Evolutionary Dynamics of Metabolic Adaptation

OMSLAG Rob Kreuger en Milan van Hoek

DRUK PrintPartners Ipskamp

ISBN 978-90-9022791-7

Evolutionary Dynamics of Metabolic Adaptation

De Evolutionaire Dynamica van Metabolische Aanpassing

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 5 maart 2008 des middags te 14.30 uur

door

Milan Johannes Adrianus van Hoek

geboren op 28 juni 1978
te Tilburg.

Promotor:
Prof. dr. P. Hogeweg

The studies described in this thesis were performed at the department of Theoretical Biology and Bioinformatics at Utrecht University.

Contents

1

General Introduction

Every organism has to eat to stay alive and to reproduce. Therefore metabolism is one of the defining properties of life. Metabolism generates the building blocks that organisms need to grow and the energy they need to conduct processes that are needed to stay alive. Energy is also needed to maintain homeostasis, which means that organisms need to keep a certain balanced internal environment in order to stay alive. Or in the words of Erwin Schrödinger: "the essential thing in metabolism is that the organism succeeds in freeing itself from all the the entropy it cannot help producing while alive." (Schrödinger, 1944).

Basic metabolism has been conserved remarkably well during evolution. The TCA-cycle for example is found across the whole Tree of Life, from archaea to multicellular eukaryotes. Although there are many differences in metabolism between species, the core metabolic pathways are found in almost every species. Therefore we can study metabolism in just a few species very precisely and still obtain relatively general results. In this thesis we study metabolism in two model organisms, the bacterium *Escherichia coli* and the budding yeast *Saccharomyces cerevisiae*. Both these organisms are unicellular organisms. Therefore, unlike multicellular organisms, these organisms need to perform all metabolic functions in one cell. This makes them easy to study and, for the purpose of this thesis, relatively easy to model.

In this thesis we study how organisms adapt their metabolism to a changing environment. This happens in essentially two ways, via metabolic regulation and evolutionary adaptation. Maybe the best known example of metabolic regulation is the regulation of the glucose level in the blood via insulin (and glucagon). When the glucose level in the blood rises, β cells in the islets of Langerhans start producing insulin, which is secreted into the blood (see for example Meugnier *et al.* (2007)). Metabolic regulation happens on two time scales. First, when the environment of a cell changes, the intracellular environment changes very fast, because fluxes through metabolic pathways will change. This occurs in the order of seconds. If the intracellular environment has changed, this may lead to a change in gene expression in the cell. In the case of glucose regulation in the blood, insulin is produced. This occurs in the order of minutes to hours.

Obviously, metabolic regulation is crucial for the survival of an organism. If the regulation of the glucose level fails, diabetes results. Metabolic regulation is therefore needed to decide under which circumstances which metabolic pathways should be active. Not every metabolic pathway will be active at a given

time for several reasons. Firstly, it is costly to produce the proteins required to activate a certain metabolic pathway. Secondly, cells have a limited volume and if all proteins needed to activate all metabolic pathways were present together, this would never fit in a single cell. Thirdly, some metabolic pathways can hinder each other in their function and therefore should not be active together.

Organisms can also adapt their metabolism via evolutionary adaptation. Obviously, evolutionary adaptation takes many generations and therefore defines a third time scale. In this thesis we study the interplay between these three time scales (the time scales of metabolic/genetic regulation and evolutionary adaptation) by studying the evolutionary dynamics of metabolic adaptation.

We do this from two extreme perspectives. In the first part of this thesis we study the evolution of a very small metabolic pathway, the *lac* operon of *E. coli*. The *lac* operon, which codes for only three proteins, regulates lactose uptake and metabolism and is a paradigm system for genetic regulation.

In the second part of this thesis we study the evolution of metabolism from an entirely different perspective. Metabolic reactions do not occur in isolation, but are linked to each other. Most metabolites that are produced are again the precursor of other metabolites. In this way all metabolites are linked into a network that consists of hundreds or thousands metabolites and interactions. In *S. cerevisiae* this network has been extensively studied. We study the evolution of the metabolic network after a whole genome duplication that is known to have occurred in *S. cerevisiae*. Fig. 1.1 gives an idea of the difference in scale between the *lac* operon of *E. coli* and the metabolic network of *S. cerevisiae*. In our model of the *lac* operon only 9 metabolites are incorporated (see Fig. 1.1 *C*), whereas in the genome-scale metabolic model 1061 metabolites are incorporated (of which only a minority is shown in Fig. 1.1 *A*).

Here we will first give an introduction to the biology of *E. coli* and the *lac* operon. In the second part of the Introduction we focus on *S. cerevisiae* and metabolic networks.

1.1 The *lac* Operon of *Escherichia coli*

E. coli is a Gram-negative, rod-shaped bacterium, belonging to the class of gammaproteobacteria, that lives in the mammalian gut. Although there are some pathogenic strains, which can for example cause diarrhea, most strains are symbiotic. *E. coli* can grow very fast both under aerobic and anaerobic circumstances and is therefore very well adapted to the environments it lives in, the gut and 'the outside world'. Because *E. coli* is so easy to grow, it is maybe the best studied organism.

