TMG, is due to noisy expression of the LacY protein [21]. And in the flagella system, fluctuations in the level of the CheR protein are responsible for variability in rotation direction [139]. These cases indicate that fluctuations in a single protein can generate phenotypic heterogeneity within an identical population of cells. So, although noise reduces fidelity in regulatory networks, it can also create diversity, that can be advantageous in variable environments [133, 134, 140].

How fluctuating protein levels affect cellular growth remains an open question. Highly expressed metabolic proteins are responsible for catalyzing reactions that lead to the synthesis of new cell material [6, 141]. In bacteria, the average levels of these costly proteins are not random, and thought to be tuned for maximal growth rate, where the level is dictated by a balance between cost of producing and benefit of having the protein [8, 10, 11, 26]. This view is supported by experiments where the average expression level of the *lac* operon rapidly adapts to gain maximal growth rate [10]. Additional support for the tuning of protein levels can be found in the optimal use of the *E. coli* metabolic network on different carbon substrates [83]. Bacterial growth is generally assumed to be proportional to the total flux through its metabolic network, which can be described in terms of the protein activities using metabolic control analysis [142, 143]. How metabolic flux and growth rate in a single cell are affected by fluctuations of its metabolic proteins, present a number of general open questions. Can fluctuations cause a reaction in the metabolic network to become rate limiting? Are fluctuations in metabolites buffered (e.g. by storage or outflux) or do local flux fluctuations propagate through the metabolic network and affect growth? Does the dilution of protein due to growth help in maintaining homeostasis of the system? Do cells have other regulatory feedbacks that counteract the detrimental effect of noise on growth? These questions, fundamental to all living cells, merit experimental investigation at the single cell level.

If variation in the level of metabolic proteins has a significant effect on growth, one would expect to observe differences in growth rate between individual cells.

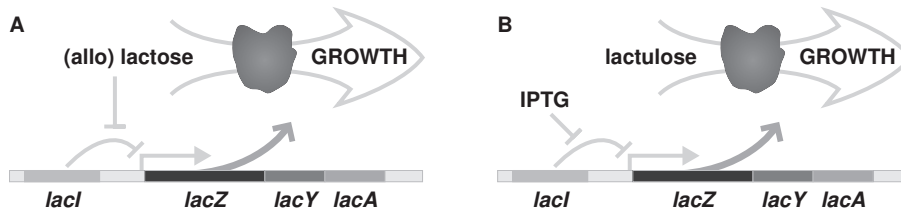


Figure 6.1: The *lac* operon of *E. coli* and decoupling of its regulation and metabolism

(A) In the natural *lac* system, allolactose (an isomer of lactose) relieves the repression of the *lac* operon, after which the *lac* operon allows consumption of lactose and subsequent growth of the cell.

(B) By using two synthetic sugars, the regulation of and the consumption by the *lac* operon can be decoupled. Lactulose is a disaccharide of galactose and fructose, for which the *lac* operon is required for degradation. However lactulose does not reduce LacI's repression, so *E. coli* cannot grow on lactulose alone, as the *lac* operon does not get expressed. IPTG cannot be degraded and used for growth, but does interact with LacI to reduce expression.

This putative heterogeneity is hidden in population growth measurements (which show an extremely constant growth rate in the exponential growth phase), due to averaging over millions of cells. The variation in many cell cycle parameters, such as interdivision time and initiation mass, has been studied extensively [144–146]. However, this variability may well be due to stochastic timing events, rather than differences in the metabolic state of cells [147]. Studies that measured single cell growth rates in terms of biomass increase, report a coefficient of variation higher than 10% [146, 148–152]. Aging of cells might explain a small part of these differences in growth rate [149], but the major underlying source has remained elusive. A central question therefore is: does noise in the expression of those proteins directly involved in biomass accumulation propagate to variations in cellular growth rate?

Here, we measure how fluctuations of enzymes at the beginning of an essential metabolic pathway correlate with the growth rate of individual cells. We focus on the *lac* system in *E. coli*, which performs the first essential steps for growth on lactose. By importing and cleaving this sugar, the *lac* proteins are responsible for all energy and carbon sources entering the cell. Single cells are followed over multiple generations using time-lapse fluorescence microscopy, allowing us to quantify the temporal cross correlation between expression and growth rate.

6.1 Decoupling of the *lac* system

The *lac* operon

In *E. coli*, the *lac* operon is responsible for the import and catabolism of lactose. The operon contains three genes: *lacZ*, *lacY* and *lacA* (see Fig. 6.1A). Only the first two genes, *lacZ* and *lacY*, are required by *E. coli* for growth on lactose. LacY is a transmembrane transporter protein that pumps lactose from the environment into

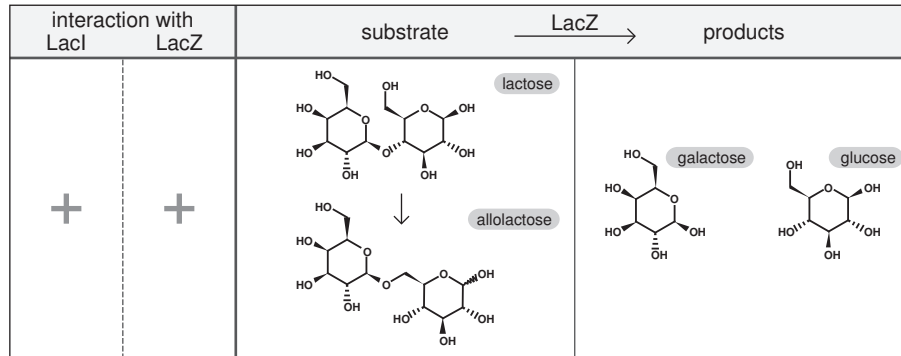


Figure 6.2: Molecular structures and interactions of glycosides in the natural *lac* system

In the natural *lac* system, lactose is cleaved by the LacZ protein into galactose and glucose [7]. A considerable fraction of the lactose is first converted into allolactose [153]. It is the interaction of this isomer of lactose with LacI that induces *lac* expression [154]. Allolactose itself is also cleaved by LacZ into glucose and galactose. This reaction has the same or higher efficiency than lactose, and no lactose is formed [155].

the cell. The intracellular lactose is subsequently cleaved by the LacZ enzyme into glucose and galactose (see Fig. 6.2). These monosaccharides can then be degraded into energy and metabolically useful fragments by other enzymes.

The role of the third protein in the *lac* operon, LacA, is still obscure. It functions as an acetyltransferase for galactosides (for example lactose), but has an extremely low affinity for its substrates [156–158]. Generally, it has been thought to be involved in detoxification [159]. More recently it has been suggested that LacA acts as a safety valve, making sure that when too much lactose is imported into the cell, some is exported to avoid a detrimental osmotic pressure [160]. In any case, *lacA* does not seem to be necessary for lactose catabolism, as knock-outs of this gene, have indistinguishable growth and behavior in laboratory conditions [161].

The expression of the *lac* operon is controlled by the LacI repressor (see Fig. 6.1A) [109]. With no lactose present in the environment, the constitutively expressed LacI protein represses the *lac* operon by binding to the DNA of the *lac* promoter. Once lactose is available, it may cause LacI to unbind, resulting in increased expression of the *lac* operon [7]. Interestingly, LacI repression can only be relieved, once a cell contains a basal level of LacY and LacZ protein [20, 21]. LacY is needed because lactose cannot diffuse passively across the cell membrane. Therefore, lactose can only affect LacI after it is actively transported into the cell. However, lactose does not directly affect LacI. It's actually an isomer of lactose, allolactose that binds LacI causing its binding reduction. This isomer is formed in an alternative reaction by LacZ. Hence both LacY and LacZ are needed before extracellular lactose can reduce the repression of LacI.

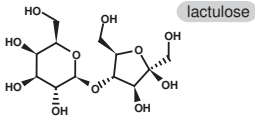
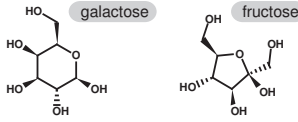
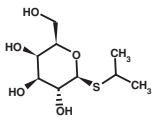

interaction with LacI LacZ		substrate	LacZ →	products
-	+	 lactulose		 galactose fructose
+	-	 IPTG		

Figure 6.3: Molecular structures and interactions of glycosides in the decoupled *lac* system
 In our decoupled *lac* system, lactulose is cleaved by LacZ into galactose and fructose, which can both be used for cell growth [162]. Lactulose does not interact with LacI to induce the *lac* operon [162]. Another glycoside, IPTG, does interact with LacZ, but cannot be degraded by LacZ [7].

Coupling between cue and substrate

The process of cellular catabolism can be divided into two functional tasks: sensing and consumption. Sensing comprises of measuring the availability of a nutrient and transferring this information to the regulation of relevant proteins. These proteins are responsible for consumption, i.e. the conversion of the nutrient into building materials and energy used for growth. In some catabolic systems both sensing and consumption are performed by the same protein [163, 164]. In *E. coli*, for example, degradation of proline is catalyzed by the PutA enzyme, which in absence of proline binds directly to its own promoter region on the DNA, repressing *putA* expression [165, 166]. In most catabolic systems, however, the two functions have been separated and are performed by separate dedicated proteins.

Catabolism of lactose by the *lac* genes is one of those systems where sensing and consumption are separated. The transcription factor LacI is responsible for sensing of lactose and subsequent regulatory steps. The enzyme LacZ, on the other hand, catalyzes the breakdown of lactose, hence its consumption. But, although sensing and consumption are performed by different proteins, the two functional tasks are still coupled: sensing and consumption each depend on lactose, as lactose acts both as cue and as substrate (see Fig. 6.1A).

From an evolutionary perspective coupling between sensing and consumption is favorable, as it ensures that the costly *lac* proteins are only produced when there is lactose to degrade. From an experimental perspective, however, the coupling can be restrictive, as one cannot probe the two functional tasks independently. More specifically, when changing the average *lac* expression by varying the amount of lactose in the environment, this also affects the activity of LacZ and LacY protein, being a function of lactose concentration following Michaelis-Menten kinetics. In order to keep the external growth conditions constant, while varying only the internal expression levels of metabolic genes, we sought to remove the constraint of

coupling between cue and substrate.

Decoupling

Several methods exist that allow alteration of *lac* expression level at constant substrate level. For one, the steady state β -galactosidase activity can be altered by genetic modifications. Such mutants have been obtained for the *lac* operon by reversion of nonsense mutations in *lacZ* [162, 167–169]. In these cases, changes in consumption level were not obtained by changing the *lac* expression levels, but by alteration of the catalytic activity of the proteins instead. In order to get mutants with different steady state expression levels, the promoter region could be targeted by directed mutations, or by experimental evolution [10]. In any case, the choice of expression levels might be limited, as the phenotypes of the produced mutants are still unpredictable. Furthermore, as the new expression level is permanent, dynamic changes cannot be applied.

We therefore attempted an alternative approach where the expression level at full lactose induction is reduced by an ‘inhibitory chemical’. Such an anti-inducer can be found in ONPF (see Table 6.1 on page 97), which stabilizes the LacI repressor when bound to DNA, resulting in lower *lac* expression [154]. Unfortunately, the concentration needed for a significant reduction in expression, also leads to an independent detrimental effect on growth (which we checked by growth on glucose), which makes this method unsuitable for our purposes. Another approach, involving the inhibition of RNA translation for specific genes by antisense agents [170, 171], might give better results, but was not tested.

As sensing and consumption are coupled by lactose, the cleanest method of decoupling consists of replacing lactose by two substances that are only part with one of these processes (one as cue and the other as substrate). For sensing such a chemical is well known: IPTG is a thio-galactoside that can effectively induce the *lac* operon, while not being degraded by LacZ, and thus not supporting growth. Conveniently, there also exists a synthetic sugar with the opposite properties: lactulose is a synthetic disaccharide consisting of fructose and galactose subunits, which is imported and degraded by the *lac* proteins (see Fig. 6.1B and Fig. 6.3). Wild-type *E. coli* cells however, cannot grow on lactulose, as the *lac* operon does not get induced by the sugar [162]. Only after induction with for example IPTG or by the appearance of a constitutively expressing mutant, growth is possible. By using lactulose with varying IPTG levels, we could force cells to use different *lac* protein levels for growth, while keeping the same external growth conditions.

6.2 Single cell measurements of protein level and growth

Fluorescence time-lapse microscopy

Most microbiological techniques for measuring protein levels involve bulk measurements, in which the behavior of individual cells are averaged out. The first method for measuring the level of LacZ protein in single cells, was based on the protein’s hydrolysis of fluorogenic substrates. The signal amplification due to each

LacZ protein producing many fluorescent product molecules, allowed samples containing only a single cell to be quantified [172]. Recently, this method was improved, allowing sensitivity for single LacZ proteins [24]. As the fluorescent molecules are not retained within the cell, this method is limited to measurement of a single cell at a time.

Another single-cell method uses fluorescent proteins to measure protein levels. The popularity of fluorescent proteins is largely due to the possibility to genetically fuse them to almost any protein, resulting in both high sensitivity and high specificity. Single-cell expression measurements can be performed using flow cytometry (see for example [173]). This technique allows automated size and fluorescence measurements of thousands of cells per second, providing excellent statistics. Unfortunately, flow cytometry only produces ‘static’ snapshots, and cannot be used for dynamic measurements of individual cells [174]. Without such dynamic data, single cell growth rates nor the dynamics of protein levels in individual cells can be obtained. Therefore, we used a technique that does allow repeated measurements of both size and fluorescence of growing cells: time-lapse fluorescence microscopy. This technique is essentially the same as normal fluorescence microscopy, but with constantly repeated sample measurement over a long period of time. Such time-lapse measurements are facilitated by automation of the microscope, but also impose some additional requirements on the sample.

In order to make quantitative measurements of cells, they need to be positioned along the microscope focal plane. Additionally, the cells need to be limited in their movement, so that they can be tracked over consecutive measurements. Therefore, the cells should not be free floating (as in batch growth), but stuck to a surface, and constrained to a single layer. Also, the sample should provide constant availability of nutrients, allowing the cells to grow exponentially during the whole experiment. This can be challenging when measuring over long periods of time or when the density of cells in the sample is very high. In our case we need data from many cells, as we are analyzing a non-deterministic (random) process that must be described in terms of statistical averages. For each experiment, the sample should therefore contain a large number of cells within the field of view of the microscope.

A recent trend in system biology is the use of microfluidic samples in order to meet these sample requirements [175]. This technique is especially useful for dynamic changes in cell environment, as it is determined by a flow of fresh growth media. Also, some designs allow the removal of cells, which is necessary in order to avoid the clogging of the sample with cells: under exponential growth, cells quickly grow to a number where both nutrient availability and single layer growth are compromised.

Here, we grew *E. coli* first in batch and then spread a few hundred of them between a small gel pad and a glass coverslip. This resulted in very sparsely distributed cells, where each cell grew into a single layer of cells (i.e. a ‘microcolony’). Depending on the local gel surface properties, *E. coli* cells started to stack out of the single layer when the microcolony reaches about 500 cells, providing sufficient data for statistical analysis. We obtained both phase contrast and fluorescence images from the growing microcolony with constant intervals (see Fig. 6.4). Prior to image acquisition, the cells were focused by automated software focusing. Phase

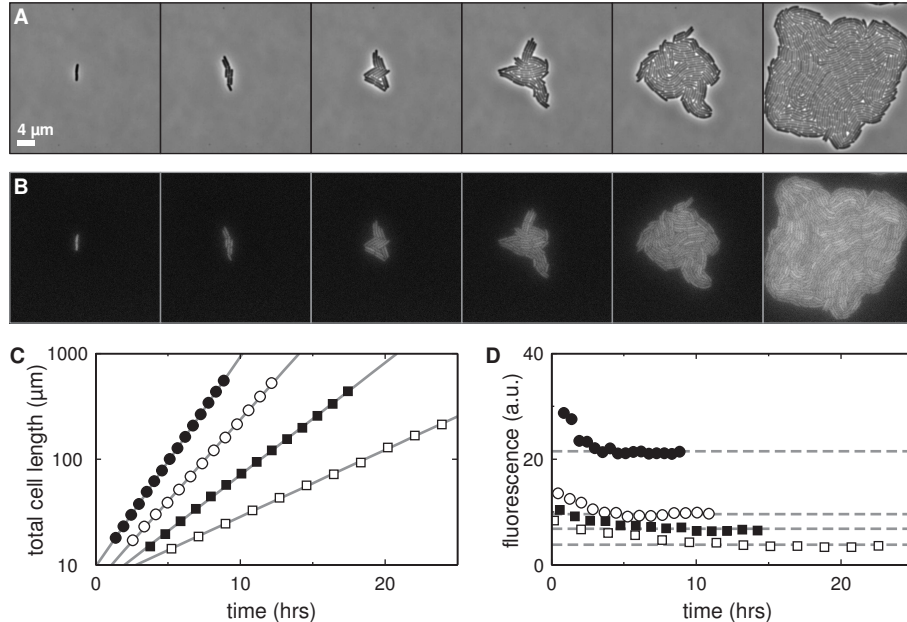


Figure 6.4: Mean growth rate and mean protein level measurement of microcolony

(A) Phase contrast images of a microcolony grown on $6 \mu\text{M}$ IPTG, corresponding to empty square (■) in panels C and D. Images that are shown are approximately 4 hours apart.

(B) Fluorescence images of the microcolony obtained at the same time as those in panel A. In these cells fluorescence is a measure of *lac* level in each cell. Although the cells are genetically identical and experiencing the same environment, there clearly exist a large variation in expression levels between the cells.

(C) By fitting the total length of the microcolony (sum of the cell lengths in the microcolony) versus time by an exponential, the mean elongation rate of the microcolony can be determined. Data is shown for *E. coli* microcolonies grown on lactulose with IPTG concentrations ranging from $200 \mu\text{M}$ (●) to $4 \mu\text{M}$ (□).

(D) The mean expression level can be determined by averaging the fluorescence of those cells after the initial decrease (fitted lines). Data is from the same microcolonies as panel C. Note that the datasets in C and D are horizontally shifted for clarity.

contrast imaging were acquired with a high frequency in between 25 to 75 images per cell cycle. Once a time-lapse series of microcolony images was obtained, the images were analyzed offline. Each phase contrast image was analyzed separately and used for cell identification and measurement of cell length (see Section 6.7).

In order to measure the expression of the *lac* operon, we used an *E. coli* mutant where the chromosomal *lacA* gene was replaced by a fluorescent protein (see Section 6.7). This transcriptional fusion ensures that fluorescence is an accurate measure of *lac* protein level. As protein noise is predominantly transcriptional, and not translational, noise properties of the fluorescent marker should be similar to the *lac* proteins. Compared to reporter systems inserted elsewhere on the

chromosome [22], the transcriptional fusion does not suffer from errors due to intrinsic noise [23]. Also, with *lacA* far away from the *lac* promoter, the regulation and expression of the *lac* operon should be minimally affected in this mutant. Furthermore, no additional *lac* promoters are added to the cell, such as in plasmid based systems, hence LacI function is likely to be unaltered. Although the level of the fluorescent protein is somewhat lower than that of the *lac* proteins, as suggested by the relative monomeric expression levels of approximately 4:2:1 for LacZ, LacY and LacA [176, 177], fluorescence was high enough for quantitative measurements.