Preferentially *E. coli* lives on glucose as carbon source. This can be seen when *E. coli* is grown in an environment with for example glucose and lactose. Under these circumstances *E. coli* first consumes glucose. When glucose is depleted, growth stops, which is called the lag-phase. After some time the cells start lactose consumption and again start growing. This phenomenon was discovered by Monod (1942), who called it diauxic growth. Monod realized that this meant that

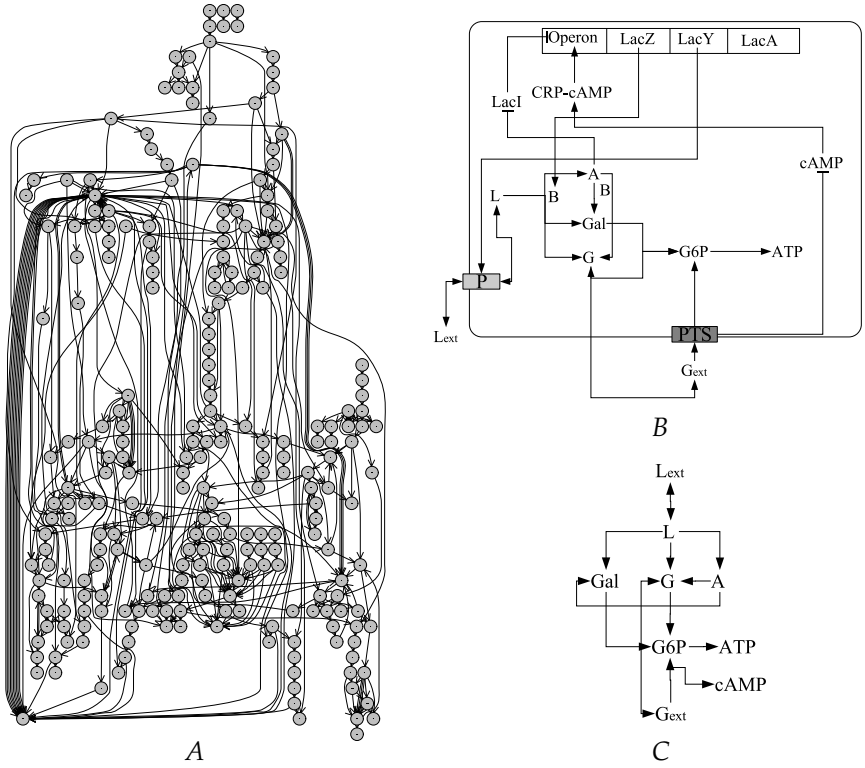


Figure 1.1: (*A*) A part of the metabolic network of *S. cerevisiae*. Nodes indicate metabolites, edges reactions. Only reactions that are active during anaerobic growth are shown. (*B*) An overview of the *lac* operon. (*C*) An overview of the metabolic network of the *lac* operon, where, as in panel *A*, only the metabolites are shown.

E. coli could adapt its metabolism to the environment (see for a good historical account Müller-Hill (1996)). Such a form of adaptation had already previously been observed in yeast by Dienert (1900), for the sugars galactose and glucose (for a good account of the history on yeast research see Barnett (2004)).

How *E. coli* is able to perform this adaptation was discovered later, by the discovery of the *lac* operon by Jacob and Monod (Jacob *et al.*, 1960). It was found that the adaptation was due to the regulation of gene expression and in this way the *lac* operon became the paradigm system of genetic regulation.

It turned out that *E. coli* has two genes that are crucial for lactose metabolism, LacZ and LacY. LacY codes for a permease protein that transports lactose into the cell and LacZ codes for β-galactosidase, which degrades lactose to glucose and galactose. It was found that the expression of both these genes is very strongly correlated and therefore it was proposed that both genes were controlled by one "operator" and formed one "operon". It was however still an open ques-

tion whether the induction by lactose was caused by positive or negative control. Pardee, Jacob and Monod found out that the *lac* operon was controlled via negative control (Pardee & Monod, 1959), i.e. in the presence of lactose, inhibition of the operon was stopped.

This model of negative control soon became a universal model of genetic regulation. It was assumed that negative control was the rule and examples of positive control were assumed to be incorrect (Beckwith, 1996; Müller-Hill, 1996). This shows how influential the work of Jacob and Monod, for which they, together with Lwoff, won the Nobel Prize in Physiology or Medicine in 1965. Interestingly, in the speech of the Nobel Committee it is stated that: "Control of gene activity is thus of a negative nature." However, soon the first example of positive control was found (Englesberg *et al.*, 1965).

Later, Savageau showed why organisms would sometimes choose for negative control, sometimes for positive control. He showed that metabolites that are often present in the environment are positively controlled, while metabolites that are seldom present are negatively controlled (Savageau, 1974, 1977). The reason for this is that proteins more easily lose binding affinity with DNA than gain it. Therefore, if a gene needs to be active most of the time, positive control mutants will lose fitness and positive control will be kept in the population. For negative control however, the mutants can take over the population. Apparently, the higher mutational load for the "positive controllers", compared to the "negative controllers" does not overrule this argument. If a gene needs to be active very seldom, the argument can be reversed to show that negative control will be selected.

After the work of Jacob and Monod, it became gradually apparent that the *lac* operon was more complicated than had been assumed (see for a good review Reznikoff (1992)). In Fig. 1.2 we give a schematic overview of how our knowledge of the *lac* operon has changed. Jacob and Monod had assumed that both regulation and transcription initiation occurred at the same site. Ippen *et al.* (1968) showed later that the operon had two distinct sites, a "promoter" site, that controlled transcription initiation and an "operator" site, that controlled regulation (Fig. 1.2 *B*).

How the regulation of the operon with respect to the glucose concentration worked was found a decade later. A signaling molecule, cAMP, senses the amount of glucose. At low glucose concentrations, the cAMP concentration is high and vice versa. cAMP binds to a protein (CAP) that induces the operon, indeed an example of positive control (Zubay *et al.*, 1970). In this way, the operon is only transcribed when lactose is available and glucose is exhausted (Fig. 1.2 *C*). Therefore, the regulation of the *lac* operon was viewed as an AND gate: lactose AND NOT glucose.

In the last three decennia it has been shown that matters are still more complicated. The existence of several operator sites has been demonstrated (Reznikoff *et al.*, 1974) (Fig. 1.2 *D*), which can stabilize the repressor-DNA complex. Furthermore, a second CAP-binding site has been found (Schmitz, 1981) (Fig. 1.2 *E*). However, this site may be non-functional. Finally, many different promoter sites have been identified (Maquat & Reznikoff, 1978; Malan & McClure, 1984)

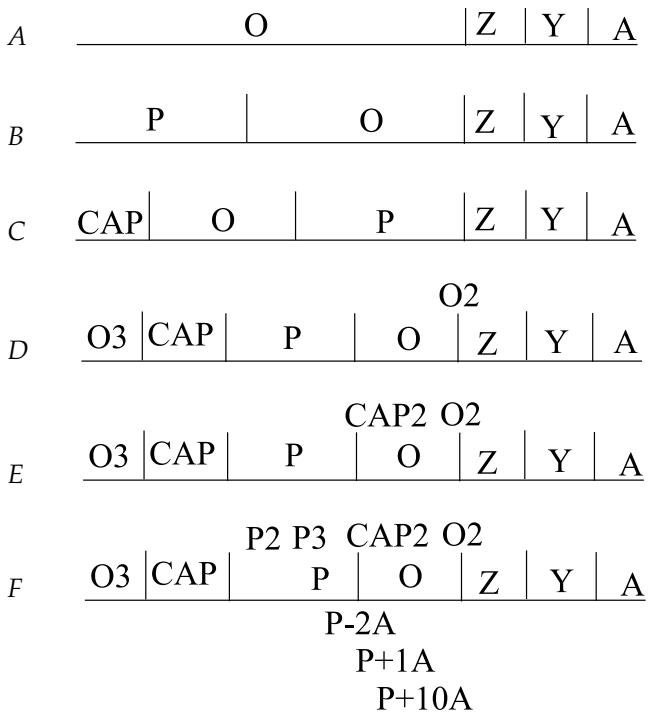


Figure 1.2: A schematic overview of the history of the *lac* operon, adapted from Reznikoff (1992) and Shapiro (1997).

(Fig. 1.2 *F*). The role of these additional promoter sites is also still unknown.

Since its discovery, the *lac* operon has become a paradigm system for genetic and metabolic regulation. The LacZ gene is widely used as a reporter gene. In this manner it is possible to study the activity of other genes. At first, the only way to induce the operon was by growing cells on lactose. Lactose is converted to allolactose by β-galactosidase, and allolactose binds to the repressor protein LacI and deactivates this repressor (Jobe & Bourgeois, 1972). Because lactose is broken down by β-galactosidase, however, the inducer concentration cannot be held constant, which made the *lac* operon hard to study for experimentalists.

Later, two inducers of the operon were found that are not degraded by β-galactosidase themselves, isopropyl β-D-thiogalactopyranoside (IPTG) and thiomethyl β-D-galactoside (TMG). Using these inducers it became possible to study the induction of the *lac* operon more precisely, because the amount of inducer could be held constant.

Furthermore, enzyme activity using lactose was very hard to detect. To solve this problem, it was found that o-nitrophenyl-β-D-galactoside (ONPG) could also be used as inducer. When this molecule is degraded by β-galactosidase, the colorless ONPG changes into the colorless galactose and the yellow o-nitrophenol.

In this way β-galactosidase activity became easy to measure.

Novick & Weiner (1957) found that the *lac* operon exhibits very interesting dynamical behavior. It was observed that there exist certain inducer concentrations for which a population of cells that is already induced, remains induced, while a population of cells that is repressed, remains repressed. This phenomenon is called hysteresis. These concentrations were called "maintenance concentrations". They explained this phenomenon by the presence of a positive feedback loop that is inherently present in the *lac* operon. The operon is turned on by an inducer, which again is transported into the cell by permease, which accounts for the positive feedback loop. Therefore, cells that are repressed, remain repressed and vice versa. It must be noted however that these experiments were performed using the artificial inducer TMG.

1.1.1 Modeling the *lac* Operon

The first mathematical models of gene regulation already date from the sixties. Models for a negative feedback loop (Goodwin, 1965; Griffith, 1968a) and a positive feedback loop were formulated (Griffith, 1968b). It was found that a negative feedback loop can cause oscillations in transcription rates, while a positive feedback can cause bistability: the existence of two stable equilibria in a system. Bistability leads to hysteresis, which mathematically explains the observations of Novick & Weiner (1957). Only recently it has been shown that positive feedback (which could also result from a double negative feedback, or any other combination of feedbacks that leads to a net positive feedback) is a necessary requirement for multistability (Soule, 2003, 2006).