The fluorescence protein inserted as a reporter was GFPmut2, which is both bright and stable [178, 179]. The maturation time of this fluorophore is less than 10 minutes [146, 180] and in our exposure conditions photobleaching was negligible. Upon illumination a fraction of the GFP proteins may form free radicals which, due to their toxicity, cause cells to grow slower. This limits the frequency that fluorescence images can be obtained for growing cells. We determined exposure conditions without significant toxicity, by comparing the growth rate of microcolonies with and without illumination. In our setup, toxicity increased with average GFP level, illumination time, illumination frequency and cell cycle time (data not shown). At full *lac* expression, we did not observe a significant effect on growth rate when cells were illuminated approximately five times per cell cycle, so we used this frequency throughout our experiments.

Each time-lapse experiment produced hundreds of digital images of cells throughout time. Extraction of relevant data from these images required cell identification followed by quantification of cell length and fluorescence. As this is a tedious and subjective task when performed by hand, this analysis was automated using software based on Matlab code kindly provided by Michael Elowitz [25]. The average fluorescence and growth rate during an experiment could be determined by analysis of only a fraction of the images. Such analysis are an alternative to 'population' measurements from batch and can reveal whether cells are growing under steady-state conditions (see Fig. 6.4)

Average protein level and growth in microcolony

The average protein level during the experiment was determined by averaging the fluorescence of all individual cells in each image. Surprisingly, we consistently observed an initial decrease in average fluorescence followed by a constant steady state level after approximately four cell generations (see Fig. 6.4D). Bleaching is negligible in our setup. Instead, this might be due to different inducer conditions in the sample, or maybe an adaptation of the cells to growth on the gel pad. Interestingly, no significant changes in cell morphology or growth rate (see Fig. 6.4) were observed during the decay in fluorescence. In a different study, cells were 'equilibrated' by growing them on a gel pad first and transferring them next to a fresh gel pad [181]. Generally, we discarded data containing the decay and only used data for further analysis once average fluorescence level had reached equilibrium.

Cells grown on lactulose with varying amount of IPTG showed different protein levels. At 200 μ M IPTG cells were highly fluorescent, and the average fluorescence gradually decreased when cells were grown on lower IPTG concentrations

(see Fig. 6.4D). We observed no bistability in the *lac* system when cells were grown on lactulose and low levels of IPTG. We did observe bistability when cells were grown on glucose and IPTG (data not shown), as has been observed before for the similar artificial inducer TMG (thiomethyl- β -D-galactoside) [20, 22]. Interestingly, cells growing on lactulose and 200 μ M IPTG were significantly more fluorescent than cells grown on either glucose with 200 μ M IPTG or lactose (data not shown). This might be due to the higher growth rate supported by these sugars, resulting in relatively less protein in the cell [6]. Also, the effective induction by lactose might be weaker than that of IPTG, as the former is degraded whereas the latter is not.

The average cell growth rate (or the microcolony growth rate) was determined by fitting an exponential through the sum of the length of all cells in the microcolony (see Fig. 6.4C). Compared to averaging the elongation rate of each separate cell in the microcolony, as done in Fig. 6.7, this method requires relatively little data analysis. As all data can be well fitted by an exponential, this suggests that the cells grow exponentially throughout the experiment.

The average growth rate at induction with 200 μ M IPTG was about 0.8 doublings per hour, corresponding to a generation time of about 75 minutes (see Fig. 6.4C). This growth rate is slightly lower than that for growth on lactose, for which we measured generation times of about 65 minutes (data not shown). This lower growth rate may be due to several causes. The *lac* enzymes may have lower catalytic efficiency for lactulose uptake and degradation, hence their expression level may be suboptimal. The lower growth rate with respect to lactose could also be due to lactulose being a less favorable sugar for growth, as the hydrolyzation of lactulose results in a fructose where lactose produces glucose (see Fig. 6.2 & 6.3). *E. coli* did not grow in minimal media containing lactulose without IPTG, confirming that lactulose does not induce *lac* expression. When cells were grown on lower IPTG concentrations, also the growth rate gradually declined. Growth rates down to 5 hours per generation could be obtained at induction with 4 μ M IPTG. Note that mutants expressing *lac* constitutively arise quickly, so care was taken when growing cells with low IPTG levels.

Removal of contamination with organic compounds

In most of our experiments, we observed constant average growth rates throughout the growth of the microcolony (see Fig. 6.4C). This suggests that the cell environment stayed constant. However, in pilot experiments where cells were grown on lactulose with very low levels of IPTG, the average growth rate drastically declined during the first few cell cycles (data not shown). Following the decline, the growth rate remained very constant again, which suggested to us that the cells might initially be growing on a contamination which was slowly depleted.

In conditions where cells can only grow slowly on the provided carbon source, it is necessary to reduce contamination of the media with carbon sources other than those under investigation. Contaminant free media is especially important when growing very dilute cultures as consumption of the contaminants will take a long time. In the case of cell growth on agarose pads, we suspected that the cells might grow on the agarose itself, as it is a galactose-based polysaccharide. Indeed, we ob-

served that *E. coli* can grow on agarose pads without a carbon source. Therefore, we developed a new technique, where cells are grown on a polyacrylamide gel that is composed of sugar-free compounds instead (see page 98 of Section 6.7). When dilute stationary cells were applied on such gels without a carbon source, we were surprised to again observe growth at a significant speed for a number of generations. Similar effect have been noted for growth in batch [182–184], and on agarose pads (Matthias Heinemann, personal communication).

Despite thorough washing of glassware and using distilled water, the cell growth persisted. Therefore, we resorted to a new approach in which the organic contaminants are consumed by *E. coli* cells before the actual growth experiment. To this purpose, we grew cells with a knocked-out *lac* operon (NCM520) on the acryl gel. These cells cannot use lactose nor lactulose as a carbon source, but were observed to grow on the contaminants. When more of these *lac*⁻ cells were applied, growth halted earlier indicating faster consumption of the contaminants (see Fig. 6.5).

When the *lac*⁻ cells had consumed the contaminants, we added fluorescent AB460 cells, and start our growth experiment. We could confirm that the contaminants had been removed, as we observed normal growth for the fluorescent *lac*⁺ cells, and no growth for *lac*⁻ cells. Also, cells grown on lactulose with very low levels of IPTG grew with a low constant average rate from the beginning onwards.

In order to apply this contamination removal method for growth experiments on other sugars such as glucose, it is necessary to use *E. coli* cells deficient for growth on the respective sugar. Such mutants may exist (e.g. see [185] for glucose), but it remains to be seen whether these mutants can consume the (unknown) organic contaminations found here. When the media supports a growth rate that is significantly higher than observed for growth on the contaminants, the removal method does not need to be applied, as the contaminants can only have negligible effect on growth.

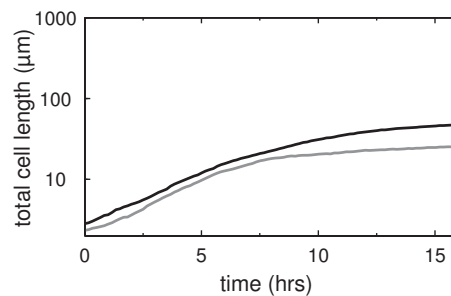


Figure 6.5: Pre-experiment organic contamination consumption

Growth curves are shown for *lac*⁻ *E. coli* microcolonies on acryl gels containing minimal medium with lactulose. When $\sim 10^2$ cells μl^{-1} were applied (black line), an initial growth rate μ of $\sim 0.43\text{h}^{-1}$ was observed. This growth occurred on an unknown carbon source and halted ($\mu < 0.05\text{h}^{-1}$) at a total cell concentration of $\sim 10^6$ cells ml^{-1} . When twice as many cells were applied (grey line), the available carbon source is consumed faster, growth halts earlier and microcolonies reach lower length.

Analysis of single cell lineages

We have shown above how we measure the average protein level and growth rate at varying conditions by analyzing only about a dozen of images from a microcolony. In order to measure how protein levels and growth rate are correlated within individual cells, we aimed to determine both these properties for single cells within the microcolony. The measurement of a cell's protein concentration is determined by extracting the cell's mean fluorescence. The growth rate of an individual cell, however, cannot be determined from a single image, as we determine it by the cell's length increase over time. In order to measure the length of a particular cell within a microcolony over time, it must be 'tracked' over subsequent images (see Section 6.7).

Cell tracking allows the construction of time traces of cell length. By also tracking cell division events, we can construct lineages, which for cell length have a typical sawtooth pattern (see Fig. 6.6 for data from a typical experiment). Considering the lineages from a single microcolony, it is important to realize that many lineages share data points in the earlier generations, as they descent from the same ancestor. When comparing the lineages extracted from a single microcolony experiment, the division of cells are initially synchronized. This synchrony is lost after a few generations, due to for instance variability in generation times and inaccuracy in the equal division of cells [186, 187]. Considering the length traces of individual cells, we observe that cell elongation is a quite constant and robust process (see Fig. 6.6C). The occasionally observed decreases in length over time are likely due to measurement errors. The length traces are straight when plotted logarithmically, suggesting length increase in a single cell is exponential, in accordance with previous observations [145]. However, the accuracy of our length measurements is limited by optical resolution, making it hard to distinguish exponential growth from bilinear growth [188].

We use the traces of length over time to determine the cell elongation rate. We do this by fitting the length measurements from a complete cell cycle with an exponential (see Fig. 6.20 in Section 6.7). As the diameter of exponentially growing cells is quite constant [189], we assume that length increase is good measure of growth rate. This technique also allows us to look at differences in growth rate within a single cell cycle by fitting the lengths from only a part of the cell cycle (see Fig. 6.20 in Section 6.7). However, for a confident measurement of a change in length, the change in length must be significantly larger than the error in the length measurement. In our experience, a length increase of about 0.4 μm is required, limiting us to a time resolution of about one third of the cell cycle. We do not observe a correlation between growth rate and the phase of the cell cycle (data not shown).

The tracking of cells also allows to construct time traces of protein level within cell lineages (see Fig. 6.6C). Using these traces the average protein level during the whole cell cycle can be determined. By extracting the traces of a cell's total fluorescence (instead of mean fluorescence), the protein production rate can be determined [25]. We, however, consider the mean fluorescence, which corresponds to the protein concentration in a cell, as we are not interested in the source of the noise, but the result of the noise in the protein level. Also these traces do not show

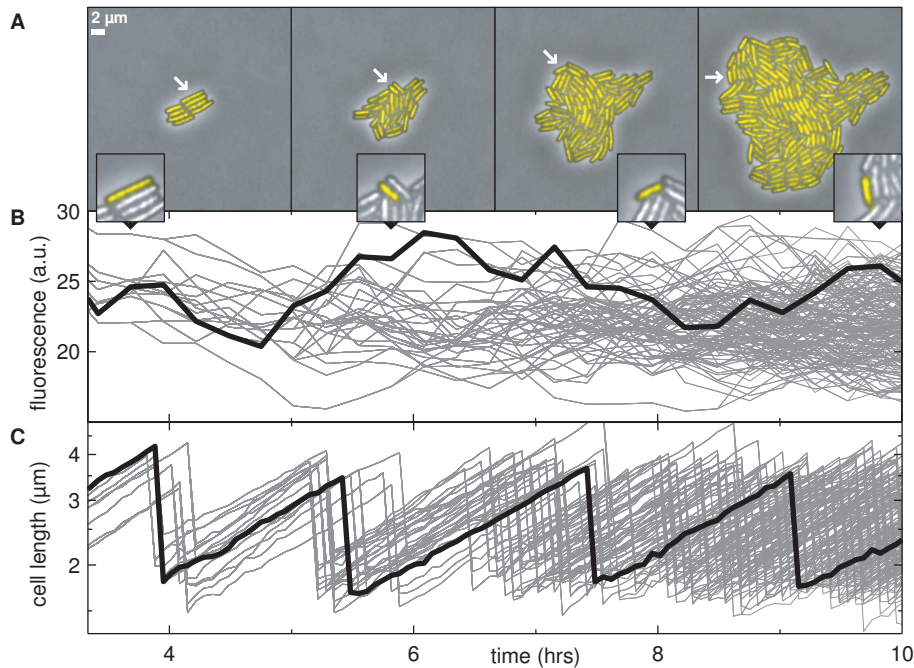


Figure 6.6: Length and *lac* level of cell lineages within a microcolony

(A) Snapshots of a typical microcolony growth experiment using fluorescence time-lapse microscopy. Phase contrast images (grayscale) are overlaid by fluorescence images (yellow) of GFP reporting for *lac* expression. Insets show magnification and highlighted cell of a single selected trajectory through 4 divisions.

(B) Fluorescence time traces for individual cell lineages obtained from the experiment shown in A. The selected cell lineage from A is shown as a bold black curve.

(C) Cell length time traces for individual cell lineages. The selected trajectory from A is shown as a bold black curve. Note that length is plotted logarithmically.

a correlation between protein level and the phase of the cell cycle.

6.3 Correlations between protein level and growth

Protein and growth noise at different metabolic states

Using single cell growth and fluorescent measurements, we can determine the amount of noise in these properties within and between cells. The cells analyzed in any particular experiment are isogenic. As we vary the amount of IPTG in the environment, we vary the average *lac* protein level. The growth of cells is dependent on the flux of nutrients converted by the *lac* proteins, so by varying the IPTG level we can force the cells into different metabolic states.

A histogram of the fluorescence data at different IPTG concentrations (see Fig. 6.7A), shows that with increasing fluorescence levels the absolute variation in flu-

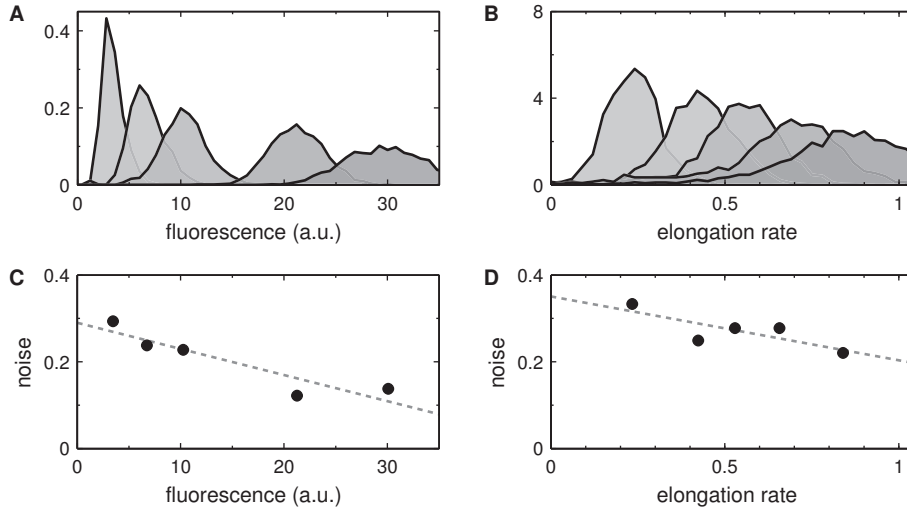


Figure 6.7: Noise in single cell *lac* level and growth rate

(A) Histograms of single cell *lac* expression levels as determined from their mean fluorescence level. Cells were grown at steady state on acryl pads with increasing levels of IPTG, from 4 μM (left histogram) increasing to 200 μM (right histogram)

(B) Histograms of single cell growth rates as determined from single cell elongation rates. Data was obtained from the same experiments as in panel A, with 4 μM IPTG for the left histogram, increasing to 200 μM for the right histogram.

(C) The *lac* expression level noise (defined as the standard deviation divided by the mean) versus the mean *lac* level. Data corresponds to panel A.

(D) Noise in single cell growth rate versus the mean population growth rate. Data corresponds to panel B. Dashed lines in panel C and D are added as guide to the eye.

orescence increase (the width of the distributions become wider). However, when we look at the noise, defined as the standard deviation divided by the mean, we observe an inverse correlation with the expression level (see Fig. 6.7C). The noise at full expression is about 10%, whereas a ten-fold lower expression level leads to a noise of about 30%. This agrees with previous work [23] where noise also depended on the average expression level.

When *E. coli* is grown exponentially in the lab, extremely constant growth rates are generally observed. As these measurements are an average over a population of cells, individual cells do not necessarily grow at a constant rate. Instead, several recent studies have shown that individual cells in exponentially growing populations vary significantly in their elongation rate [149, 150, 152]. We have quantified the elongation rates within a population of cells growing at different rates (see Fig. 6.7B), and found very large variations. Also, a similar trend as for protein level can be observed for growth rate: increasing growth rate leads to an increase in the variation (see Fig. 6.7D). However, when considering the noise, again the inverse trend is found. The noise decreases from 35% at low growth rates (0.25 doublings per hour) to 20% at high growth rates (0.8 doublings per hour).

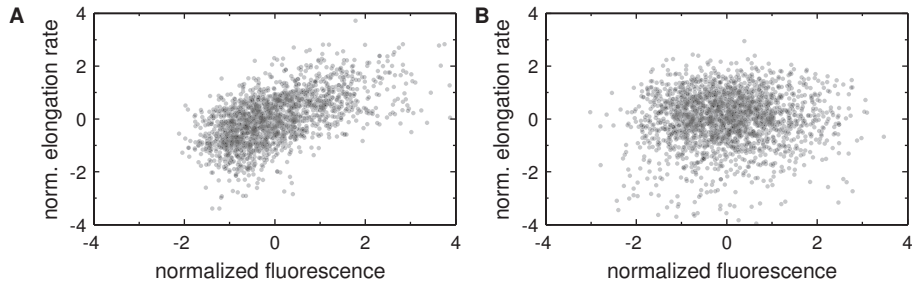


Figure 6.8: Correlation between single cell *lac* expression and growth rate

In these scatter plots, each point represents single cell fluorescence and elongation rate at one point in time. For comparison the data is normalized by its mean and standard deviation.

(A) Data for microcolony growing on 4 μ M IPTG having a low mean expression level. This data set has a correlation coefficient (R) of 0.56.

(B) Data for microcolony growing on 200 μ M IPTG having a high mean expression level. Its correlation coefficient is much smaller ($R=0.05$).

Correlation in single cells

When cells are grown at full *lac* induction level, there is significant variation in both protein level and growth rate. Surprisingly, however, there is hardly any correlation observed between these properties (see Fig. 6.8B, $R \approx 0.05$). This suggests that in these conditions the fluctuations in the *lac* proteins have little effect on cellular growth. It can also mean that fluctuations are buffered out. This can be because the *lac* proteins are not limiting, and one or more other processes limit growth. It can also be the case, that there is no direct dependency between growth and expression level, but a more complex dynamic relation, which is hidden in this analysis. At the same time, while cells vary in growth rate, this does not directly seem to affect protein levels in the cell.

When cells are grown on lower IPTG levels, both the noise in protein level and the noise in growth rate increases. In this case we do observe a strong correlation between the two properties ($R \approx 0.56$), suggesting that random temporal fluctuations in *lac* result in temporal fluctuations in growth rate (see Fig. 6.8A).