Because the *lac* operon is so well known, it also has a long modeling history. Many models have been formulated that specifically describe genetic regulation of the *lac* operon, all focusing on different aspects of the operon. Babloyantz & Sanglier (1972) were the first to formulate a mathematical model that describes the potential for bistability in the *lac* operon. Most mathematical models have focused on induction by artificial inducers (Barkley *et al.*, 1975; Lee & Bailey, 1984), some considered induction by lactose (Van Dedem & Moo-Young, 1975).

More recently, mathematical models have been published with elaborate parameter estimations (Chung & Stephanopoulos, 1996; Wong *et al.*, 1997; Yildirim & Mackey, 2003; Yildirim *et al.*, 2004; Santillan & Mackey, 2004). Because the *lac* operon has been so intensively studied, many parameters of the system have been experimentally measured. Still however, different experiments sometimes yield very different parameter estimates and caution is needed. Even more recently, stochasticity in gene expression is incorporated in mathematical models of the *lac* operon (Kierzek *et al.*, 2001; Vilar *et al.*, 2003; Vilar & Leibler, 2003)

1.2 *Saccharomyces cerevisiae*

The budding yeast *S. cerevisiae* is a unicellular fungus. Originally the habitat of *S. cerevisiae* is the surface of rotting fruit. As long as 7000 years *S. cerevisiae* has

been used to make wine (Mortimer, 2000). It is also used for baking and brewing, which is where its name, which literally means "sugar mold of beer", comes from. It has even been proposed that *S. cerevisiae* is entirely domesticated and that natural strains are migrants from human-associated fermentation, although there is also evidence for the opposite (Fay & Benavides, 2005).

S. cerevisiae reproduces by budding, in contrast to some other yeast species that reproduce via fission. *S. cerevisiae* is able to reproduce both sexually and asexually. Haploid and diploid cells can reproduce asexually by budding, while two haploid cells can mate to form a diploid cell. Diploid cells again can form haploid spores, which completes the cycle.

Apart from a long history in producing food, *S. cerevisiae* also has a relatively long scientific history, much longer than *E. coli*. As early as 1789, Lavoisier described the chemistry of fermentation by yeast. At that time, however, yeast was not considered to be a living organism (Barnett, 1998). Later, Schwann postulated that yeast is a plant, that lives on glucose and excretes ethanol. In 1861 Louis Pasteur disproved the idea of "spontaneous generation" of yeast and showed that yeast, just as other living organisms, is formed by means of reproduction (Bordenave, 2003). He also showed that yeast was responsible for the fermentation of wine, instead of being a consequence of it. Finally Pasteur showed that in the presence of oxygen, fermentation stops, the Pasteur effect.

Genetic studies of yeast have already been performed in 1935 by Winge (1935). Because *S. cerevisiae* is so easy to grow and is non-pathogenic, it has become maybe the best studied eukaryotic organism. Its genome was the first eukaryotic genome to be completely sequenced (Goffeau *et al.*, 1996). Many databases exist on the Internet where information on the genome, proteome or metabolome of *S. cerevisiae* can be found (e.g. the following websites: MIPS; SGD; YMPD (for the URLs, see the Bibliography)).

1.2.1 Whole Genome Duplication in *Saccharomyces cerevisiae*

Since the whole genome of *S. cerevisiae* has been sequenced, it has become possible to study its genome evolution. It became soon apparent that a whole genome duplication (WGD) occurred in the ancestry of *S. cerevisiae* (Wolfe & Shields, 1997; Kellis *et al.*, 2004). Such a whole genome duplication is not a feature unique to yeast. Evidence for WGDs have been found in plants (Initiative, 2000; Bowers *et al.*, 2003), vertebrates (Dehal & Boore, 2005), teleost fishes (Amores *et al.*, 1998; Taylor *et al.*, 2001) and ciliates (Aury *et al.*, 2006).

In contrast to single gene duplications or segmental duplications, WGDs open a whole range of evolutionary possibilities, because pathways are duplicated as a whole, instead of only parts of a pathway. This has led to the believe that WGDs are responsible for morphological complexity in plants and animals (Furlong & Holland, 2002; Freeling & Thomas, 2006).

In *S. cerevisiae*, the WGD has been associated with the rapid emergence of new yeast species (Scannell *et al.*, 2006). Furthermore, it has been proposed that the WGD caused the transition to the fermentative lifestyle of *S. cerevisiae* (Liti & Louis, 2005; Piskur *et al.*, 2006; Merico *et al.*, 2007). In any case, it is clear that

a whole genome duplication can have enormous consequences on the metabolic capacities of an organism.

1.2.2 Metabolism of *Saccharomyces cerevisiae*

In contrast to other yeast species, when growing in a glucose-rich environment *S. cerevisiae* ferments glucose, even when oxygen is abundant, which is called the Crabtree effect and is the opposite of the Pasteur effect. It has been assumed that this gives a competitive advantage to other yeast species, because, although fermentation is very inefficient, it allows *S. cerevisiae* to grow very fast. Furthermore it produces large amounts of ethanol, to which *S. cerevisiae* is better adapted than other yeast species. Because *S. cerevisiae* naturally lives on rotting fruit, glucose-rich environments are often experienced in the evolution of yeast. At low glucose concentrations, *S. cerevisiae* however does not produce ethanol but only uses aerobic respiration to consume glucose.