Steady state relation between protein level and growth

One putative explanation for the difference in correlation between *lac* level and growth rate between high and low expression levels, could be whether *lac* levels are limiting for growth. We have already concluded that lower expression levels lead to lower growth rates, but we sought to find the exact relation between these properties here. Therefore we analyzed the average growth rate and average protein levels under different conditions (Fig. 6.9). The data could be well fitted by a Monod

growth equation [191] with the addition of a term for maintenance:

$$\mu = \frac{(\mu_{\max} + m) \cdot E}{K + E} - m$$

, where μ is growth rate, μ_{\max} the maximal attainable growth rate, m the maintenance rate, E the *lac* expression level, and K a constant corresponding to the *lac* level at which the growth rate is half the maximum. Note that, although K is analogous to the Michaelis-Menten constant (K_M), it does not correspond to a single enzyme, but should be interpreted as the affinity of the complete cell towards the substrate [192]. The fit shown in Fig. 6.9 has $\mu_{\max} = 0.93 \text{ h}^{-1}$, $m = 0.23 \text{ h}^{-1}$, and $K = 5.8$ in arbitrary units.

The traditional Monod growth curve describes how growth rate depends on the nutrient concentration in the environment [191]. Interestingly, a formula with the same form can capture the dependency of growth on internal *lac* level (see Fig. 6.9). This suggests that the nutrient flux going through the *lac* system depends on both external nutrient concentration and internal protein concentration in a similar manner.

The basic idea of maintenance is that cells constantly consume some energy for functions other than production of new cell material, analogous to overhead in a business. Traditional maintenance derivations are based on double reciprocal plots of yield and growth rate [193, 194]. Here we determined the maintenance rate with a different method, also allowing us to quantify the minimal amount of *lac* necessary to produce the maintenance energy:

$$E_{\min} = \frac{K \cdot m}{\mu_{\max}}$$

In our system, the threshold *lac* level, E_{\min} , for growth on minimal medium with lactulose is about 5% of the level at full induction.

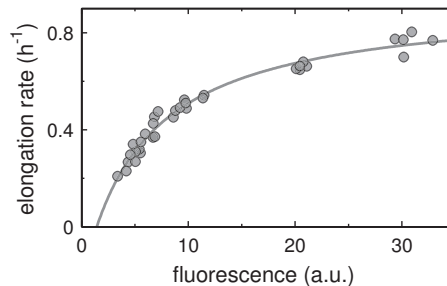


Figure 6.9: Mean growth rate versus *lac* expression level at steady state follows Monod growth

Each point represents the mean elongation rate versus the mean fluorescence of a microcolony grown in steady state at varying IPTG concentrations. Mean elongation rates were determined by fitting the total length of the microcolonies over time, as shown in figure 6.4. The fitted line is a Monod growth curve taking maintenance energy into account (see main text for details). The data was fitted using the EzyFit toolbox [190] for Matlab.

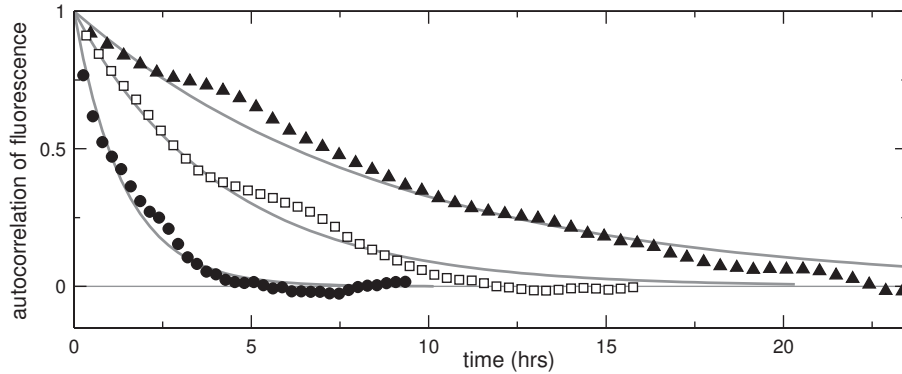


Figure 6.10: Autocorrelation of *lac* expression level in single cells

Branch-corrected composite autocorrelations were calculated with equation 6.5 (see page 106) using data from hundreds of fluorescence time traces. Data is shown for 4 μM (\blacktriangle), 6 μM (\square) and 200 μM IPTG (\bullet). By fitting the autocorrelation with an exponential $2^{-t/\tau}$, the typical correlation time can be determined (see Fig. 6.12).

6.4 Temporal correlations in cell lineages

Autocorrelations

The analysis performed so far has been on dynamical data from single cells, but only instantaneous correlations were determined. As we have a large data set of traces of dynamical data, other analysis techniques can be applied which take time correlations into account. The most basic one is the autocorrelation, which provides a measure of the similarity between observations as a function of the time delay between them [195]. Using this function it is possible to determine how a property is likely to develop over time. For example, if the property at some time point is significantly higher than average, how long does it, on average, take before the property is lower than average. For random processes the autocorrelation resembles an exponential decay, from which the autocorrelation time can be determined by an exponential fit. The autocorrelation time (or similarly the noise frequency) is an indication of how long it takes before a property has a random value again, and can thus be used to characterize fluctuations [25, 196].

Figure 6.10 shows the autocorrelations as determined at different metabolic regimes for growth on lactulose (points). All autocorrelation curves are well fitted with an exponential decay function (lines). In all cases, the autocorrelation approaches zero, which is expected because no correlations should exist at long time delays. As has been observed before, the correlation time for protein expression levels is quite long [25], which means that cells within a homogeneous population differ for more than a complete cell cycle. At a growth rate of about 0.75 generations per hour (full *lac* expression level) it takes about 5 hours before the *lac* level is completely randomized again. As has been observed before [196], at lower growth rates this might take twice or even four times as long (see Fig. 6.10). So in different

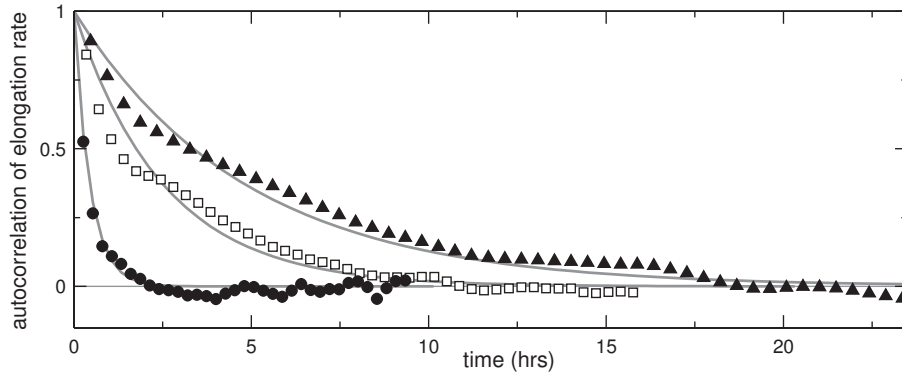


Figure 6.11: Autocorrelation of single cell growth rate

Branch-corrected composite autocorrelations were calculated with equation 6.5 (see page 106) using data from hundreds of elongation rate time traces. Data is shown for 4 μM (\blacktriangle), 6 μM (\square) and 200 μM IPTG (\bullet). The characteristic time scale of the fluctuations can be obtained by fitting the autocorrelation with an exponential $2^{-t/\tau}$ (see Fig. 6.12).

metabolic regimes, the correlation time can be quite different.

How fast growth rate fluctuates within single *E. coli* cells has remained unclear so far. Here we can analyze traces of growth rate within single cells. In order to determine growth rates at the beginning and at the end of the cell cycle, we made use of data that could be extracted using the lineage information. The elongation rate at the end of the cell cycle, could be determined by considering the lengths of the cell's two daughter cells at the beginning of their cell cycle (see Section 6.7). This method allowed extraction of growth time traces without steps around division events. As can be seen in Figure 6.11, the autocorrelation of growth is similar to that of the *lac* protein levels, except that the correlation time are about twice as short (compare to Fig. 6.10). Again the correlations can be nicely fitted by an exponential (dark lines). When *E. coli* cells are grown at full *lac* expression levels, and thus a metabolic state supporting a high growth rate, it takes about 2 hours before the growth rate of single cells are randomized. At lower *lac* levels, and thus also lower growth rates, this takes four to eight time as long.

When we plot the autocorrelation times for growth on lactulose in different metabolic regimes, a correlation with the average doubling time of the microcolony is found (see Fig. 6.12). The autocorrelation time of growth is strongly correlated with growth rate itself. Also the autocorrelation time of the *lac* protein levels are correlated with growth, suggesting that the dilution rate of proteins is a dominating factor [196].

When quantifying delay times between two measured properties, it is important that the measurements are direct. Both the determination of growth and protein level are special cases. For growth it is important to realize that it is never an instantaneous measurement, as we are looking at the rate that length, biomass or another property relating to cell size, increases. So a minimum of two measure-

ments are necessary with some time delay in between them. As size measure, we use the cell length as determined from phase contrast images. These measurements are limited in their accuracy by the small size of cells, and the limit of light wavelength diffraction. As the size also increases quite slowly, it is necessary to use length measurements with a long time delay in between them, for the growth rate determination. This means that the growth measurement is an average around a time point.

The fluorescence measurement has another feature that limits its accuracy. In principle we want to measure the concentration of active *lac* proteins within the cell. The LacZ protein is active within about a minute after its creation (translation by ribosomes from mRNA) [197]. LacY protein is inserted co-translationally into the cell membrane, and although it is still unknown how long it takes before it becomes functional, this will probably also be in the order of minutes. However, fluorescent proteins are not active immediately after folding, as a chemical reaction within the protein has to form a fluorophore first. This process is called the ‘maturation’ of the protein. For some fluorescent proteins this process can take up to hours [179, 198]. The fluorescent protein that we are using (GFPmut2) has a relatively quick maturation time of less than 10 minutes [146, 180]. Within a resolution of ten minutes, these processes can thus be assumed to occur at similar timescales [199].

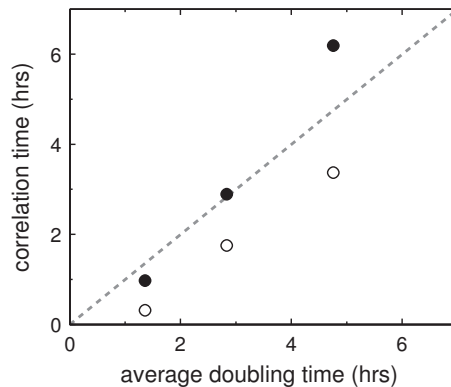


Figure 6.12: Correlation times of *lac* expression level and growth rate depend on the average cell doubling time

The characteristic autocorrelation time of single cell fluorescence is plotted against against the average cell doubling time at varying IPTG concentrations (●). The same is done for the characteristic autocorrelation time of single cell elongation rate (○). The autocorrelation times ($t_{1/2}$) were determined from exponential decay fits as shown in Fig. 6.10 and Fig. 6.11. The dotted line corresponds to the situation where correlation time equals the doubling time.

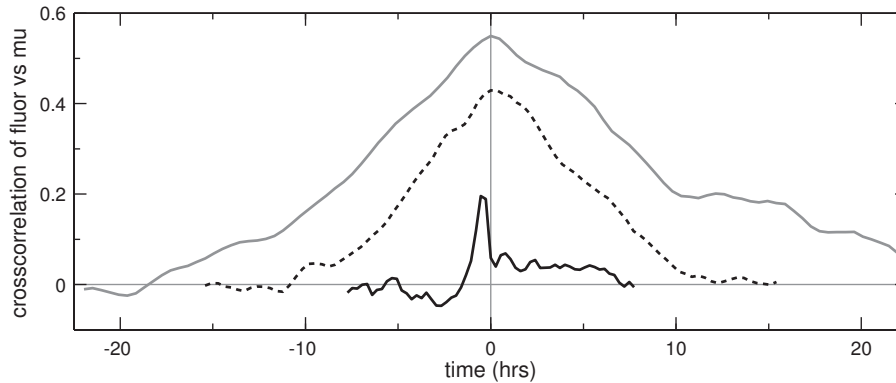


Figure 6.13: Cross-correlation between single cell *lac* expression and growth rate

Branch-corrected composite cross-correlations were calculated with equation 6.5 (see page 106) using data from hundreds of time traces. Data is shown for cells grown on lactulose with 4 μM (grey solid line), 6 μM (black dotted line) and 200 μM IPTG (black solid line).

Cross-correlation

In a cross-correlation analysis, the correlation between two properties are determined over time. In these analyses, a time delay can be found in the correlation between the two properties. This would mean that when property A is high, it takes some time before property B is high. This is a strong indication that it is property A that is affecting property B. As we will see, it sometimes reveals new correlations, hidden when no delay is considered.

When there is a causal relation between two processes and it takes a significant amount of time for the affected process to respond, a cross-correlation will look different depending on the type of causality. In the most simple model of our system there can be four different causal relations (see Fig. 6.15). In this model there are two single cell properties, the *lac* protein level and the growth rate, which can affect each other both positively and negatively. For each ‘arrow’ in the model, a process can be imagined (see legend of Fig. 6.15), and each causal relationship would show up in a different quadrant of a cross-correlation curve.

In Figure 6.13 the cross-correlation between growth and *lac* protein level are plotted under different growth conditions. The cross-correlation for growth on lactulose with low *lac* levels (grey solid line) resembles a pyramid with the maximal correlation when there is no time delay. The value at $t=0$ is actually the same as the R that can be determined from Fig. 6.8B. The fact that there is also a strong cross-correlation at both positive and negative time delays, can be due to the long autocorrelation times of both growth and protein level.

The cross-correlation between growth and *lac* level was also determined for cells with somewhat higher levels of *lac* protein. At higher protein levels (black dotted line in Fig. 6.13), the cross-correlation is lower than before, but effectively keeps the same shape. Again more weight seems to lie at positive time delays.

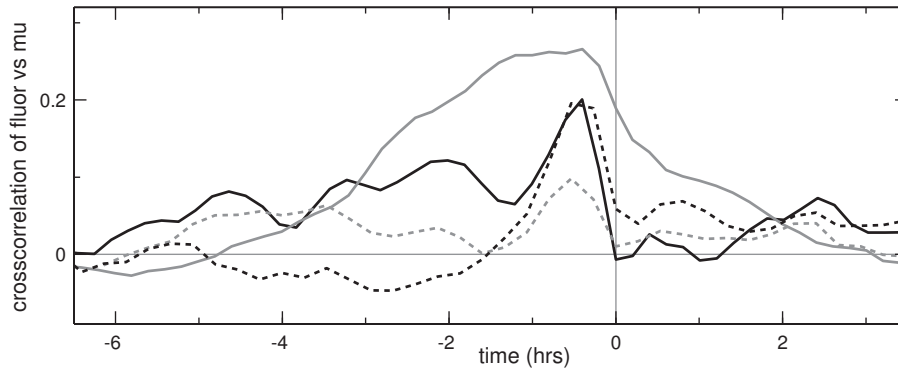


Figure 6.14: Time delay in cross-correlation between single cell *lac* expression and growth rate

Branch-corrected composite cross-correlations were calculated with equation 6.5 (see page 106) using data from hundreds of time traces. Data is shown for 2 independent experiments where cells were grown on lactulose with 200 μM IPTG (dotted lines). Additionally, data from two additional experiments are shown, with cells grown on lactose (black solid line) and cells grown on glucose with 20 μM IPTG (grey solid line, average of two micro-colonies from the same experiment).

When cells were grown on lactulose at full *lac* expression level, no significant correlation without time delay was found (see Fig. 6.8A). Surprisingly, the cross-correlation of these data reveal a significant correlation between growth and protein level at a negative time delay (black solid line in Fig. 6.13). In these growth conditions, fluctuations in cellular growth rate result in correlated protein fluctuations about 20 minutes later. These results suggest that cellular growth, or some process related to growth, affects the expression level of the *lac* proteins.

The delay in correlations between growth and *lac* level was also observed in other growth conditions (See Fig. 6.15). Cells grown on lactulose with about 30% lower *lac* expression level (grey dotted line in Fig. 6.15) showed a somewhat weaker correlation, but still had the same delay. Interestingly, cells grown on lactose also resulted in the same time delay (black solid line). Growth on lactose should be very similar to growth on lactulose with high IPTG levels, but some differences in both sensing and consumption are to be expected. Both lactose and IPTG can fully induce the *lac* operon [7], but whereas (allo)lactose is also consumed by the cell, the imported IPTG is not. Also, the consumption of lactulose is expected to be less efficient, as the *lac* proteins did not adapt to this artificial galactoside, and its product contains the energetically less favorable fructose, instead of glucose (compare Fig. 6.2 and Fig. 6.3). These differences may explain the observation that on lactose the average growth rate was somewhat higher and the average *lac* level significantly lower than on lactulose with IPTG (data not shown).

A time delay between growth and *lac* levels was also observed in a very different growth condition. Cells grown on glucose with a moderate level of IPTG (20 μM), showed a less pronounced peak in the cross-correlation, but also a significant

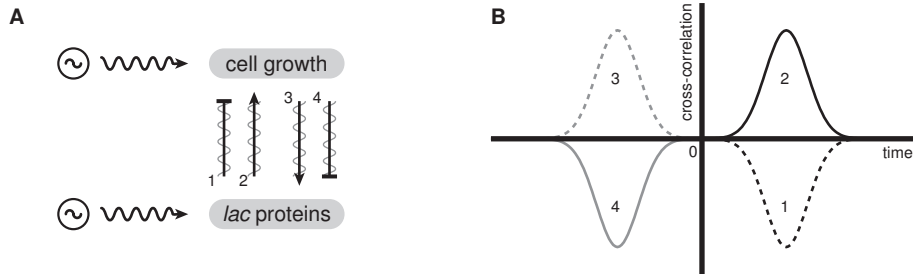


Figure 6.15: Interdependencies between cell growth and protein level result in specific time delays in cross-correlation.

(A) In a minimal model, cell growth and protein level can have both positive and negative effect on each other (solid lines). Cell growth can be affected by the protein level with regard to the cost of producing them (1), and the benefit of having them (2). The protein level is affected by cell growth by protein production (3), and dilution (4). The fluctuations due to noise sources (wiggly lines) can be either buffered or propagated with some characteristic time delay.

(B) Resulting time dependent cross-correlation between protein level and cell growth for each separate interaction. Dips are caused by negative effects (1 & 4), whereas peaks are due to positive effects (2 & 3). The direction of the interaction determines the sign of the time delay. The magnitude of the time delay depends on unknown processes such as buffering.

negative time delay (grey solid line in Fig. 6.14). In this growth condition, the *lac* proteins cannot contribute to growth (effectively the arrow number two in Fig. 6.15 is absent). The observed correlation is therefore likely a general feature, where fluctuations in growth affect all protein levels in the cell.

6.5 Conclusion

Only recently it has been realized that the molecular randomness to which all living cells are subjected, results in large and long fluctuations of molecular components in the cell [23, 25]. The magnitude of variation is remarkable, prompting the question how cells are affected by all these fluctuations. For example, these fluctuations are likely to also occur in all cellular components that are responsible for cell growth, from any essential enzyme producing precursor metabolites up to the assembly machinery of the cell envelope [6]. In *E. coli* a vast metabolic network is responsible for the catabolism (breakdown) of nutrients and the anabolism (build up) of useful precursor molecules. Fluctuations in the abundance of intermediate metabolites in these networks do not propagate and can therefore only have little effect on cell growth [135]. But, whether fluctuations in enzyme levels affect the growth of individual cells has remained unclear so far. Whether they do, must depend on the magnitude and duration of these dynamic fluctuations, but also on the unknown way how fluctuations propagate through the complex cellular metabolic network leading to growth. Here, we investigated how the naturally occurring fluctuations in the level of essential metabolic *lac* proteins affect the growth of individ-

ual cells.