When glucose is depleted, *S. cerevisiae* shifts its metabolism to the aerobic metabolism of ethanol. This is again a form of diauxic shift as we also discussed in *E. coli*. It has been shown that this diauxic shift in *S. cerevisiae* is associated with a massive change in gene expression (DeRisi *et al.*, 1997), which indicates the importance of genetic regulation in yeast metabolism.

Metabolic Network of *Saccharomyces cerevisiae*

The whole set of metabolic reactions and metabolites of an organism is called its metabolic network. A metabolic network consists of different pathways, such as glycolysis, TCA-cycle, pentose phosphate pathway, which are all linked to each other.

The modeling of metabolic networks has a long history. Already in 1960 a computer model of the glycolysis in tumor cells was published (Chance *et al.*, 1960). Since then, many mathematical models describing metabolic pathways have been published.

Since the sequencing of whole genomes, it has become possible to construct the "complete" metabolic network of organisms. This has first been done for *Haemophilus influenzae* (Edwards & Palsson, 1999), *E. coli* (Edwards & Palsson, 2000) and *Helicobacter pylori* (Schilling *et al.*, 2002). The first genome-scale metabolic network of a eukaryote was of *S. cerevisiae* (Forster *et al.*, 2003). Later, this metabolic network was extended to include compartmentalization (Duarte *et al.*, 2004a) and genetic regulation (Herrgard *et al.*, 2006).

Modeling Metabolic Networks

There are several mathematical techniques presently available to model metabolic networks. Using differential equations for every metabolite, for example, it is in theory possible to model every metabolic network. When a metabolic network becomes very large however, such a detailed, parameter-rich model is no

longer a very sensible choice, because normally the amount of unknown parameters will be huge (an exception is the human red blood cell, see for example Joshi & Palsson (1989)). Therefore, when metabolic networks become very large, different approaches are needed. One of these approaches is Flux Balance Analysis (FBA).

FBA is a constraint-based modeling approach. This means that, instead of trying to describe a system dynamically, FBA uses constraints (stoichiometric, physical etc.) to narrow down the possible flux distributions in a certain network. By assuming that the metabolic network is in equilibrium, FBA ignores the dynamics of the network (which vastly decreases the total amount of parameters) and furthermore constraints the possible flux distributions. This is done in the following way: the whole metabolic network can be described by a set of differential equations

$$\frac{d\vec{x}(t)}{dt} = S \cdot \vec{v}(t), \tag{1.1}$$

where $\vec{x}$ is a vector describing the concentrations of all metabolites in the system, S is the stoichiometric matrix and $\vec{v}$ is the vector describing all fluxes in the network. The construction of a metabolic network is equivalent to finding the stoichiometric matrix S. The stoichiometric matrix indicates which metabolites can be converted into which and gives the corresponding stoichiometric constants. When all reactions are in equilibrium, Eq. 1.1 simply amounts to

$$S \cdot \vec{v} = 0. \tag{1.2}$$

This is a set of linear equations and is therefore easy to solve. However, the number of equations (which is equal to the number of metabolites) is generally smaller than the number of variables (which is equal to the number of reactions). Therefore, the system is under-determined and there are infinitely many solutions.

The idea of FBA now is to add additional constraints in order to narrow down the solution space, such as maximal flux values for certain reactions. Furthermore, in FBA the solution space is narrowed down by optimizing a certain function of variables of the system, which is called the objective function. If this objective function is a linear function of the variables (the fluxes through the system) ($Z = \vec{c} \cdot \vec{v}$, with $\vec{c}$ the vector which defines the weight of every flux to the objective function), this constitutes a linear programming (LP) problem, which can be relatively easily solved. Common objective functions are biomass formation and the amount of ATP produced. It has been shown that optimizing for biomass can correctly predict gene deletion results (Edwards & Palsson, 2000; Schilling *et al.*, 2002; Famili *et al.*, 2003) and metabolic by-product secretion (Famili *et al.*, 2003).

1.3 This Thesis

In this thesis we study the evolution of metabolic adaptation from the single gene to the level of the whole metabolic network. In this way we approach the problem of evolution of metabolism from two extreme perspectives. In **part I** we start at

the single gene level by studying the evolution of the *lac* operon. In **part II** we study evolution of metabolism from the perspective of the whole organism by studying the evolution of the metabolic network of *S. cerevisiae* after its whole genome duplication.

The *lac* operon, as we explained in this introduction, is a very small subsystem of the metabolic network of *E. coli*. Because it is so well-studied, we can model the *lac* operon in much detail. This gives us insight into how the regulation of even one operon is not a trivial task. Setty *et al.* (2003) found that the regulation of the *lac* operon is not the classical AND gate as we explained in section 1.1 (lactose AND NOT glucose). They measured the promoter activity of the *lac* operon as a function of two variables, lactose and glucose. The promoter function they observed was very intricate: it had four different plateau values and the switches were shallow, such that there existed concentrations with intermediate expression levels.