Steady state growth rate depends on average *lac* level

We used an *E. coli* strain in which we could measure the cellular concentration of the *lac* proteins, which are responsible for the first essential steps in the catabolism of lactose or lactulose. By decoupling the regulation and the metabolic function of these proteins, we could vary the average *lac* level which led to corresponding changes in steady state growth rate. Dean et al. already showed by means of genetic mutants growing in nutrient-limited chemostats, that the relation between *lac* activity and growth rate follows a concave function [168, 169]. We found a dependency of steady state growth on the average enzyme expression that followed Monod growth [191] (see Fig. 6.9). We also quantified the minimal amount of *lac* proteins that are necessary before any growth is possible (indicating *E. coli*'s maintenance energy [193, 194]). Interestingly, changes in average *lac* level of the magnitude observed for spontaneous fluctuations, would eventually (in steady state) result in significant changes in growth rate.

Recently, there has been renewed interest in describing the expression of protein to be evolutionary tuned by a cost-benefit relation [8, 10, 11, 26]. This theory assumes that expression of a particular protein is costly, which in general reduces the growth rate of the cell. Hence, when there is no use for the protein, there is a strong selective pressure to reduce the expression. When there is a use for the protein, and in some way it helps the growth of the cell, and actually increases the growth rate, there is a selective pressure to increase its expression. These combining factors result in a cost-benefit relation, which theoretically, and indirectly have been determined to have a concave form, with a negative second derivative [10, 11, 26]. Its concave form is due to extremely high expression levels not having benefit for the cell, and only causing cost, hence resulting in a decrease of fitness.

The data presented in Fig. 6.9 can be regarded as benefits minus the costs of the *lac* system for growth on lactulose. Although we do see a growth rate decrease when cells are growing on glucose with high levels of IPTG, we do not see this for growth on lactulose. This means that the benefit of additional *lac* proteins remains larger than the cost of their expression [8, 10]. Artificially increasing the copy number of the *lac* genes could perhaps move the system into the regime where additional expression is detrimental. Note, that this cost-benefit curve is for growth on lactulose, and not on lactose, for which the *lac* system has evolved. Although the *lac* proteins can consume lactulose, their expression and catalytic activity are not optimized for it. Most probably, when grown on lactose the expression level of the *lac* proteins at full induction, are closer to the peak of the cost-benefit function than for growth on lactulose.

Protein and growth fluctuations depend on cell cycle duration

Quantification of *lac* fluctuations at different average expression levels showed noise to rise from 10% at high levels, to 30% at lower levels (Fig. 6.7). Similar noise levels were observed at these expression levels when cells were grown on glucose, where

growth rate did not depend on expression level [23]. This suggests that the protein noise amplitude is not affected by cell doubling time. In contrast, the protein noise correlation time does depend on growth rate. At full induction levels, we observed a correlation time of *lac* protein of about 1 hour. As has been observed before, this is quite similar to the cell division time [25, 196]. At lower *lac* expression levels the correlation time increased dramatically, but remained close to the cell division time at those conditions (see Fig. 6.12).

Variation in growth rate between single cells had thus far been sporadically quantified [150–152]. We quantified it in terms of elongation rates, revealing a surprisingly large variation in growth rates of around 25%. To our knowledge, the duration of growth rate fluctuations have thus far never been reported. Here, we found it also to be correlated with the average cell doubling time, but also significantly shorter than the correlation time of protein fluctuations, especially at high growth rates (see Fig. 6.12).

Protein fluctuations can propagate to growth

Having characterized the fluctuations in *lac* level and growth, we addressed the question whether the temporal *lac* fluctuations propagate to growth in a similar fashion as in steady state. Indeed, at low expression level, the regime where *lac* is limiting, a strong correlation between protein level and growth was found (see Fig. 6.8A). This suggests that the fluctuations in protein levels can propagate through the metabolic network, transmit to growth and that there is little or no buffering against these *lac* fluctuations. A cross-correlation analysis did not reveal a time delay associated with this propagation (see grey solid and black dotted lines in Fig. 6.13). Such a delay could be expected, because the *lac* proteins do not directly affect growth: after the *lac* proteins have imported nutrients and converted them into monosaccharides, a multitude of processes still need to be performed, before the nutrients can actually contribute to growth. Although the cross-correlation does hint at a small positive delay (more weight at positive time points compared to negative time points), the strongest correlations were always found at time delays of exactly zero. Therefore, if a time delay is associated with the fluctuation propagation, it must be below our time resolution. Taking a small delay in the maturation of our fluorescence reporter into account, the propagation delay could still be up to 10 minutes.

At high *lac* expression levels, protein variation was still of a magnitude that would have resulted in small but significant growth changes at steady-state (compare Fig. 6.7A and Fig. 6.9). But in contrast to when low *lac* levels are low, we did not observe a strong correlation between protein level and growth rate at high expression levels (see Fig. 6.8B), implying that here *lac* fluctuations do not propagate to growth. There could be some unknown reason why buffering does occur at high *lac* levels, but it is more likely that the lack of propagation is due to a lower ‘control coefficient’ of *lac* on the flux through its metabolic path. In other words: when there are few *lac* proteins, small changes in their level have a large effect on growth, as *lac* is responsible for the most limiting step in the growth process. But when *lac* levels are high, it is less limiting, and small changes in *lac* level therefore have much lower

effect on growth. Interestingly, the variation in cell growth rates were still very large in this regime, prompting the intriguing question what noisy process is dominating growth here.

Growth fluctuations propagate to protein level

When we analyzed the cross correlation between protein level and growth at conditions where *lac* was fully expressed, we found a surprising result that remains hidden in static measurements. A significant correlation with a maximum at a negative time delay was observed (see black solid line in Fig. 6.13). As the time delay is negative, the causality of the correlation must be from growth to protein level. It has been observed before that higher growth rates lead to higher protein levels (predominantly for the ribosomes) [6, 200]. However, so far these observations were always for steady state growth conditions. In our single cell experiments, the changes in growth rate are natural fluctuations which last for only a short amount of time (tens of minutes in these cases). As cells are growing in steady-state conditions, and fluctuations are only short-lived, dramatic changes in cell composition are unlikely. However, we observe this delayed growth effect on protein level for growth on different sugars, suggesting it is a general feature for exponentially growing cells (see Fig. 6.14).

Interestingly, our observations of changes in growth rate preceding changes in protein level, contrast with measurements in yeast where changes in gene expression were observed to precede changes in growth rates [201]. The latter observations can be explained by environmental sensing that feed-forward to gene expression levels, only later followed by correlated changes in growth rate. Possibly both feedback and feed-forward mechanisms cause correlations between growth rate and protein level.

We did not observe the delayed growth effect in growth conditions where *lac* levels are low. This could be due to the low growth rates at these conditions, at which *E. coli* generally contains a surplus of ribosomes [202]. Alternatively, correlations between *lac* level and cell growth might be dominated by propagations of *lac* fluctuations to growth, masking the (delayed) effect of growth on overall protein level in the cell.

Why does *E. coli* not grow faster?

Interestingly, when a population of *E. coli* cells is growing exponentially with a very constant rate, many cells are actually growing faster than the population rate. The question arises now whether cells could in any way change such that they would all grow at this higher growth rate, resulting in a higher average population growth rate. As *E. coli* experiences a strong evolutionary pressure to maximize its growth rate, there must be some reason why it has not adapted to reduce its growth rate fluctuations. Although our experiments do not clarify this issue, we can suggest several reasons why growth fluctuations remain and *E. coli* does not grow faster.

First of all, it could be costly to reduce variations in growth rate. Assuming that growth rate fluctuations are caused by fluctuations in gene expression and other

cellular processes, cells could incorporate mechanisms that reduce these fluctuations. However, the formation of such mechanisms would cost energy and cell mass, resulting in a reduction of the average growth rate. So there could be a trade-off between growth speed gained by noise reduction and growth speed lost due to the cost associated with this noise reduction. Second, population fitness is not only increased by maximizing the growth rate in a constant environment, but also by effective responses to fluctuations in the environment. As heterogeneity within a population can be beneficial in the face of environmental changes [26, 56], reduction in growth fluctuations may also have detrimental effects on fitness. Third, it could be that even if cells could reduce noise, the average steady-state growth would not increase. In that case, the higher than average growth rates due to noise, are possible by the build-up of energy or cellular components during a period of lower than average growth.

Fluctuations in expression capacity are not due to growth rate fluctuations

Much effort has been put in finding the underlying source of protein variation. So far, it has been shown that some part originates from noisy gene expression due to stochastic molecular interactions (binding and unbinding events) at the promoter of the gene ('intrinsic noise') [23]. Another noise source comes from the random partitioning of molecules at cell division [203]. Both these noise sources can be significant when the level of the protein is very low, but otherwise they are negligible. Surprisingly, the majority of the variation in protein levels is not gene specific, and is exhibited as a more cell-wide concerted phenomenon ('extrinsic noise') [23, 204, 205]. Some of this variation is due to cell aging and progression through the cell cycle stage [149, 204]. But the main factors responsible for this variation in 'expression capacity' ('extrinsic noise' as experienced by different promoters) have not been identified yet. However, it is likely to be related to differences in transcriptional and translational components, and possibly also to the cell's metabolic state [206]. Our results may aid in painting a better picture of noise sources in the cell.

Considering that *E. coli*'s dry weight consists of 55% protein [6], the growth rate of an individual cell might depend on similar factors as its expression capacity. Interestingly, in those experiments where *lac* levels were high, fluctuations in *lac* are presumably caused for the greater part by extrinsic noise [23]. Assuming that intrinsic noise was negligible, the variation in *lac* level would effectively correspond to the cell's expression capacity. But, although we found noise in growth and noise in protein level to correlate, this effect was quite weak for high *lac* levels, even when a time delay was taken into account (see Fig. 6.14). If both growth rate and expression capacity would largely depend on the same factors, we should have observed a very large correlation between the two. In other words: the large variation in growth rate correlates weakly with protein noise, and therefore is likely to depend on other factors than those responsible for variation in the extrinsic noise of *lac*.

When expression matters, the *lac* operon is not bistable

Soon after the discovery of gratuitous inducers, such as IPTG and TMG, a surprising phenomenon of the regulation of the *lac* operon was observed. When a little bit of inducer was added to cells that had not been exposed to the inducer before, an extremely heterogeneous population of cells developed [20]. Part of the population did not show any induction of the *lac* operon, while the other part showed a very high expression. This phenomenon, called ‘all-or-nothing’ behavior or ‘bistability’, is due to a auto-stimulation of *lac* induction: LacY proteins pump inducer into the cell, resulting in *lac* induction, hence more LacY proteins [20,207]. When cells have been unexposed to inducer, sporadic and stochastic expression of more than a hundred LacY proteins per cell, are required in order to turn the operon ‘on’ [21]. Note that high external inducer concentrations, result in sufficient passive diffusion over the membrane to induce the *lac* operon.

Interestingly, bistability of the *lac* system has been demonstrated with artificial inducers, TMG and IPTG, but never with the natural inducer (allo)lactose (see Fig. 6.2) [20, 22]. It is difficult to test for bistability on lactose, as the lactose concentration supplied needs to be so low, that it would rapidly be depleted due to consumption. Some studies have argued that the all-or-none behavior does not occur for growth on lactose [22, 208, 209]. Essentially, there are two reasons why the *lac* operon may not be bistable when cells are growing on lactose. First, in contrast to artificial inducers, lactose is not only imported by the LacY protein but also degraded by the LacZ protein, effectively counteracting the auto-stimulation of *lac* induction. Second, lactose does not only act as inducer, but its import and degradation eventually leads to cellular growth. Therefore, an increase of the LacY proteins, leads to an increase in growth, leading to a decrease of LacY proteins due to dilution, again counteracting the auto-stimulation.

In our experiments we observed all-or-none behavior for growth on glucose with 20 μ M IPTG within a single microcolony (data not shown). However, when cells were grown on lactulose, all-or-none induction was absent and graded responses were observed at all IPTG concentrations. In these experiments, the *lac* operon is still induced with a gratuitous inducer that is not degraded. But growth of the cell does depend on expression of the *lac* proteins. So, these results suggests that the *lac* operon is not bistable, when cells need *lac* expression for growth. It would be interesting to grow *E. coli* in a flow cell with extremely low concentrations of lactose. Such experiments could establish once and for all, whether expression of *lac* is bistable under natural conditions.

Consequences for modeling of metabolic networks

So far, most models of metabolic networks consider steady state conditions, where the average behavior of each cell remains constant over time. Such modeling efforts have been very successful in describing average population behavior [210–212]. However, now that single cells studies reveal large heterogeneities within population of cells, both in protein level and in growth rate, it can be questioned whether steady state models capture the behavior of metabolic networks correctly. Espe-

cially when fluctuations in growth and a single protein have a significant effect on each other, as we have shown in this study, the behavior of the metabolic network is likely not captured in a steady state model. Therefore, use of the interdependency between protein level and growth rate should be of great value to the modeling of dynamic out of equilibrium metabolic networks. Furthermore, our results of how cellular growth and protein level affect each other, can be used to describe the dynamical behavior of metabolic networks and growth.

An interesting example of how growth dependency on protein level can affect molecular network studies, is presented by Tănase-Nicola and ten Wolde [26]. They show that population quantities (which are generally measured experimentally) are not necessarily equal to the time average as experienced by a lineage of cells (which are generally modeled in molecular network studies). This effect is due to the fact that in a time snapshot of the population there is an overrepresentation of cells with levels optimal for growth, as they make more offspring due to their higher growth rate. Data from this study can be used to quantify the discrepancy between population and lineage quantities.

6.6 Future work

Capturing data in simple model

The experimental results presented in this chapter revealed complex interdependencies between fluctuations in cellular growth and protein levels. Although the observations can be understood qualitatively from known cellular processes, new insights could be obtained by capturing them quantitatively in a model. A dynamical model should at least include the cost and benefit of *lac* expression for cell growth, and the production and dilution of proteins due to growth (see Fig. 6.15). It will be interesting to see whether this minimal model can capture the experimentally observed noise levels, correlations and delay times, at different metabolic states.

A model that captures our results will allow predictions to be made about how protein levels and growth fluctuate under other conditions. But most importantly, it could help to explain what limits the growth rate of individual cells and is causing growth fluctuations. Other observations that might be understood from the behavior of the model are the significantly lower autocorrelation time of growth rate with respect to protein levels, and the absent delayed effect of growth on protein level when *lac* is limiting. Furthermore, it could test some hypotheses that are hard to approach experimentally. For example, does a fluctuating dilution rate cause homeostasis of proteins that are limiting for growth? By having the proteins in the model dilute with either a constant average rate, or a fluctuating rate dependent on cell growth, this question could be explored.

Protein production rate and regression analysis

Apart from a model to capture our experimental observations, additional analysis of our lineage data could perhaps provide new insights. In this study, we performed

analyses that are very powerful in finding statistical relevant correlations in noisy data. However, these analyses also have their limitations, as they do not distinguish between positive or negative deviations from the average, nor take dynamical information into account. If, for example, there would be a difference in how growth and protein level correlate depending on whether a fluctuation is higher or lower than the average level, this would remain hidden in our current correlation analysis. Such a difference is very well possible in our low induction experiments, where it is the question whether all other metabolic proteins responsible for growth on lactose, adjust to the new metabolic state. If they have adjusted completely, a decrease in *lac* level would result in an immediate reduction in flux, while an increase in *lac* level would result in a delayed flux increase, as all other components have to adjust first. The curved correlation cloud in Figure 6.8A hints that this might be the case, and a more sophisticated (regression) analysis might confirm this. In addition, some new insights might emerge from a frequency domain analysis on the lineage data [196,213].

Furthermore, it would be interesting to determine the *lac* production rate instead of the *lac* level in the cell. Whereas the growth rate of a cell should depend on the *lac* concentration and not on *lac* production, it is to be expected that the production rate more directly responds to the metabolic state, and perhaps the growth rate of the cell. Therefore, the way how protein level depends on growth rate, might better be understood by looking at the *lac* production rate, than at the *lac* level. Note, that protein production rate is not the same as the derivative of our currently determined protein concentration, as the latter also depends on increase of cell size.

How does noise affect the fitness of a population of cells?

Whether and how protein fluctuations affect the fitness of an organism, depends on how they affect a population of cells over time. Mathematical models predict that fluctuations can be both detrimental and beneficial for the average population growth rate [11,26]. When the average protein level has not been tuned to its environment and deviates sufficiently from the optimal one, noise can be favorable, as there is a larger benefit from the fraction of the population closer to the optimal protein level, than the loss from the fraction that is far away from it [26]. However, when protein levels have adapted to constant conditions, the absence of noise would lead to each individual being able to have optimal concentrations of all its components, resulting in a higher mean fitness [11]. As our decoupled *lac* system can provide both conditions, i.e. average protein expression at either optimal or non-optimal level, it could be very useful in experimental confirmations of these predictions. More specifically, it would be interesting to compare the average growth rates of mutants with different levels of protein noise [214] on both conditions.

Origin and function of growth rate fluctuations

As we noted before, it is unknown what process or which processes are causing the growth rate fluctuations in single cells. Being the ‘engine of the cell’, it would be interesting to construct a fluorescent reporter for ribosome levels, as their level might dictate growth rate. Another hypothesis poses that growth on minimal medium is limited by amino acid supply, rather than by ribosome capacity [215]. Therefore, it would be interesting to see whether single cell growth rate variations in *E. coli* decrease when the medium is supplemented with amino acids. We observed a lower autocorrelation time of the growth fluctuations when *E. coli* was grown on glucose. Therefore the systematic measurement of growth rate fluctuations on a range of sugars and media, could perhaps clarify the origin of the fluctuations.

6.7 Materials and methods

Strains

Growth experiments were performed using derivatives of *E. coli* MG1655 (*rph-1 ilvG- rfb-50*). For pre-experiment consumption of organic contaminants (see below), we used strain NCM520 ($\Delta lacAYZ$), obtained from the Coli Genetic Stock Center (CGSC). All measurements on *lac* expression were performed with strain AB460 ($\Delta lacA::gfp-cat$), which was created by Alex Böhm and kindly provided by Martin Ackermann. AB460 was constructed from MG1655 by replacing the *lacA* gene with *GFPmut2* [178] and chloramphenicol resistance using the protocol described by Datsenko and Wanner [216].