These findings appear to be in contrast with the findings of Novick & Weiner (1957), who describe enzyme induction in the *lac* operon as an all-or-nothing phenomenon. It must be noted however that Setty *et al.* (2003) measured the population average of the transcription rates. Therefore, part of the findings of Setty *et al.* (2003) could be explained by population heterogeneity.

Nevertheless, this raises the question whether the *lac* operon is such an intricate promoter function. To answer this question, in **chapter 2** we studied the evolution of the *lac* operon in a spatial, fluctuating environment. We find that in such an environment, indeed a promoter function similar to the one observed by Setty *et al.* (2003) evolves. Furthermore we find that the intermediate expression levels in this promoter function are of crucial importance, because they allow for a graded response with respect to the extracellular environment. We found that such a graded response allows for a much more rapid response to the environment than a discontinuous (bistable) response.

However, such a bistable response has been observed in experiments by Novick & Weiner (1957) and recently been confirmed by Ozbudak *et al.* (2004). We also show in **chapter 2** that these results can be explained when we realize that the artificial inducer TMG was used in these experiments. In contrast to lactose, artificial inducers are not degraded by β-galactosidase and therefore the positive feedback loop is much stronger for artificial inducers than for lactose. A strong positive feedback loop can cause bistability and therefore a discontinuous instead of a graded response to the environment. The *in silico* evolved promoter function indeed behaved bistable with respect to artificial inducers, but not with respect to lactose.

In **chapter 2** we used a deterministic model for *lac* operon dynamics. However, gene expression is inherently stochastic and indeed mRNA and protein numbers can be very low in bacterial cells. Furthermore, it has been proposed that stochasticity can render bistability advantageous, because it allows switching between the two equilibria. This again leads to population heterogeneity in the population, which can be beneficial under certain circumstances (see for example Thattai & Van Oudenaarden (2004)).

Therefore it is important to study the effect of stochasticity on the evolution

of the *lac* operon, the subject of **chapter 3**. We find that the *lac* operon evolves to minimize stochasticity in gene expression. This is achieved by evolving a higher repressed transcription rate than in the deterministic model. Again it turns out that we cannot understand the evolution and dynamics of the *lac* operon if we assume that it can be either "on" or "off". We also show that stochasticity in the *lac* operon is higher when induced by artificial inducers than by lactose, again due to the fact that the positive feedback loop is stronger for artificial inducers.

In **part I** we have shown that the precise dynamics of even such a small system as the *lac* operon can only be understood by detailed evolutionary modeling and that the precise form of genetic regulation can be crucial for the dynamics. In **part II** we are interested in the behavior of the whole metabolic network of *S. cerevisiae*. However, dynamical modeling of a whole metabolic network, as we explained in section 1.2.2, is unfeasible. Therefore we can only study the steady state flux distributions of the network. However, when we study metabolic networks we are not interested in the precise dynamics of every reaction, but we want to know which reactions are active under certain circumstances, which can be studied using a steady state assumption.

In **part II** we study the evolution of yeast after its whole genome duplication (WGD) that occurred approximately 100 million years ago. We do this from two different perspectives, mutation and selection. In **chapter 4** we study what kind of mutations are responsible for the massive gene loss that occurred after the WGD in yeast and whether this pattern of gene loss can be understood by the mutational mechanism alone.

It turns out that we can and we find that the pattern of gene deletions can be entirely explained by the mutational dynamics. Deletion of stretches of base pairs (in the order of a few hundred base pairs) cause most genes to be deleted on their own, but some genes to be lost simultaneously. Furthermore we show that the sizes of base pair deletions in pseudogenes in *S. cerevisiae* are in the right order of magnitude to be responsible for the simultaneous gene deletions.

In **chapter 5** we study the evolution of a metabolic network after a WGD, from a selectionist point of view. We are interested in how a WGD can change the metabolic network of a cell. In yeast, for example, the WGD occurred roughly simultaneously with the first appearance of the angiosperms (flowering plants) and it has been speculated that the WGD gave yeast the possibility to specialize on glucose-rich fruits (Conant & Wolfe, 2007).

Therefore, we developed a metabolic model, based on FBA, that simulates the WGD and subsequent gene loss in the metabolic network of *S. cerevisiae*. This model is based on a previously published genome-scale metabolic model of *S. cerevisiae* (Duarte *et al.*, 2004a). Using this model we find that we can satisfactorily predict the evolutionary outcome of the WGD in *S. cerevisiae*.

We find that genes that are never retained in duplicate in our model have a smaller probability to be retained in duplicate in *S. cerevisiae* than genes that are often retained in duplicate in our model. Furthermore we find that transporter and glycolysis genes are retained more often in duplicate after WGD, both in the data and in our model, which causes an increase in glycolytic flux (see also Conant & Wolfe (2007)). Finally we found that, when a cell is not yet perfectly

adapted to an environment, a WGD can infer an immediate fitness advantage. All these results confirm the hypothesis that the WGD has helped yeast to adapt to the newly arisen environment of glucose-rich fruits.

Part I

Evolution of the *lac* Operon of *Escherichia coli*

2

In Silico Evolved *lac* Operons Exhibit Bistability for Artificial Inducers, but Not for Lactose

M.J.A. van Hoek and P. Hogeweg
Theoretical Biology/Bioinformatics Group, Utrecht University
Padualaan 8, 3584 CH Utrecht, The Netherlands.