MG1655 is derived from the original K-12 strain which was obtained from ‘nature’: a stool sample of a diphtheria patient in Palo Alto, CA in 1922 [3]. During its historical derivation from K-12, MG1655 has been rid of the lambda bacteriophage and F plasmid [217], but also acquired some unwanted mutations relevant for growth experiments: *rph-1* is a 1 base pair deletion at the end of *rph*, resulting in decreased *pyrE* expression and a mild growth defect (about 10% in minimal media) due to internal uracil (pyrimidine) starvation [218]. The wild-type growth rate can be restored by supplementing the medium with uracil [218, 219]. *ilvG-* is a frameshift that knocks out acetohydroxy acid synthase II, resulting in growth inhibition by valine [220]. *rfb-50* is an IS5 insertion that results in the absence of O-antigen synthesis and sensitivity to phage P1 [221]. Potentially, the MG1655 derivative we use, also carries a large deletion around the *fnr* (fumarate-nitrate respiration) regulatory gene, resulting in growth defects in anaerobic respirations [219]. This *fnr-* mutant (now named CGSC 8003) was dispatched by the Coli Genetic Stock Center as MG1655 (CGSC 6300) until October 2003, when it was replaced with a *fnr+* MG1655. In our growth conditions only the *rph-1* mutation is relevant, for which reason we supplement our media with uracil (see below).

Media

Cells were grown in M9 minimal medium (47.7 mM Na₂HPO₄, 25 mM KH₂PO₄, 9.3 mM NaCl, 17.1 mM NH₄, 2.0 mM MgSO₄, 0.1 mM CaCl₂) supplemented with 0.2 mM uracil (see above). As carbon and energy source either 0.1% lactulose (=2.9 mM) or 0.1% glucose (=5.6 mM) was added. When indicated IPTG was added to the medium (0-200 μM). Suppliers of chemicals can be found in Table 6.1.

Cell growth in batch culture

Cells were grown in 50mL flasks in a 37°C water bath with vigorous shaking. First, TY medium was inoculated with cells from -80°C glycerol stock, and grown until 0.02-0.50 OD (optical density at 600 nm, 1 cm path length). Cells were then diluted back to have OD ≈ 0.01 in TY, and subsequently diluted into M9 medium with glucose or lactose at 1:200, 1:800 and 1:3200 for growth O/N. These flasks contained approximately 5 · 10², 10³ and 5 · 10³ cells ml⁻¹, respectively. The following day the O/N culture with the highest OD while still in exponential growth (OD < 0.20) was used to inoculate medium with the designated medium (containing IPTG and either glucose or lactulose). Cells were grown exponentially for at least another 17 hours, and if necessary diluted back again, until growth rate became stable.

This protocol allows reproducible cultivation of bacteria without cells being washed, nor reaching stationary phase. The initial rich media is both diluted out and eaten up by the cells. When growing *E. coli* on lactulose with low IPTG (and thus low expression levels), constitutive *lac* mutants quickly take over the population. Therefore care must be taken to only grow cells in this medium for as short as

chemical	supplier, product number	alternative description
Na ₂ HPO ₄ ·2H ₂ O	Merck, 1.06580	disodium hydrogen phosphate dihydrate
KH ₂ PO ₄	Merck, 1.04873	potassium dihydrogen phosphate
NaCl	Merck, 1.16224	sodium chloride
NH ₄ Cl	Merck, 1.01145	ammonium chloride
MgSO ₄ ·7H ₂ O	Prolabo, 1.05886	magnesium sulfate heptahydrate
CaCl ₂ ·2H ₂ O	Merck, 1.02382	calcium dichloride dihydrate
uracil	Sigma, 41128	2,4-pyrimidinediol
glucose·H ₂ O	Merck, 1.08342	D-(+)-glucose monohydrate
lactose·H ₂ O	Fluka, 61339	D-(+)-lactose monohydrate
lactulose	Fluka, 61360	4-O-β-D-galactopyranosyl-β-D-fructofuranose
IPTG	Sigma, I5502	isopropyl-β-D-1-thiogalactopyranoside
ONPF	Sigma, N3253	2-nitrophenyl-β-D-1-fucopyranoside
acrylamide	Bio-Rad, 161-0148	40% acrylamide / bis solution 37.5 : 1
ammonium persulfate	Sigma, A-9164	(NH ₄) ₂ S ₂ O ₈
TEMED	Bio-Rad, 161-0800	tetra-methyl-ethylene-diamine
repe silane	Amersham, 17-1332-01	2% Si(CH ₃) ₂ Cl ₂ in octa-methyl-cyclo-octasilane
agarose	Roche, 11.388.991.001	multipurpose agarose
low melt agarose	Sigma, A4018	low gelling temperature agarose type VII

Table 6.1: Details of chemicals used in this study. Shown are the chemical names used in main text, together with their supplier and product number, followed by a more detailed description of the chemical.

possible.

Polyacrylamide gel pads

In order to force cells to grow into a single plane microcolony, they need to be confined. Traditionally this is accomplished by growing them between a glass coverslip and a solid matrix pad of agarose or gelatin (eg. [149]). Sugar-based gels, like agarose and gelatin, are ill-suited for growth experiments where carbon sources or carbon intake are varied. Therefore, polyacrylamide (PA) gels were employed, whose polymer subunits cannot be used by *E. coli* as a carbon source.

PA gels were made by mixing 1.25 ml 40% acrylamide, 3.7 mL water, 50 μ L fresh 10% ammonium persulfate, and 5 μ L TEMED together, after which 900 μ L was poured into a silanized cavity glass slide (see Fig. 6.16 and Table 6.1). The slide was immediately covered with a 24 mm x 60 mm coverslip and left to polymerize at room temperature for about half an hour. After removal of the coverslip, the gel was cut in pieces of 5 mm x 10 mm, and transferred to a flask with water. To get rid of unpolymerized chemicals, the gels were washed by transferring them repeatedly to fresh water. During preparation and storage, all glassware and chemicals were kept sterile, to avoid growth of cell contaminations sticking to the gel. Although the PA gels are safe to handle, note that unpolymerized acrylamide is toxic and can be absorbed through the skin.

The day before an experiment, a piece of gel was transferred several times to large volumes of the designated media. For experiments with growth on lactulose, the polyacrylamide gel was pre-grown with *lac*⁻ cells to remove contaminants (see page 76).

Apart from not being sugar-based, PA gels have several other advantages over traditional agarose gels. First of all, PA gels are quite firm, and therefore easy to handle. This allows disassembling sample slides without tearing the gel, and maneuvering of different gels into the same sample holder. Second, after washing, the PA gels have increased a fraction in size, which allows firm clamping in between the glass slides in the sample holder. Furthermore, PA gels can be stored for weeks, have good wettability, and *E. coli* microcolonies often grow with more separation between cells compared to standard agarose gels.

Time-lapse microscopy

For microscopy, cells were diluted back to OD \approx 0.005 and transferred to a microscope incubation chamber (Solent), allowing precise 37°C temperature control. 1 μ L of culture was applied on a small pre-warmed PA gel inside a glass cavity slide and sealed with a pre-warmed coverslip (see Fig. 6.16A). The sample slide, containing a large volume of oxygen, was placed in a metal clamp to ensure tight sealing (see Fig. 6.16B).

Imaging was performed with an inverted microscope (Nikon, TE2000), equipped with 100X oil objective (Nikon, Plan Fluor NA 1.3), cooled CCD camera (Photometrix, CoolSnap HQ), xenon lamp with liquid light guide (Sutter, Lambda LS), GFP filter set (Chroma, 41017), computer controlled shutters (Sutter, Lambda

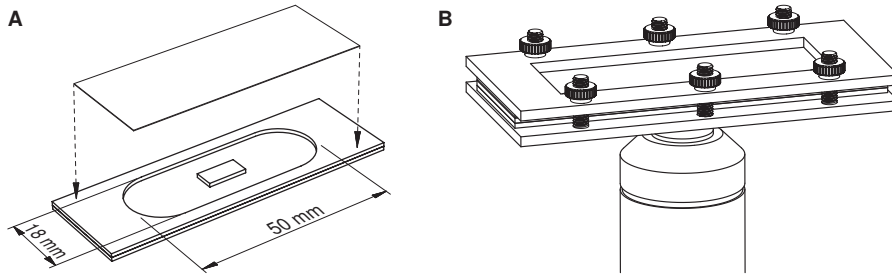


Figure 6.16: Sample holder for time-lapse microscopy

(A) Glass microscope slide with the center part cut out (manufactured in-house) was stuck on top of a normal slide using a little amount of silicon grease (Dow Corning) to form a cavity glass slide. Such slides were used to polymerize gels, or, as shown here, to hold a PA gel piece with cells. A coverslip would close off the sample chamber.

(B) To keep the sample chamber closed, it was put upside down in a metal clamp (manufactured in-house) and placed on the microscope.

10-3 with SmartShutter) and automated stage (Märzhäuser, SCAN IM 120 x 100). An additional intermediate 1.5X magnification was used, resulting in images with pixel size corresponding to a length of 41 nm.

Up to 9 of the very sparse cells on the gel were manually selected, and followed for up to 30 hours using MetaMorph microscope control software. Care was taken to have positions properly spaced, to avoid cells being affected by the illumination of neighboring positions. Phase contrast images (200 ms exposure time with GIF filter, 3 images $-0.2 \mu\text{m}$, $0 \mu\text{m}$ and $+0.2 \mu\text{m}$ offset from focus) were recorded with exact intervals every 1.5 - 4 min. Fluorescence images were recorded every 11.5 - 28 min, with 2x2 binning and either 500 ms or 1000 ms exposure. The low exposure frequency ensured that photodamage and photobleaching was negligible during the experiment. To correct for focus drift, MetaMorph's adjust focus function was used on the center of 2x2 binned phase contrast images, before each image was taken.

Cell segmentation and tracking

Images were analyzed offline using custom Matlab (MathWorks) programs. The analysis consisted of 3 parts: outlining and tracking of cells, fluorescence extraction, and length measurement.

The outline of cells was determined by first averaging the 3 phase contrast images. From this image, edges were determined using a Laplacian of Gaussian filter. In essence, this algorithm looks for maxima and minima in the derivative of the image. Next the background was separated from the cells, and clumps of cells were cut based on concavity and phase contrast maxima (see Fig. 6.17).

Although these automated analysis steps functioned properly for over 90% of the cells, some cells were not identified, or segmented incorrectly. Especially at the

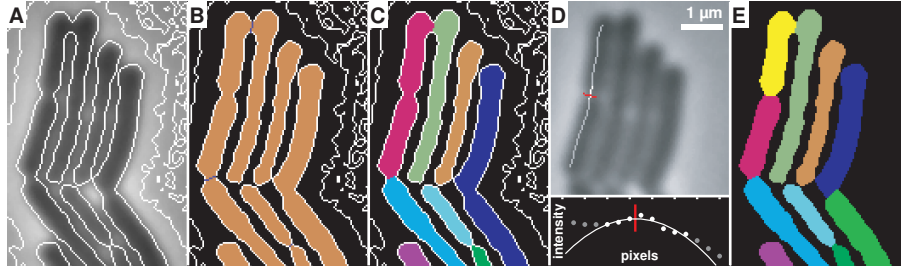


Figure 6.17: Segmentation of cells using phase contrast images

(A) Phase contrast image of bacterial cells with an overlay of its Laplacian of Gaussian filter (white pixels).

(B) Once cells are distinguished from background, obvious separations between cells are determined by finding concave points on opposite sides along the edges of cells. Cells are segmented by a straight line between these points (blue pixels).

(C) Individual cells can now be distinguished (different colors), but not all cell separations are found.

(D) Remaining cell clumps are separated by finding optima (red line) in the phase contrast values along a line through the cell clump. The line is obtained by morphological thinning of the segmented area, and only those optima near cell areas with concave edges, or user-defined locations, are accepted.

(E) The resulting segmentation image, with each cell drawn in a different color.

end of experiments, with high numbers of cells and little separation between them, considerable manual intervention was required. The outline of these problematic cells were determined by forcefully running the same analysis procedures at user-defined locations.

Once each frame was segmented, lineages of cells were traced by a simple tracking algorithm that searches for nearby cells in successive frames. Although the increase in size of the microcolony was taken into account, ~ 10 frames per cell generation were required for proper tracking, as cells move unpredictably within an expanding microcolony. A complete history of each cell lineage in the microcolony was unambiguously determined up until the microcolony expanded beyond the field of view, or when a second layer of cells would form. Generally this meant 9 generations of lineages, consisting of ~ 1000 complete cell cycles from a single microcolony. After segmentation and tracking it was straightforward to a cell's location, its orientation within the microcolony, and the age of its cell poles [222].

Cell fluorescence

Before the fluorescence for a particular cell was extracted, the fluorescence image was corrected for imaging artifacts. First a background image (I_b) was subtracted, correcting for sensor noise and background signal. Next, the image was normalized by a shading image (I_s), correcting for uneven illumination of the sample. Given the

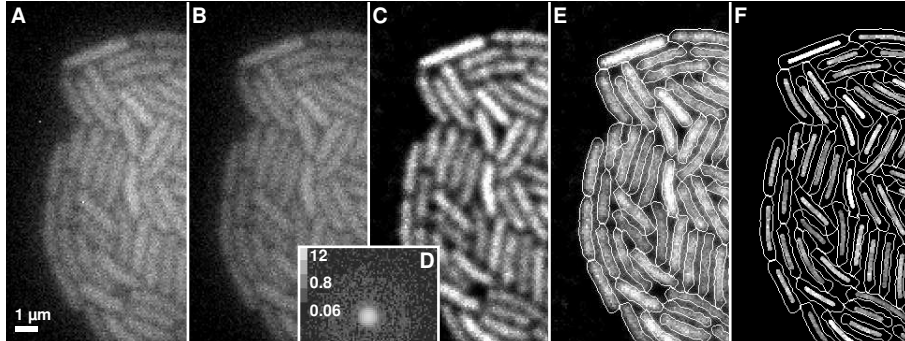


Figure 6.18: Determination of average protein levels from fluorescence images

(A) The original fluorescence image.

(B) Fluorescence image after background correction (pixel specific for camera) and shading correction (pixel specific for complete imaging system). Note that cells at the edge of the microcolony seem less fluorescent as they get less scatter fluorescence from adjoining cells.

(C) The fluorescence image after deconvolution, which probabilistically reassigns each imaged photon to its original pixel position [223].

(D) Image of part of the point spread function (PSF) used for the deconvolution. Note that the color scale is logarithmic and values indicate 10^{-3} times the weight of each pixel. The PSF was determined by imaging fluorescent microspheres of subpixel size. Multiple images from different microspheres were aligned and averaged.

(E) Using the cell segmentation (white outlines, determined from the phase contrast images) the total cell fluorescence can be extracted. Note that cells at the edge of the microcolony are segmented as thicker cells, hence having more background pixels within their perimeter.

(F) In order to get an accurate mean fluorescence value for each cell, only pixels within a box of fixed width inside the cell perimeter were averaged. In order to get the box on top of the center of the cell, the box was allowed to move a few pixels until it found the maximum mean fluorescence.

original image I , the calibrated output image I_c is given by:

$$I_c = \frac{I - I_b}{I_s - I_b}$$

Both the background and the shading image were determined by averaging tens of images made under experimental conditions.

The fluorescence image had a small alignment offset with the phase contrast image. This was corrected by finding the best fit of fluorescence on top of segmented cells, and shifting the fluor image accordingly. The last correction of the fluorescence image was a deconvolution step to correct the blur produced by the imaging system. For this Matlab's Lucy-Richardson algorithm was used with an experimentally determined point-spread function.

Once this corrected image was obtained, the total fluorescence of a cell could be determined by extracting those pixels that were within the cell outline, as determined from the phase contrast image. In order to get the mean fluorescence of a cell, the total fluorescence needs to be divided by the size of the cell. The size mea-

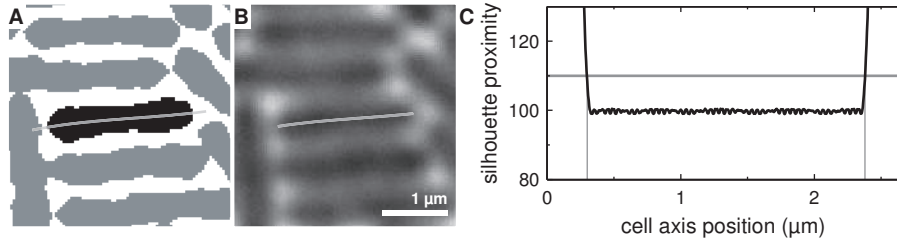


Figure 6.19: Single cell length measurement

(A) The axis of a cell is determined by fitting a third degree line through the silhouette (segmentation) of the cell.

(B) The axis between the cell poles plotted on top of the phase contrast image of the cell. The length of a cell is the length of the cell axis between these cell poles, which are determined in panel C.

(C) The poles along the cell axis are determined by first measuring the silhouette proximity along the axis. The silhouette proximity is defined as the total distance of the closest 25 segmentation pixels. Within the cell silhouette the total distance consistently remains 100 μm , but near the cell poles it rapidly increases. The location of the cell poles were taken at a silhouette proximity of 110 μm .

surement, however, is also used for the determination of the elongation rate, with which the mean fluorescence is correlated later.

In order to have an independent measurement of mean fluorescence, we resorted to the following method. A box with fixed width of 0.4 μm , beginning and ending 0.3 μm from the cell poles, was aligned on top of each cell (see Fig. 6.18). The mean fluorescence (F) for the complete cell was determined as the mean of the pixels within this box (F_{box}) with subtraction of the background fluorescence (F_{b} , determined from pixels outside microcolony) and the cell autofluorescence (F_{a} , determined from fluorescence of MG1655 cells without any GFP). The mean fluorescence obtained in this way, was normalized such that images with different exposure times (t_{e} in ms) and resolution (θ in $\mu\text{m} \cdot \text{pixel}^{-1}$) could be compared:

$$F = \frac{F_{\text{box}} - F_{\text{b}} - F_{\text{a}}}{t_{\text{e}} \theta^2}$$

Cell length

During its cell cycle, *E. coli* cells increase with only $\sim 2 \mu\text{m}$ in length. For the determination of growth rate differences within the cell cycle, it is therefore necessary to have a high precision length measurement. Although the pixel size in our optical system corresponds to a length of $\sim 0.04 \mu\text{m}$, it's not straightforward to obtain such precision in length measurements.

An important limitation lies in the optical resolution of imaging systems (shortest distance that can still be distinguished as separate entities). In light microscopy, this resolving power is limited to $\sim 0.25 \mu\text{m}$, as it depends on the numerical aperture of available objectives and the wavelength of applicable light. The diffraction

causing this limited resolution, results in fuzzy looking cell edges in phase contrast images (see Fig. 6.19). Fortunately, subresolution precision can still be obtained by careful analysis of phase contrast intensity profiles along the cell edges (see Fig. 6.19 and [224]). Although this works very well for single cells, applying this technique to cells in microcolonies presents some difficulties.

In phase contrast microscopy, differences in density and composition within a sample can be measured as the light interacts with the sample it passes through. Unfortunately, this means that the absolute phase contrast intensity at any point is highly dependent on the surrounding of that point, as the light will also pass through there. Within a microcolony this poses a problem, as the surrounding of cells can be quite different. Some cells are positioned freely (especially at the periphery), whereas others are squeezed together.

Another difficulty arises in experiments with slow growing cells. The shape of *E. coli* cells can be dramatically altered by growing them in confined spaces [225, 226]. Only very minor changes in shape are generally observed during growth in microcolonies. However, when cells are growing very slowly, we observe many cells growing into a curved shape.