Biophys J. **91(8)**: 2833-43 (2006 Oct)

Abstract

Bistability in the *lac* operon of *Escherichia coli* has been widely studied, both experimentally and theoretically. Experimentally, bistability has been observed when *E. coli* is induced by an artificial, nonmetabolizable, inducer. However, if the *lac* operon is induced with lactose, the natural inducer, bistability has not been demonstrated. We derive an analytical expression that can predict the occurrence of bistability both for artificial inducers and lactose. We find very different conditions for bistability in the two cases. Indeed, for artificial inducers bistability is predicted, but for lactose the condition for bistability is much more difficult to satisfy. Moreover, we demonstrate that *in silico* evolution of the *lac* operon generates an operon that avoids bistability with respect to lactose, but does exhibit bistability with respect to artificial inducers. The activity of this evolved operon strikingly resembles the experimentally observed activity of the operon. Thus our computational experiments suggest that the wild-type *lac* operon, which regulates lactose metabolism, is not a bistable switch. Nevertheless, for engineering purposes, this operon can be used as a bistable switch with artificial inducers.

2.1 Introduction

Since the discovery of the *lac* operon of *Escherichia coli* (Jacob *et al.*, 1960), it has been a model study for genetic regulation. The *lac* operon simultaneously regulates the transcription of three genes, LacZ, LacY, and LacA. Only LacZ and LacY are important for lactose utilization. LacZ codes for β-galactosidase, the protein responsible for lactose degradation and LacY codes for a membrane permease protein, which transports the lactose into the cell.

The expression of the *lac* operon depends on the internal concentration of two molecules, allolactose and cAMP. Allolactose is derived from lactose, while glucose influx into the cell represses cAMP. Both allolactose and cAMP induce the operon and therefore, classically, the *lac* operon is described as a Boolean function: lactose AND NOT glucose.

Both the permease and β-galactosidase are needed to produce allolactose, and allolactose again induces the operon. Therefore, there is an inherent positive feedback loop in the system that can cause bistability (Griffith, 1968b). This bistability is caused by a fold bifurcation.

This bistable behavior has already been observed by Novick & Weiner (1957). It was observed that a genetically identical population of *E. coli* can be heterogeneous in its activity of the operon, while the environment is the same for the whole population. These experiments were performed using an artificial inducer of the operon. These inducers, such as isopropyl β-D-thiogalactopyranoside (IPTG) and thiomethyl β-D-galactoside (TMG), cannot be metabolized by β-galactosidase, in contrast to lactose, the natural inducer.

Recent advances in experimental techniques that allow direct measurement of the promoter activity have shed new light on the *lac* operon. The activity of the operon in living cells has been measured (Setty *et al.*, 2003). It was found that the operon is an intricate function of the cAMP and inducer concentration, with four different threshold values and plateau transcription levels. Furthermore, shallow switches were observed, which need approximately a 10-fold change in cAMP or inducer concentration, all in contrast to the classical AND-gate. Furthermore, bistability of the *lac* operon on the single cell level has been observed (Ozbudak *et al.*, 2004). This study confirmed previous findings that the operon behaves bistably with respect to artificial inducers. However, interestingly, bistability could not be demonstrated when growing on lactose, the natural inducer of the operon. Finally, it has been shown that *E. coli* can optimize its operon activity in only a few hundred generations (Dekel & Alon, 2005).

Several theoretical models, using detailed parameter values, have been developed that explain the occurrence of bistability in the *lac* operon, both for induction by artificial inducers (Chung & Stephanopoulos, 1996) and by lactose (Yildirim & Mackey, 2003). These detailed theoretical models have two important drawbacks. First, the outcome depends highly on the parameter values used. Notwithstanding extensive experimental work, these parameter values are still not very reliable. Second, these detailed models are not analytically solved, and therefore it is difficult to know which parts of the model contribute to the bistability. Therefore, the modeling is still far from conclusive.

Here we try to avoid these shortcomings in two ways. First, by making an approximation of our model we are able to find an analytical expression that predicts under which conditions operons are bistable, both for artificial inducers and lactose. Second, by *in silico* evolution of the *lac* operon in a fluctuating environment of glucose and lactose, we let cells adapt their operons to the assumed biochemical parameters and their environment. In this way we can study under which environmental circumstances bistability will evolve.

The analytical expression we find indeed predicts bistability to occur for artificial inducers, whereas it also predicts that for lactose it is much more difficult to have a bistable switch. In our evolutionary simulations we find that, when starting with a bistable population, evolution drives the population out of the bistable region. When growing on artificial inducers, however, these evolved operons do behave bistably. Furthermore, the phase diagram of the *in silico* evolved *lac* operon is quantitatively very similar to the experimentally observed phase diagram (Ozbudak *et al.*, 2004).

We performed evolutionary simulations in different environments. Sometimes bistability at very high glucose concentrations was observed, but for lower glucose concentrations bistability was always avoided.

Our findings put the occurrence of bistability in the *lac* operon in a different perspective. Relative to lactose, bistability is avoided, and bistability relative to artificial inducers is a side-effect of evolution on lactose, the natural inducer of the *lac* operon.