The precision of length measurements in microcolonies depends largely on dealing with these difficulties. In order to obtain an accurate length under these conditions, we applied the following approach. The outline of every cell is determined (during segmentation) by an edge detection that is essentially based on finding the steepest parts in intensity profiles. Next, the cell axis is determined, by fitting a third degree polynomial $f(x)$ through the silhouette of the cell:

$$f(x) = ax^3 + bx^2 + cx + d$$

We define the length of the cell, as the length of this axis from one pole (x_0) to the other (x_1). The exact pole positions along the axis are determined by taking into account the cap of the cell (see Fig. 6.19). The cell length (L) can then be obtained by numerical integration between pole x_0 and x_1 of the cell:

$$L = \int_{x_0}^{x_1} \sqrt{1 + f'(x)^2}$$

Judging from the smoothness of length traces (see Fig. 6.6) the automated length measurements within microcolonies had a precision around the pixel size of $\sim 0.04 \mu\text{m}$.

Single cell growth rate

As no significant changes in cell width (except those due to changing curvature of the cell) are observed during growth experiments, we assume that cell size is proportional to cell length. In that case, single cell growth rate can be determined by obtaining the cell elongation rate. Although there have been some observations of *E. coli* cells growing bilinearly [188, 227], most studies support exponential elongation of cells [228, 229]. This fits with our own observations, so we determined the elongation rate by an exponential fit of length over time:

$$L(t) = L_0 e^{\mu_e \cdot \ln(2) \cdot t} = L_0 2^{\mu_e \cdot t}$$

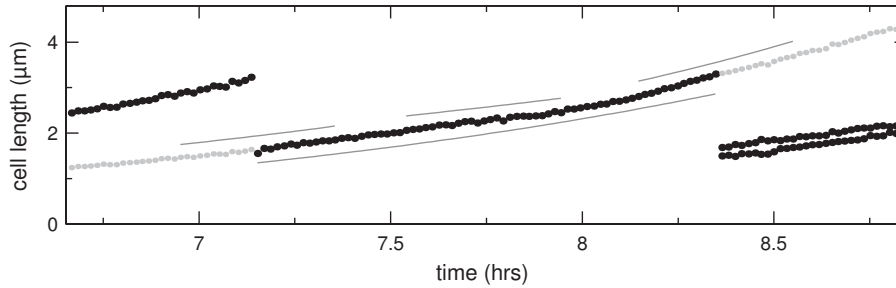


Figure 6.20: Growth rate of a single cell

The length of a single cell, its parent and its offspring plotted over time (dark circles). The average elongation rate of the cell can be determined by fitting an exponential through all its length measurements over time (grey fit below circles). Elongation rates of the cell at particular time points are determined by fitting data within a particular window around that time point (grey fits above circles). At the beginning and end of each cell cycle, length data of the parent or the offspring are used (grey circles, see main text). Note that the fits are horizontally shifted for clarity.

where the elongation rate μ_e is effectively the same as the doubling rate (see Fig. 6.20). Using a power with base 2, instead of the widely used natural base e , allows easy computation of doubling time T_d .

$$T_d = \frac{1}{\mu_e}$$

The elongation rate can either be determined for a single cell and its complete cell cycle, or alternatively for time points within its cell cycle. In the latter case a window needs to be specified around the time point of interest. The definition of this window presents some constraints. The size of the time window determines how many datapoints used for elongation rate determination are shared between neighboring time points. Therefore, for independent measurements of growth rate with a high time resolution, a small window is required. However, an accurate determination of the elongation rate also requires a length increase of the cell that is significantly larger than the precision of its length measurement. In other words: in a timespan with little cell growth, the error in the length measurement dominates the elongation rate. Given an average birth length L_b and the minimal required length increase L_i , the required time window T_w is given by:

$$T_w = T_d * \log_2\left(\frac{L_b + L_i}{L_b}\right)$$

We set the minimal length increase required for our time window to $0.5 \mu\text{m}$. With an average birth length of $2 \mu\text{m}$, this corresponds to a window of approximately a third of the cell's doubling time. Note that the time window is determined independently for each experiment, as the average doubling times differ up to 4 fold between experiments.

Retrieving length data in the window around time points at the beginning or at the end of a cell cycle is problematic, as the window expands beyond the cell's life. Neglecting unavailable datapoints within the window, results in large errors in determination of the elongation rate. An alternative, where the window is moved until it fits the cell's datapoints, results in constant elongation rates at $\frac{1}{6}$ part of the beginning and $\frac{1}{6}$ part of the end of the cell cycle. Therefore, we applied a technique where we extend the datapoints for a cell's length using the length of the mother (L_m), sister (L_s) and daughter cells (L_{d1} & L_{d2}) (see Fig. 6.20)

At the end of a cell's cycle additional data points are added by summation of the lengths of the two daughter cells. At the beginning of a cell's cycle points are added by using a fraction of the length of the mother cell. Although division in *E. coli* is quite symmetric, there are still some length differences between sister cells. Therefore not exactly half of the mother cell length was taken, but a ratio based on their birth lengths. The extended length data for a cell is thus defined by:

$$L^*(t) = \begin{cases} \frac{L_0}{L_0 + L_{0,sister}} L_m(t) & \text{if } t < t_0 \\ L(t) & \text{if } t_0 \leq t \leq t_1 \\ L_{d1}(t) + L_{d2}(t) & \text{if } t > t_1 \end{cases}$$

Cross-correlation

In order to correlate variation in expression level (F) and growth rate (μ_e) over time, we first extract a discrete time signal for each single cell lineage (see Fig. 6.6 and Fig. 6.21). Each lineage has N data values which are separated by sampling interval Δt , such that data value n originates from time point $t = n * \Delta t$. We define the noise in the lineage signals as the difference between the signal and the population mean:

$$\epsilon_n = F_n - \frac{1}{M} \sum_{m=1}^M F_{n,m} \quad \mu_n = \left(\frac{dL}{dt}\right)_n - \frac{1}{M} \sum_{m=1}^M (\mu_e)_{n,m}$$

where M is the total number of cells at each time point from all lineages. The cross covariance between ϵ and μ within a single lineage at time-lag $r\Delta t$, is then defined by [195]:

$$C_{\epsilon\mu}(r\Delta t) = \begin{cases} \frac{1}{N-r} \sum_{n=1}^{N-r} (\epsilon_n \mu_{n+r}) & \text{if } r \geq 0 \\ C_{\mu\epsilon}(-r\Delta t) & \text{if } r < 0 \end{cases} \quad (6.1)$$

In order to get the cross-correlation, $R_{\epsilon\mu}$, the covariance is normalized by the standard deviation, σ , of the signals:

$$R_{\epsilon\mu}(r\Delta t) = \frac{C_{\epsilon\mu}(r\Delta t)}{\sigma_\epsilon \sigma_\mu} \quad , \text{ with}$$

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2} \quad \bar{x} = \frac{1}{N} \sum_{n=1}^N (x_n)$$

As the mean of ϵ and μ are 0, their standard deviation is defined by their autocovariance at $r = 0$, giving cross-correlation:

$$R_{\epsilon\mu}(r\Delta t) = \frac{C_{\epsilon\mu}(r\Delta t)}{\sqrt{C_{\epsilon\epsilon}(0)C_{\mu\mu}(0)}} \quad (6.2)$$

In order to get the autocorrelation of either protein level or growth rate, the autocovariance is calculated, which is a special case of equation 6.1 where the records coincide:

$$C_{\epsilon\epsilon}(r\Delta t) = \frac{1}{N-|r|} \sum_{n=1}^{N-|r|} (\epsilon_n \epsilon_{n+|r|})$$

Note that an alternative formula for the covariance is sometimes used (see e.g. [196]), which does not correct for decreasing number of datapoints at higher $r\Delta t$:

$$C_{\epsilon\mu}(r\Delta t) = \begin{cases} \sum_{n=1}^{N-r} (\epsilon_n \mu_{(n+r)}) & \text{if } r \geq 0 \\ C_{\mu\epsilon}(-r\Delta t) & \text{if } r < 0 \end{cases}$$

At higher $r\Delta t$, there is less data available to calculate a reliable correlation value. The resulting increase in the error can be masked by this formula, as it constrains the cross-correlation to approach 0 at higher $r\Delta t$. Here, however, we use equation 6.1 instead, which is more accurate, but can give unreliable estimates at high $r\Delta t$.

Composite cross-correlation from branched data

Equation 6.1 estimates the cross covariance of the process underlying data from a single lineage. A better estimate can be obtained by combining data from multiple lineages. The composite cross covariance of M lineages is defined by:

$$C_{\epsilon\mu}^M(r\Delta t) = \begin{cases} \frac{1}{M} \frac{1}{N-r} \sum_{m=1}^M \left[\sum_{n=1}^{N-r} (\epsilon_{n,m} \mu_{(n+r),m}) \right] & \text{if } r \geq 0 \\ C_{\mu\epsilon}^M(-r\Delta t) & \text{if } r < 0 \end{cases} \quad (6.3)$$

As our lineages are extracted from a branched data set, many pairs of points are used multiple times (see lineages IV and V in Fig. 6.21). In order to get a composite cross-correlation that best estimates the real underlying process, it's important to only use comparisons between unique pairs of points. We can correct for multiple contributions of the same pair of datapoints, by weighing each pair based on the number of lineages the datapoints are used in (λ):

$$w_{n,m,r} = \frac{1}{\lambda_{(n+r),m}} \quad (6.4)$$

This results in the branch-corrected composite cross covariance:

$$C_{\epsilon\mu}^M(r\Delta t) = \begin{cases} \frac{1}{w_{tot}(r)} \sum_{m=1}^M \left[\sum_{n=1}^{N-r} (\epsilon_{n,m} \mu_{(n+r),m} w_{n,m,r}) \right] & \text{if } r \geq 0 \\ C_{\mu\epsilon}^M(-r\Delta t) & \text{if } r < 0 \end{cases} \quad (6.5)$$

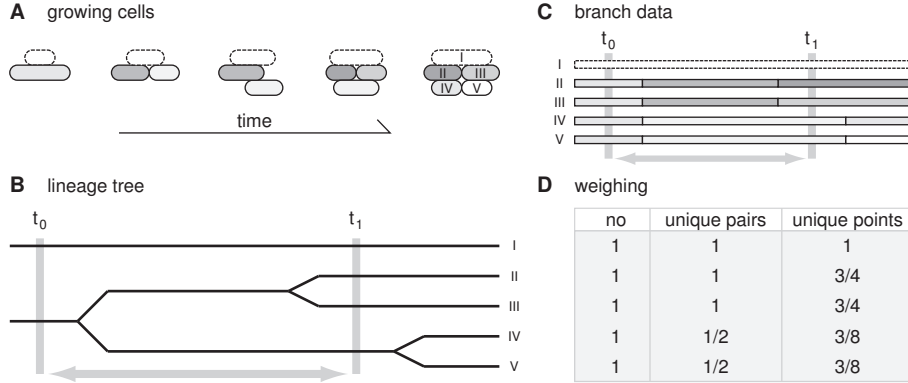


Figure 6.21: Extracting and weighing lineages from a branched data set

(A) Depiction of a growing microcolony over time, starting with 2 cells on the left and growing into 5 cells on the right.

(B) A lineage tree of the data shown in A. The tree starts with two lines (left), indicating the two starting cells, and at each division the line splits, resulting in five cells at the end (right).

(C) Five lineages can be extracted from the data. Note that most lineages share part of their data. When correlating data points from t_0 with t_1 , one pair consists from completely independent data points (lineage I). Two lineages provide exactly the same pairs of data points (lineages IV and V), and two lineages only share a data point at t_0 (lineages II and III).

(D) Different types of weighing for the correlation of data points from t_0 with t_1 . Equation 6.3 does not apply any weighing scheme, such that weighing of each lineage is set to one. Only comparisons between unique data pairs are obtained by weighing with equation 6.4. Lineages II and III are not completely independent, which can be corrected for by weighing with equation 6.6.

$$w_{tot}(r) = \sum_{m=1}^M \sum_{n=1}^{N-r} w_{n,m,r}$$

This branch correction weighs each comparison between two unique data points only once (similar to Dunlop et. al. [181]). However, there is an additional issue involving the dependency of data points in a branched data set, that is not addressed with this weighing correction. Although each comparison between data points is unique now, there are many occurrences where a single datapoint is used in pairs with many other data points (see lineages II and III in Fig. 6.21). These combinations are weighed equally compared to pairs of totally independent data points. A possible way to correct for this is to weigh reused datapoints for 50% (and thus their pairs for 75%):

$$w_{n,m,r}^* = \begin{cases} \frac{1}{\lambda_{(n+r),m}} & \text{if } \lambda_{n,m} = 1 \\ \frac{3}{4 \cdot \lambda_{(n+r),m}} & \text{if } \lambda_{n,m} > 1 \end{cases} \quad (6.6)$$

This latter weighing method was applied in this study.

Bibliography

- [1] Schaechter M, Gorbach SL (1996) The intestinal ecology of *Escherichia coli* revisited. *ASM News* **62**:304.
- [2] Koch A (1987) Phosphate metabolism and cellular regulation in microorganisms., Washington, D.C.: ASM Press. pp. 300-305.
- [3] Bachmann B (1996) *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology., Washington, D.C.: ASM Press. pp. 2460-2488.
- [4] Purcell EM (1977) Life at low Reynolds number. *Am J Phys* **45**:3-11.
doi:10.1119/1.10903
- [5] Schrödinger E (1944) What is life? Cambridge: Cambridge Univ. Press.
- [6] Ingraham JL, Maaløe O, Neidhardt FC (1983) Growth of the Bacterial Cell. Sunderland, Massachusetts: Sinauer Associates Inc.
- [7] Müller-Hill B (1996) The lac Operon. Berlin: Walter de Gruyter.
- [8] Koch AL (1983) The protein burden of lac operon products. *J Mol Evol* **19**:455-462.
doi:10.1007/BF02102321
- [9] Harder W, Dijkhuizen L (1982) Strategies of mixed substrate utilization in microorganisms. *Philos Trans R Soc Lond, B, Biol Sci* **297**:459-480.
doi:10.1098/rstb.1982.0055
- [10] Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**:588-592.
doi:10.1038/nature03842
- [11] Kalisky T, Dekel E, Alon U (2007) Cost-benefit theory and optimal design of gene regulation functions. *Phys Biol* **4**:229-245.
doi:10.1088/1478-3975/4/4/001
- [12] Weickert MJ, Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem* **267**:15869-15874.
<http://www.jbc.org/content/267/22/15869.full.pdf>
- [13] Swint-Kruse L, Matthews KS (2009) Allosterity in the LacI/GalR family: variations on a theme. *Curr Opin Microbiol* **12**:129-137.
doi:10.1016/j.mib.2009.01.009
- [14] Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, Müller-Hill B (1990) Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. *EMBO J* **9**:615-621.
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC551714/>
- [15] Nguyen CC, Saier MH (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett* **377**:98-102.
doi:10.1016/0014-5793(95)01344-X
- [16] Pauling L, Zuckerkandl E (1963) Chemical paleogenetics; molecular "restoration studies" of extinct forms of life. *Acta Chem Scand* **17**:S9-S16.
doi:10.3891/acta.chem.scand.17s-0009
- [17] Campbell JH, Lengyel JA, Langridge J (1973) Evolution of a second gene for beta-galactosidase in *Escherichia coli*. *Proc Natl Acad Sci USA* **70**:1841-1845.
doi:10.1073/pnas.70.6.1841
- [18] Poelwijk FJ, Heijning P, de Vos MGJ, Kiviet DJ, Tans SJ (2010) Optimality and the evolution of transcriptionally regulated gene expression. Submitted to *BMC Syst. Biol.*

Bibliography

- [19] Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* **240**:421-433. doi:10.1006/jmbi.1994.1458
- [20] Novick A, Weiner M (1957) Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci USA* **43**:553-566. doi:10.1073/pnas.43.7.553
- [21] Choi PJ, Cai L, Frieda K, Xie XS (2008) A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**:442-446. doi:10.1126/science.1161427
- [22] Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. *Nature* **427**:737-740. doi:10.1038/nature02298
- [23] Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* **297**:1183-1186. doi:10.1126/science.1070919
- [24] Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**:358-362. doi:10.1038/nature04599
- [25] Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science* **307**:1962-1965. doi:10.1126/science.1106914
- [26] Tanase-Nicola S, ten Wolde PR (2008) Regulatory control and the costs and benefits of biochemical noise. *PLoS Comput Biol* **4**:e1000125. doi:10.1371/journal.pcbi.1000125
- [27] Darwin C (1859) *On the Origin of Species by Means of Natural Selection*. London: Murray.
- [28] Smith JM (1970) Natural selection and the concept of a protein space. *Nature* **225**:563-564. doi:10.1038/225563a0
- [29] Malcolm BA, Wilson KP, Matthews BW, Kirsch JE, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**:86-89. doi:10.1038/345086a0
- [30] Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA (1990) The ribonuclease from an extinct bovid ruminant. *FEBS Lett* **262**:104-106. doi:10.1016/0014-5793(90)80164-E
- [31] Ugalde JA, Chang BSW, Matz MV (2004) Evolution of coral pigments recreated. *Science* **305**:1433. doi:10.1126/science.1099597
- [32] Thornton JW (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* **5**:366-375. doi:10.1038/nrg1324
- [33] Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Int Cong Genet* **1**:356-366. <http://www.blackwellpublishing.com/ridley/classictexts/wright.asp>
- [34] Gillespie JH (1991) *The Causes of Molecular Evolution*. Oxford: Oxford Univ. Press.
- [35] Kauffman SA (1993) *The Origins of Order: Self-organization and Selection in Evolution*. Oxford: Oxford Univ. Press.
- [36] Gavrillets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton: Princeton Univ. Press.
- [37] van Nimwegen E, Crutchfield JP (2000) Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull Math Biol* **62**:799-848. doi:10.1006/bulm.2000.0180

-
- [38] Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**:1165-1174.
doi:10.1111/j.0014-3820.2005.tb01768.x
- [39] 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.
doi:10.1371/journal.pcbi.0020058
- [40] Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* **47**:713-719.
<http://www.genetics.org/cgi/reprint/47/6/713>
- [41] Weinreich DM, Delaney NE, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**:111-114.
doi:10.1126/science.1123539
- [42] DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**:678-687.
doi:10.1038/nrg1672
- [43] Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* **103**:5869-5874.
doi:10.1073/pnas.0510098103
- [44] Hurley JH, Dean AM, Koshland DE, Stroud RM (1991) Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and NADP+ complexes. *Biochemistry* **30**:8671-8678.
doi:10.1021/bi00099a026
- [45] Hurley JH, Chen R, Dean AM (1996) Determinants of cofactor specificity in isocitrate dehydrogenase: structure of an engineered NADP+ → NAD+ specificity-reversal mutant. *Biochemistry* **35**:5670-5678.
doi:10.1021/bi953001q
- [46] Kalodimos CG, Bonvin AMJJ, Salinas RK, Wechselberger R, Boelens R, et al. (2002) Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA-binding domain. *EMBO J* **21**:2866-2876.
doi:10.1093/emboj/cdf318
- [47] Kopke Salinas R, Folkers GE, Bonvin AMJJ, Das D, Boelens R, et al. (2005) Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. *Chem-biochem* **6**:1628-1637.
doi:10.1002/cbic.200500049
- [48] Koradi R, Billeter M, Wäijthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**:51-5, 29-32.
doi:10.1016/0263-7855(96)00009-4
- [49] Lunzer M, Miller SP, Felsheim R, Dean AM (2005) The biochemical architecture of an ancient adaptive landscape. *Science* **310**:499-501.
doi:10.1126/science.1115649
- [50] Zhu G, Golding GB, Dean AM (2005) The selective cause of an ancient adaptation. *Science* **307**:1279-1282.
doi:10.1126/science.1106974
- [51] Bridgman JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**:97-101.
doi:10.1126/science.1123348
- [52] Barkai N, Leibler S (1997) Robustness in simple biochemical networks. *Nature* **387**:913-917.
doi:10.1038/43199
- [53] Kirschner M, Gerhart J (1998) Evolvability. *Proc Natl Acad Sci USA* **95**:8420-8427.
doi:10.1073/pnas.95.15.8420
- [54] Kitano H (2004) Biological robustness. *Nat Rev Genet* **5**:826-837.
doi:10.1038/nrg1471
-

Bibliography

- [55] Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J (2004) Robustness of cellular functions. *Cell* **118**:675-685.
doi:10.1016/j.cell.2004.09.008
- [56] Thattai M, van Oudenaarden A (2004) Stochastic gene expression in fluctuating environments. *Genetics* **167**:523-530.
doi:10.1534/genetics.167.1.523
- [57] Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309**:2075-2078.
doi:10.1126/science.1114383
- [58] Arnold FH, Wintrobe PL, Miyazaki K, Gershenson A (2001) How enzymes adapt: lessons from directed evolution. *Trends Biochem Sci* **26**:100-106.
doi:10.1016/S0968-0004(00)01755-2
- [59] Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* **4**:457-469.
doi:10.1038/nrg1088
- [60] Couñago R, Chen S, Shamoo Y (2006) In vivo molecular evolution reveals biophysical origins of organismal fitness. *Mol Cell* **22**:441-449.
doi:10.1016/j.molcel.2006.04.012
- [61] Lenski RE, Travisano M (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci USA* **91**:6808-6814.
doi:10.1073/pnas.91.15.6808
- [62] Stephens SG (1951) Possible significance of duplication in evolution. *Adv Genet* **4**:247-265.
doi:10.1016/S0065-2660(08)60237-0
- [63] Ohno S (1970) Evolution by Gene Duplication. New York: Springer-Verlag.
- [64] Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151-1155.
doi:10.1126/science.290.5494.1151
- [65] Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. *Nat Genet* **36**:492-496.
doi:10.1038/ng1340
- [66] Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**:1234-1244.
doi:10.1093/nar/gkg210
- [67] Bray D, Lay S (1994) Computer simulated evolution of a network of cell-signaling molecules. *Biophys J* **66**:972-977.
doi:10.1016/S0006-3495(94)80878-1
- [68] Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA* **101**:580-585.
doi:10.1073/pnas.0304532101
- [69] Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**:101-113.
doi:10.1038/nrg1272
- [70] Sengupta AM, Djordjevic M, Shraiman BI (2002) Specificity and robustness in transcription control networks. *Proc Natl Acad Sci USA* **99**:2072-2077.
doi:10.1073/pnas.022388499
- [71] 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.
doi:10.1126/science.298.5594.824
- [72] Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**:119-124.
doi:10.1098/rspb.1994.0058

- [73] Fukami-Kobayashi K, Tateno Y, Nishikawa K (2003) Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol Biol Evol* **20**:267-277.
doi:10.1093/molbev/msg038
- [74] Lehming N (1990) Regeln für Protein/DNA-Erkennung (PhD Thesis). Universität zu Köln.
- [75] Hollis M, Valenzuela D, Pioli D, Wharton R, Ptashne M (1988) A repressor heterodimer binds to a chimeric operator. *Proc Natl Acad Sci USA* **85**:5834-5838.
doi:10.1073/pnas.85.16.5834
- [76] MacArthur S, Brookfield JFY (2004) Expected rates and modes of evolution of enhancer sequences. *Mol Biol Evol* **21**:1064-1073.
doi:10.1093/molbev/msh105
- [77] Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**:2052-2058.
doi:10.1101/gr.1252603
- [78] Lynch M (2005) Simple evolutionary pathways to complex proteins. *Protein Sci* **14**:2217-2225.
doi:10.1110/ps.041171805
- [79] Francino MP (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**:573-577.
doi:10.1038/ng1579
- [80] Jürgens C, Strom A, Wegener D, Hettwer S, Wilmanns M, et al. (2000) Directed evolution of a ($\beta\alpha$)₈-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc Natl Acad Sci USA* **97**:9925-9930.
doi:10.1073/pnas.160255397
- [81] O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem Biol* **6**:R91-R105.
doi:10.1016/S1074-5521(99)80033-7
- [82] Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biol* **3**:RESEARCH0008.
doi:10.1186/gb-2002-3-2-research0008
- [83] Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**:186-189.
doi:10.1038/nature01149
- [84] Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* **4**:42.
doi:10.1186/1471-2148-4-42
- [85] Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3**:e130.
doi:10.1371/journal.pbio.0030130
- [86] Inagaki Y, Doolittle WF, Baldauf SL, Roger AJ (2002) Lateral transfer of an EF-1 α gene: origin and evolution of the large subunit of ATP sulfurylase in eubacteria. *Curr Biol* **12**:772-776.
doi:10.1016/S0960-9822(02)00816-3
- [87] Stoebel DM (2005) Lack of evidence for horizontal transfer of the lac operon into Escherichia coli. *Mol Biol Evol* **22**:683-690.
doi:10.1093/molbev/msi056
- [88] Bhan A, Galas DJ, Dewey TG (2002) A duplication growth model of gene expression networks. *Bioinformatics* **18**:1486-1493.
doi:10.1093/bioinformatics/18.11.1486
- [89] Wagner A (2003) How the global structure of protein interaction networks evolves. *Proc Biol Sci* **270**:457-466.
doi:10.1098/rspb.2002.2269

Bibliography

- [90] Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, et al. (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci USA* **101**:8414-8419. doi:10.1073/pnas.0402940101
- [91] Weber W, Fussenegger M (2002) Artificial mammalian gene regulation networks—novel approaches for gene therapy and bioengineering. *J Biotechnol* **98**:161-187. doi:10.1016/S0168-1656(02)00130-X
- [92] Farmer WR, Liao JC (2000) Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat Biotechnol* **18**:533-537. doi:10.1038/75398
- [93] Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. *Proc Natl Acad Sci USA* **99**:16587-16591. doi:10.1073/pnas.252535999
- [94] Miller JH (1972) *Experiments in Molecular Genetics*. New York: Cold Spring Harbor Laboratory Press.
- [95] Sadler JR, Novick A (1965) The properties of repressor and the kinetics of its action. *J Mol Biol* **12**:305-327. doi:10.1016/S0022-2836(65)80255-8
- [96] Sadler JR, Sasmor H, Betz JL (1983) A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc Natl Acad Sci USA* **80**:6785-6789. doi:10.1073/pnas.80.22.6785
- [97] Betz JL, Sasmor HM, Buck F, Insley MY, Caruthers MH (1986) Base substitution mutants of the lac operator: in vivo and in vitro affinities for lac repressor. *Gene* **50**:123-132. doi:10.1016/0378-1119(86)90317-3
- [98] Dubertret B, Liu S, Ouyang Q, Libchaber A (2001) Dynamics of DNA-protein interaction deduced from in vitro DNA evolution. *Phys Rev Lett* **86**:6022-6025. doi:10.1103/PhysRevLett.86.6022
- [99] Korona R, Nakatsu CH, Forney LJ, Lenski RE (1994) Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proc Natl Acad Sci USA* **91**:9037-9041. doi:10.1073/pnas.91.19.9037
- [100] Fisher RA (1950) The "Sewell Wright Effect". *Heredity* **4**:117-119. doi:10.1038/hdy.1950.8
- [101] Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**:11-45. doi:10.1016/S0022-5193(87)80029-2
- [102] 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. doi:10.1016/S0022-5193(05)80094-3
- [103] 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. doi:10.1016/S0022-5193(89)80019-0
- [104] Kim Y (2007) Rate of adaptive peak shifts with partial genetic robustness. *Evolution* **61**:1847-1856. doi:10.1111/j.1558-5646.2007.00166.x
- [105] Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**:383-386. doi:10.1038/nature05451
- [106] Buckling A, Wills MA, Colegrave N (2003) Adaptation limits diversification of experimental bacterial populations. *Science* **302**:2107-2109. doi:10.1126/science.1088848
- [107] Wright S (1931) Evolution in Mendelian Populations. *Genetics* **16**:97-159. <http://www.genetics.org/cgi/reprint/16/2/97>

- [108] Whitlock MC, Phillips PC, Moore FBG, Tonsor SJ (1995) Multiple Fitness Peaks and Epistasis. *Annu Rev Ecol Syst* **26**:601-629.
doi:10.1146/annurev.es.26.110195.003125
- [109] Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**:318-356.
doi:10.1016/S0022-2836(61)80072-7
- [110] Adler K, Beyreuther K, Fanning E, Geisler N, Gronenborn B, et al. (1972) How lac repressor binds to DNA. *Nature* **237**:322-327.
doi:10.1038/237322a0
- [111] Barker A, Fickert R, Oehler S, Müller-Hill B (1998) Operator search by mutant Lac repressors. *J Mol Biol* **278**:549-558.
doi:10.1006/jmbi.1998.1729
- [112] Bell CE, Lewis M (2001) The Lac repressor: a second generation of structural and functional studies. *Curr Opin Struct Biol* **11**:19-25.
doi:10.1016/S0959-440X(00)00180-9
- [113] 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.
doi:10.1016/0022-2836(78)90025-6
- [114] 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.
doi:10.1016/0022-2836(77)90056-0
- [115] Eismann ER, Müller-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.
doi:10.1016/S0022-2836(05)80262-1
- [116] Fickert R, Müller-Hill B (1992) How Lac repressor finds lac operator in vitro. *J Mol Biol* **226**:59-68.
doi:10.1016/0022-2836(92)90124-3
- [117] Lehming N, Sartorius J, Niemöller 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.
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC553756/>
- [118] 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.
doi:10.1016/S0022-2836(77)80034-X
- [119] 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.
doi:10.1006/jmbi.1996.0479
- [120] Poelwijk FJ, Tanase-Nicola S, Kiviet DJ, Tans SJ (2010) Reciprocal sign epistasis is a necessary condition for multipeaked fitness landscapes. Submitted to *J. Theor. Biol.*
- [121] 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.
doi:10.1126/science.1135308
- [122] Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**:515-519.
doi:10.1038/nature08249
- [123] Cowperthwaite MC, Meyers LA (2007) How Mutational Networks Shape Evolution: Lessons from RNA Models. *Annu Rev Ecol Evol Syst* **38**:203-230.
doi:10.1146/annurev.ecolsys.38.091206.095507
- [124] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* **106**:67-72.
doi:10.1073/pnas.0805923106

Bibliography

- [125] Segré D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* **37**:77-83.
doi:10.1038/ng1489
- [126] Mildvan AS (2004) Inverse thinking about double mutants of enzymes. *Biochemistry* **43**:14517-14520.
doi:10.1021/bi048052e
- [127] DeLano WL The PyMOL Molecular Graphics System.
- [128] Lewis M (2005) The lac repressor. *C R Biol* **328**:521-548.
doi:10.1016/j.crvi.2005.04.004
- [129] 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.
doi:10.1016/0022-2836(79)90074-3
- [130] Miller JH (1984) Genetic studies of the lac repressor. XII. Amino acid replacements in the DNA binding domain of the Escherichia coli lac repressor. *J Mol Biol* **180**:205-212.
doi:10.1016/0022-2836(84)90438-8
- [131] 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.
doi:10.1063/1.3453602
- [132] Forrest S, Mitchell M (1993) Foundations of Genetic Algorithms 2, San Mateo, CA: Morgan Kaufmann. pp. 109-126.
- [133] McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet* **15**:65-69.
doi:10.1016/S0168-9525(98)01659-X
- [134] McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA* **94**:814-819.
doi:10.1073/pnas.94.3.814
- [135] Levine E, Hwa T (2007) Stochastic fluctuations in metabolic pathways. *Proc Natl Acad Sci USA* **104**:9224-9229.
doi:10.1073/pnas.0610987104
- [136] Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**:840-846.
doi:10.1038/nature04785
- [137] Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, et al. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**:636-643.
doi:10.1038/ng1807
- [138] Stiel GM, Garcia-Ojalvo J, Liberman LM, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* **440**:545-550.
doi:10.1038/nature04588
- [139] Korobkova E, Emonet T, Vilar JMG, Shimizu TS, Cluzel P (2004) From molecular noise to behavioural variability in a single bacterium. *Nature* **428**:574-578.
doi:10.1038/nature02404
- [140] Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**:451-464.
doi:10.1038/nrg1615
- [141] Neidhardt FC, Ingraham JL, Schaechter M (1990) Physiology of the Bacterial Cell: A Molecular Approach. Sunderland, Massachusetts: Sinauer Associates Inc.
- [142] Dean AM (1994) Fitness, flux and phantoms in temporally variable environments. *Genetics* **136**:1481-1495.
<http://www.genetics.org/cgi/content/abstract/136/4/1481>
- [143] Fell D (1997) Understanding the Control of Metabolism. London: Portland Press.

- [144] Koppes LH, Woldringh CL, Nanninga N (1978) Size variations and correlation of different cell cycle events in slow-growing *Escherichia coli*. *J Bacteriol* **134**:423-433.
<http://jb.asm.org/cgi/content/abstract/134/2/423>
- [145] Cooper S (1991) *Bacterial Growth and Division*. San Diego: Academic Press.
- [146] Adiciptaningrum AM (2009) Phase Variation of Type 1 Fimbriae: a Single Cell Investigation (PhD Thesis). Universiteit van Amsterdam.
- [147] Rahn O (1932) A chemical explanation of the variability of the growth rate. *J Gen Physiol* **15**:257-277.
doi:10.1085/jgp.15.3.257
- [148] Schaechter M, Williamson JP, Hood JR, Koch AL (1962) Growth, cell and nuclear divisions in some bacteria. *J Gen Microbiol* **29**:421-434.
doi:10.1099/00221287-29-3-421
- [149] Stewart EJ, Madden R, Paul G, Taddei F (2005) Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol* **3**:e45.
doi:10.1371/journal.pbio.0030045
- [150] Strovas TJ, Sauter LM, Guo X, Lidstrom ME (2007) Cell-to-cell heterogeneity in growth rate and gene expression in *Methylobacterium extorquens* AM1. *J Bacteriol* **189**:7127-7133.
doi:10.1128/JB.00746-07
- [151] Reshes G, Vanounou S, Fishov I, Feingold M (2008) Timing the start of division in *E. coli*: a single-cell study. *Phys Biol* **5**:046001.
doi:10.1088/1478-3975/5/4/046001
- [152] Tsuru S, Ichinose J, Kashiwagi A, Ying BW, Kaneko K, et al. (2009) Noisy cell growth rate leads to fluctuating protein concentration in bacteria. *Phys Biol* **6**:036015.
doi:10.1088/1478-3975/6/3/036015
- [153] Jobe A, Bourgeois S (1972) lac Repressor-operator interaction. VI. The natural inducer of the lac operon. *J Mol Biol* **69**:397-408.
doi:10.1016/0022-2836(72)90253-7
- [154] Müller-Hill B, Rickenberg HV, Wallenfels K (1964) Specificity of the induction of the enzymes of the lac operon in *Escherichia coli*. *J Mol Biol* **10**:303-318.
doi:10.1016/S0022-2836(64)80049-8
- [155] Huber RE, Wallenfels K, Kurz G (1975) The action of beta-galactosidase (*Escherichia coli*) on allolactose. *Can J Biochem* **53**:1035-1038.
doi:10.1139/o75-142
- [156] Zabin I, Kepes A, Monod J (1962) Thiogalactoside transacetylase. *J Biol Chem* **237**:253-257.
<http://www.jbc.org/content/237/1/253>
- [157] Musso RE, Zabin I (1973) Substrate specificity and kinetic studies on thiogalactoside transacetylase. *Biochemistry* **12**:553-557.
doi:10.1021/bi00727a031
- [158] Lewendon A, Ellis J, Shaw WV (1995) Structural and mechanistic studies of galactoside acetyltransferase, the *Escherichia coli* LacA gene product. *J Biol Chem* **270**:26326-26331.
<http://www.jbc.org/content/270/44/26326>
- [159] Andrews KJ, Lin EC (1976) Thiogalactoside transacetylase of the lactose operon as an enzyme for detoxification. *J Bacteriol* **128**:510-513.
<http://jb.asm.org/cgi/content/abstract/128/1/510>
- [160] Danchin A (2009) Cells need safety valves. *Bioessays* **31**:769-773.
doi:10.1002/bies.200900024
- [161] Wilson TH, Kashket ER (1969) Isolation and properties of thiogalactoside transacetylase-negative mutants of *Escherichia coli*. *Biochim Biophys Acta* **173**:501-508.
doi:10.1016/0005-2736(69)90014-5
- [162] Dean AM (1995) A molecular investigation of genotype by environment interactions. *Genetics* **139**:19-33.
<http://www.genetics.org/cgi/content/abstract/139/1/19>

Bibliography

- [163] Commichau FM, Stülke J (2008) Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression. *Mol Microbiol* **67**:692-702.
doi:10.1111/j.1365-2958.2007.06071.x
- [164] Wray LV, Zalieckas JM, Fisher SH (2005) *Bacillus subtilis* glutamine synthetase controls gene expression through a protein-protein interaction with transcription factor TnrA. *Cell* **107**:427-435.
doi:10.1016/S0092-8674(01)00572-4
- [165] Brown ED, Wood JM (1992) Redesigned purification yields a fully functional PutA protein dimer from *Escherichia coli*. *J Biol Chem* **267**:13086-13092.
<http://www.jbc.org/content/267/18/13086>
- [166] Ostrovsky de Spicer P, Maloy S (1993) PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator. *Proc Natl Acad Sci USA* **90**:4295-4298.
doi:10.1073/pnas.90.9.4295
- [167] Hartl DL, Dykhuizen DE, Dean AM (1985) Limits of adaptation: the evolution of selective neutrality. *Genetics* **111**:655-674.
<http://www.genetics.org/cgi/content/abstract/111/3/655>
- [168] Dean AM, Dykhuizen DE, Hartl DL (1986) Fitness as a function of beta-galactosidase activity in *Escherichia coli*. *Genet Res* **48**:1-8.
doi:10.1017/S0016672300024587
- [169] Dykhuizen DE, Dean AM, Hartl DL (1987) Metabolic flux and fitness. *Genetics* **115**:25-31.
<http://www.genetics.org/cgi/content/abstract/115/1/25>
- [170] Goh S, Boberek JM, Nakashima N, Stach J, Good L (2009) Concurrent growth rate and transcript analyses reveal essential gene stringency in *Escherichia coli*. *PLoS ONE* **4**:e6061.
doi:10.1371/journal.pone.0006061
- [171] Good L, Sandberg R, Larsson O, Nielsen PE, Wahlestedt C (2000) Antisense PNA effects in *Escherichia coli* are limited by the outer-membrane LPS layer. *Microbiology (Reading, Engl)* **146** (Pt 10):2665-2670.
<http://mic.sgmjournals.org/cgi/content/abstract/146/10/2665>
- [172] Rotman B (1961) Measurement of activity of single molecules of beta-D-galactosidase. *Proc Natl Acad Sci USA* **47**:1981-1991.
doi:10.1073/pnas.47.12.1981
- [173] Becskei A, Kaufmann BB, van Oudenaarden A (2005) Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat Genet* **37**:937-944.
doi:10.1038/ng1616
- [174] Longo D, Hasty J (2006) Dynamics of single-cell gene expression. *Mol Syst Biol* **2**:64.
doi:10.1038/msb4100110
- [175] Bennett MR, Hasty J (2009) Microfluidic devices for measuring gene network dynamics in single cells. *Nat Rev Genet* **10**:628-638.
doi:10.1038/nrg2625
- [176] Brown JL, Brown DM, Zabin I (1967) Thiogalactoside transacetylase. Physical and chemical studies of subunit structure. *J Biol Chem* **242**:4254-4258.
<http://www.jbc.org/content/242/18/4254>
- [177] Jones TH, Kennedy EP (1969) Characterization of the membrane protein component of the lactose transport system of *Escherichia coli*. *J Biol Chem* **244**:5981-5987.
<http://www.jbc.org/content/244/21/5981>
- [178] Cormack BP, Valdivia RH, Falkow S (1996) FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**:33-38.
doi:10.1016/0378-1119(95)00685-0
- [179] Shaner NC, Steinbach PA, Tsien RY (2005) A guide to choosing fluorescent proteins. *Nat Methods* **2**:905-909.
doi:10.1038/nmeth819

- [180] Zaslaver A, Bren A, Ronen M, Itzkovitz S, Kikoin I, et al. (2006) A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nat Methods* **3**:623-628.
doi:10.1038/nmeth895
- [181] Dunlop MJ, Cox RS, Levine JH, Murray RM, Elowitz MB (2008) Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet* **40**:1493-1498.
doi:10.1038/ng.281
- [182] Bigger JW, Nelson JH (1941) The growth of coliform bacilli in distilled water. *J Pathol Bacteriol* **53**:189-206.
doi:10.1002/path.1700530204
- [183] Postgate JR, Hunter JR (1962) The survival of starved bacteria. *J Gen Microbiol* **29**:233-263.
doi:10.1099/00221287-29-2-233
- [184] Shehata TE, Marr AG (1971) Effect of nutrient concentration on the growth of Escherichia coli. *J Bacteriol* **107**:210-216.
<http://jb.asm.org/cgi/content/abstract/107/1/210>
- [185] Sakai T, Nakamura N, Umitsuki G, Nagai K, Wachi M (2007) Increased production of pyruvic acid by Escherichia coli RNase G mutants in combination with cra mutations. *Appl Microbiol Biotechnol* **76**:183-192.
doi:10.1007/s00253-007-1006-9
- [186] Powell EO (1956) Growth rate and generation time of bacteria, with special reference to continuous culture. *J Gen Microbiol* **15**:492-511.
doi:10.1099/00221287-15-3-492
- [187] Trueba FJ (1982) On the precision and accuracy achieved by Escherichia coli cells at fission about their middle. *Arch Microbiol* **131**:55-59.
doi:10.1007/BF00451499
- [188] Reshes G, Vanounou S, Fishov I, Feingold M (2008) Cell shape dynamics in Escherichia coli. *Biophys J* **94**:251-264.
doi:10.1529/biophysj.107.104398
- [189] Marr AG, Harvey RJ, Trentini WC (1966) Growth and division of Escherichia coli. *J Bacteriol* **91**:2388-2389.
<http://jb.asm.org/cgi/content/abstract/91/6/2388>
- [190] Moisy F (2009). "Ezyfit toolbox for Matlab Version 2.30. <http://www.fast.u-psud.fr/ezyfit/>.
- [191] Monod J (1949) The Growth of Bacterial Cultures. *Annu Rev Microbiol* **3**:371-394.
doi:10.1146/annurev.mi.03.100149.002103
- [192] Kovárová-Kovar K, Egli T (1998) Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics. *Microbiol Mol Biol Rev* **62**:646-666.
<http://mmb.asm.org/cgi/content/abstract/62/3/646>
- [193] Marr AG, Nilson EH, Clark DJ (1963) The Maintenance Requirement of Escherichia Coli. *Ann N Y Acad Sci* **102**:536-548.
doi:10.1111/j.1749-6632.1963.tb13659.x
- [194] Pirt SJ (1965) The maintenance energy of bacteria in growing cultures. *Proc R Soc Lond, B, Biol Sci* **163**:224-231.
doi:10.1098/rspb.1965.0069
- [195] Bendat JS, Piersol AG (2000) Random Data: Analysis and Measurement Procedures. New York, NY, USA: John Wiley & Sons, Inc.
- [196] Austin DW, Allen MS, McCollum JM, Dar RD, Wilgus JR, et al. (2006) Gene network shaping of inherent noise spectra. *Nature* **439**:608-611.
doi:10.1038/nature04194
- [197] Boezi JA, Cowie DB (1961) Kinetic studies of beta-galactosidase induction. *Biophys J* **1**:639-647.
doi:10.1016/S0006-3495(61)86913-0

Bibliography

- [198] Verkhusha VV, Akovbian NA, Efremenko EN, Varfolomeyev SD, Vrzheschch PV (2001) Kinetic analysis of maturation and denaturation of DsRed, a coral-derived red fluorescent protein. *Biochemistry Mosc* **66**:1342-1351.
doi:10.1023/A:1013325627378
- [199] Dong GQ, McMillen DR (2008) Effects of protein maturation on the noise in gene expression. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**:021908.
doi:10.1103/PhysRevE.77.021908
- [200] Klumpp S, Zhang Z, Hwa T (2009) Growth rate-dependent global effects on gene expression in bacteria. *Cell* **139**:1366-1375.
doi:10.1016/j.cell.2009.12.001
- [201] Levy S, Barkai N (2009) Coordination of gene expression with growth rate: a feedback or a feed-forward strategy? *FEBS Lett* **583**:3974-3978.
doi:10.1016/j.febslet.2009.10.071
- [202] Maaløe O (1979) *Biological Regulation and Development*, New York: Plenum Press. pp. 487-542.
- [203] Berg OG (1978) A model for the statistical fluctuations of protein numbers in a microbial population. *J Theor Biol* **71**:587-603.
doi:10.1016/0022-5193(78)90326-0
- [204] Colman-Lerner A, Gordon A, Serra E, Chin T, Resnekov O, et al. (2005) Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**:699-706.
doi:10.1038/nature03998
- [205] Volfson D, Marciniak J, Blake WJ, Ostroff N, Tsimring LS, et al. (2006) Origins of extrinsic variability in eukaryotic gene expression. *Nature* **439**:861-864.
doi:10.1038/nature04281
- [206] Maheshri N, O'Shea EK (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct* **36**:413-434.
doi:10.1146/annurev.biophys.36.040306.132705
- [207] Maloney PC, Rotman B (1973) Distribution of suboptimally induced beta-D-galactosidase in *Escherichia coli*. The enzyme content of individual cells. *J Mol Biol* **73**:77-91.
doi:10.1016/0022-2836(73)90160-5
- [208] van Hoek MJA, Hogeweg P (2006) In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. *Biophys J* **91**:2833-2843.
doi:10.1529/biophysj.105.077420
- [209] Ninfa AJ, Mayo AE (2004) Hysteresis vs. graded responses: the connections make all the difference. *Sci STKE* **2004**:pe20.
doi:10.1126/stke.2322004pe20
- [210] Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol* **4**:133-140.
doi:10.1046/j.1462-2920.2002.00282.x
- [211] Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* **97**:5528-5533.
doi:10.1073/pnas.97.10.5528
- [212] Fendt SM, Buescher JM, Rudroff F, Picotti P, Zamboni N, et al. (2010) Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol Syst Biol* **6**:356.
doi:10.1038/msb.2010.11
- [213] Cox CD, McCollum JM, Austin DW, Allen MS, Dar RD, et al. (2006) Frequency domain analysis of noise in simple gene circuits. *Chaos* **16**:026102.
doi:10.1063/1.2204354
- [214] Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, et al. (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell* **24**:853-865.
doi:10.1016/j.molcel.2006.11.003

- [215] Elf J, Ehrenberg M (2005) Near-critical behavior of aminoacyl-tRNA pools in *E. coli* at rate-limiting supply of amino acids. *Biophys J* **88**:132-146.
doi:10.1529/biophysj.104.051383
- [216] Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* **97**:6640-6645.
doi:10.1073/pnas.120163297
- [217] Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453-1462.
doi:10.1126/science.277.5331.1453
- [218] Jensen KF (1993) The *Escherichia coli* K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. *J Bacteriol* **175**:3401-3407.
<http://jbs.asm.org/cgi/content/abstract/175/11/3401>
- [219] Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, et al. (2003) Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J Bacteriol* **185**:5611-5626.
doi:10.1128/JB.185.18.5611-5626.2003
- [220] Lawther RP, Calhoun DH, Adams CW, Hauser CA, Gray J, et al. (1981) Molecular basis of valine resistance in *Escherichia coli* K-12. *Proc Natl Acad Sci USA* **78**:922-925.
doi:10.1073/pnas.78.2.922
- [221] Liu D, Reeves PR (1994) *Escherichia coli* K12 regains its O antigen. *Microbiology (Reading, Engl)* **140 (Pt 1)**:49-57.
doi:10.1099/13500872-140-1-49
- [222] De Paepe M, Taddei F (2006) Viruses' life history: towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol* **4**:e193.
doi:10.1371/journal.pbio.0040193
- [223] Wallace W, Schaefer LH, Swedlow JR (2001) A workingperson's guide to deconvolution in light microscopy. *BioTechniques* **31**:1076-8, 1080, 1082 passim.
- [224] Itan E, Carmon G, Rabinovitch A, Fishov I, Feingold M (2008) Shape of nonseptated *Escherichia coli* is asymmetric. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**:061902.
doi:10.1103/PhysRevE.77.061902
- [225] Takeuchi S, DiLuzio WR, Weibel DB, Whitesides GM (2005) Controlling the shape of filamentous cells of *Escherichia coli*. *Nano Lett* **5**:1819-1823.
doi:10.1021/nl0507360
- [226] Männik J, Driessen R, Galajda P, Keymer JE, Dekker C (2009) Bacterial growth and motility in sub-micron constrictions. *Proc Natl Acad Sci USA* **106**:14861-14866.
doi:10.1073/pnas.0907542106
- [227] Grover NB, Woldringh CL, Zaritsky A, Rosenberger RF (1977) Elongation of rod-shaped bacteria. *J Theor Biol* **67**:181-193.
doi:10.1016/0022-5193(77)90192-8
- [228] Kubitschek HE, Woldringh CL (1983) Cell elongation and division probability during the *Escherichia coli* growth cycle. *J Bacteriol* **153**:1379-1387.
<http://jbs.asm.org/cgi/content/abstract/153/3/1379>
- [229] Koch AL (1993) Biomass growth rate during the prokaryote cell cycle. *Crit Rev Microbiol* **19**:17-42.
doi:10.3109/10408419309113521

Summary

The *lac* Operon: Fluctuations, Growth and Evolution

This thesis is concerned with two distinct fundamental research questions that are both investigated using the *E. coli lac* system. In the first chapters we investigate what the shape of biological fitness landscapes look like. Chapter 2 reviews recent progress in measurement of empirical fitness landscapes, and introduces the open questions in evolution that they may answer, such as why particular evolutionary paths are taken. In this chapter, we also introduce the concept of epistasis as a useful description of the local shape of fitness landscapes. In chapter 3 we describe existing *in vivo* measurements on *lac* repressor and operator mutants and show how these can be used to construct a fitness landscape of *lac* regulation. Using computer simulations we simulate mutational pathways and reveal that new regulatory interactions can easily evolve. Chapter 4 deals with the local structure of the *lac* landscape. We determine that the landscape is multi-peaked and, consistent with earlier predictions, show the presence of reciprocal sign epistasis. We conclude our analysis of the *lac* landscape in chapter 5 with a more global analysis of its structure, focusing on which landscape features are important for evolution. This study reveals that the essential features of the *lac* landscape can be sufficiently captured by modeling the presence or absence of additivity between functional residues.

In chapter 6 we turn to another fundamental research question: how do random molecular fluctuations in the number proteins in a single cell propagate to its growth? Again, we use the *E. coli lac* system to investigate this question. But whereas the first part of this thesis consists of theoretical simulations of *lac* regulation, here we perform laboratory experiments on *E. coli* cells that require use of their *lac* enzymes for growth. By means of automated and highly sensitive fluorescence microscopy, we measure both fluctuations in *lac* level and in growth rate in individual growing cells. These experiments show that fluctuations in the growth rate of single cells can be linked to protein fluctuations, but also reveal a intricate dynamic interdependency between these two properties.

Samenvatting

Het *lac* Operon: Fluctuaties, Groei en Evolutie

De cellen waar wij en alle andere levende wezens uit zijn opgebouwd, bestaan uit een enorme hoeveelheid moleculen die schijnbaar willekeurig tegen elkaar aan het botsen zijn. Het is verbazingwekkend dat de interactie van deze niet-levende deeltjes leidt tot al het complexe leven op aarde. Met fundamenteel onderzoek proberen wij te begrijpen hoe dit allemaal kan. Soms leidt zulk fundamenteel onderzoek tot grote technologische ontwikkelingen. Meestal draagt het voornamelijk bij aan een beter begrip van de natuur, en steevast aan een set nieuwe uitdagende open vragen.

In dit proefschrift is een enkel moleculair model systeem onderzocht: het *lac* operon. Dit specifieke systeem komt voor in de *E. coli* bacterie, een klein eencellig organisme. Varianten van het systeem zijn echter terug te vinden in vele organismen, waaronder de mens. Het *lac* systeem is verantwoordelijk voor slechts één taak van de cel: het consumeren van lactose. Het bestaat daarvoor uit gespecialiseerde eiwitten die lactose detecteren en andere die de lactose afbreken zodat de cel het kan gebruiken om te groeien. Door decennia van onderzoek is er al ontzettend veel over het *lac* operon bekend. Het is daardoor uitermate geschikt om fundamentele vragen te onderzoeken. Dit proefschrift maakt hier optimaal gebruik van door meteen twee erg verschillende onderzoeksvragen te belichten.

In het eerste gedeelte van dit proefschrift wordt het *lac* operon gebruikt om evolutionaire vragen te onderzoeken: Hoe zijn complexe moleculaire systemen geëvolueerd? Kunnen we begrijpen waarom ze überhaupt kunnen evolueren? En kunnen we voorspellen hoe ze in de toekomst zullen evolueren? Omdat evolutie een ontzettend veelomvattend proces is, zijn zulke vragen niet simpel te beantwoorden. Een methode om inzicht in de evolutionaire mogelijkheden van een systeem te krijgen is door het effect van veranderingen (mutaties) te meten. Met behulp van zulke metingen kan een zogenaamd fitness landschap van het systeem gemaakt worden. Dit landschap, met pieken en dalen, beschrijft wat het effect van verschillende opvolgende mutaties zijn. Door zo'n gemeten landschap te analyseren kunnen we de evolutionaire mogelijkheden van het systeem kwantificeren.

In hoofdstukken 2 t/m 5 onderzoeken wij hoe fitness landschappen van natuurlijke systemen eruit zien. We vergelijken recent gemeten landschappen (hoofdstuk 2) en simuleren het proces van evolutie op het fitness landschap van het *lac* operon (hoofdstuk 3). Een analyse van de structuur van het *lac* landschap toont aan dat deze meerdere pieken bevat (hoofdstuk 4). Deze kunnen we relateren aan lokale vormen in het landschap die met het concept van epistase beschreven kunnen worden. We sluiten onze analyse af met een globale analyse van het *lac* landschap, waarbij we op zoek gaan naar structuren in het landschap die van belang zijn voor succesvolle evolutie (hoofdstuk 5). Tot op zekere hoogte blijken deze structuren

met behulp van simpel model te kunnen worden beschreven.

In het laatste gedeelte van dit proefschrift (hoofdstuk 6) wordt het *lac* operon gebruikt voor een heel ander soort onderzoeksvraag, namelijk gericht op de moleculaire werking van het systeem. Hier kijken we hoe de moleculaire werking van het *lac* systeem beïnvloed wordt door fluctuaties in de onderdelen (moleculen) waaruit het systeem bestaat. Deze fluctuaties, inherent aan moleculaire processen, treden op in alle cellen. Het is echter nog onduidelijkheid of de groei van cellen door deze fluctuaties wordt beïnvloed. Daarom hebben we fluctuaties in de groeisnelheid van enkele cellen tegelijkertijd gemeten met fluctuaties in de hoeveelheid van de *lac* eiwitten. Het blijkt dat de groeisnelheid van *E. coli* cellen inderdaad beïnvloed wordt door fluctuaties in het *lac* operon.

Dankwoord

DAAN KIVIET
Amsterdam, oktober 2010



List of Publications

Daniel J. Kiviet[Ⓜ], Frank J. Poelwijk[Ⓜ], Daniel M. Weinreich, and Sander J. Tans
Empirical fitness landscapes reveal accessible evolutionary paths
Nature **445**:383-386 (2007) - Chapter 2

Daniel J. Kiviet[Ⓜ], Frank J. Poelwijk[Ⓜ], and Sander J. Tans
Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data
PLoS Comput Biol **2**:e58 (2006) - Chapter 3

Alexandre Dawid, Daniel J. Kiviet, Manjunatha Kogenu, Marjon de Vos, and Sander J. Tans
Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape
Chaos **2**:026105 (2010) - Chapter 4

Daniel J. Kiviet, Pieter Rein ten Wolde, and Sander J. Tans
Simple rules underlie an empirically determined genotype-phenotype landscape
in preparation - Chapter 5

Daniel J. Kiviet, Philippe Nghe, and Sander J. Tans
Noise propagation in metabolic networks
in preparation - Chapter 6

Frank J. Poelwijk, Philip Heijning, Marjon G.J. de Vos, Daniel J. Kiviet, and Sander J. Tans
Optimality and the evolution of transcriptionally regulated gene expression
submitted to BMC Syst Biol

Frank J. Poelwijk, Sorin Tănase-Nicola, Daniel J. Kiviet, and Sander J. Tans
Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes
under review at J Theor Biol

[Ⓜ] Joined first authors

Curriculum Vitae

Daniel J. Kiviet

- 1979.11.15 Born in Amsterdam, the Netherlands.
- 1991-1997 Secondary school, Stedelijke Scholengemeenschap Middelburg.
- 1997-2002 M.Sc. Chemistry (cum laude),
Universiteit van Amsterdam.
Graduate research in the Molecular Microbial Physiology group
of prof. dr. K. J. Hellingwerf.
- 2002-2003 Research analyst,
Universiteit van Amsterdam,
Molecular Biology & Microbial Food Safety group of prof. dr. S. Brul.
- 2003-2003 Graduate Diploma in Information Technology (cum laude),
Australian National University.
Research under supervision of dr. H. Gardner
- 2004-2010 Ph.D. research,
FOM Institute for Atomic and Molecular Physics,
Biophysics group of prof. dr. S. J. Tans.