2.2 Methods

2.2.1 The Dynamics of the *lac* Operon.

We developed a differential equation model, based on the model of Wong *et al.* (1997), to describe the dynamics of the *lac* operon. The model consists of 10 differential equations describing glucose and lactose metabolism and regulation. Here we shortly describe the model equations. In the Supplementary Material the rationale of the model is explained in more detail.

Five of these 10 differential equations are important for the bistability in this system, namely the equations describing mRNA, M; β-galactosidase, B; permease, P; internal lactose, L; and allolactose, A. mRNA production is modeled after Setty *et al.* (2003), while the other four equations are modeled after Wong *et al.* (1997).

$$PA(A,C) \equiv V_1 \frac{1 + \frac{V_2(C/k_C)^n}{1+(C/k_C)^n} + \frac{V_3}{1+(A/k_A)^m}}{1 + \frac{V_4(C/k_C)^n}{1+(C/k_C)^n} + \frac{V_5}{1+(A/k_A)^m}} \quad (2.1)$$

$$\frac{dM}{dt} = min(PA(A,C), V_{mRNA,max}) - (\gamma_M + \mu)M \quad (2.2)$$

$$\frac{dB}{dt} = k_B M - (\gamma_B + \mu)B \quad (2.3)$$

$$\frac{dP}{dt} = k_P M - (\gamma_P + \mu)P \quad (2.4)$$

$$\frac{dL}{dt} = P(\frac{k_{Lac,in}L_{ext}}{K_{Lac,in} + L_{ext}} - \frac{k_{Lac,out}L}{K_{Lac,out} + L}) - B\frac{(k_{cat,Lac} + k_{Lac-Allo})L}{L + K_{m,Lac}} - (\gamma_L + \mu)L \quad (2.5)$$

$$\frac{dA}{dt} = B\frac{k_{Lac-Allo}L}{L + K_{m,Lac}} - B\frac{k_{cat,Allo}A}{A + K_{m,Allo}} - (\gamma_A + \mu)A \quad (2.6)$$

$PA(A,C)$ is defined as the promoter activity as a function of allolactose (A) and cAMP (C) and is a two-dimensional Hill-function, with coefficients m and n. cAMP is dependent on glucose uptake, while allolactose is dependent on lactose uptake. Glucose uptake is determined by the phosphoenolpyruvate: carbohydrate phosphotransferase system (PTS) (Postma *et al.*, 1993). This system transports and phosphorylates external glucose, and cAMP production is repressed by this process; therefore, the internal glucose concentration is inversely related to the cAMP concentration.

Equation 2.2 describes transcription and degradation of mRNA. We impose a maximal transcription rate, similar to the maximal *in vivo* transcription rate (Malan *et al.*, 1984) (see Supplementary Material), to avoid unrealistically high transcription rates during the evolutionary simulations.

The value μ is defined as the growth rate of the cell. We assume first-order degradation for all chemicals. Equations 2.3 and 2.4 describe translation of mRNA to β-galactosidase and permease. Equation 2.5 describes reversible lactose influx by permease, where L_{ext} is the external lactose concentration, and the degradation of internal lactose by β-galactosidase.

Inducer exclusion is not taken into account in our model for two reasons. First, lactose efflux does not depend on the external glucose concentration (Wong *et al.*, 1997), because inducer exclusion only affects lactose influx. We will later show that only lactose efflux determines bistability in the *lac* operon. Secondly, although bistability is not affected by inducer exclusion, it does have an effect on the evolution of the promoter function. When inducer exclusion would be taken into account, cells would never experience high internal glucose and lactose concentrations simultaneously. Therefore, there would be no evolutionary pressure on this part of the promoter function. Because we did not want to impose any form of regulation, we did not take inducer exclusion into account.

Equation 2.6 describes production and degradation of allolactose by β-galactosidase. Note that, in contrast to Wong *et al.* (1997), we added an operon-independent degradation rate of lactose and allolactose. Without these terms, bistability would only depend on the growth rate. When the growth rate then would be zero, bistability would not be possible, because the lactose and allolactose steady states would be independent of B and P.

Apart from these five differential equations that determine bistability, five more differential equations are needed to describe glucose uptake, metabolism, and growth. β-galactosidase degrades lactose to glucose and galactose. Internal glucose can be phosphorylated or, if the internal glucose concentration is very high (Hogema *et al.*, 1999), excreted to the medium. This gives for the internal glucose concentration

$$\begin{aligned}\frac{dG}{dt} &= \frac{k_{cat,Lac}BL}{L+K_{m,Lac}} + \frac{k_{cat,Allo}BA}{A+K_{m,Allo}} \\ &\quad - \frac{k_{cat,Glu}G}{G+K_{m,Glu}} - k_{Glu,out}(G-G_{ext}) - \mu G. \end{aligned} \tag{2.7}$$

Besides internal glucose, glucose-6-phosphate is modeled separately. External glucose is transported inside the cell and phosphorylated by the PTS. Galactose formed by lactose degradation is also converted to glucose-6-phosphate. The glucose formed by lactose degradation is phosphorylated, as modeled in Eq. 2.7. Finally, glucose-6-phosphate is degraded by respiration and fermentation, such that glucose-6-phosphate is first respired and if the respiratory pathways are saturated, overflow is fermented (Andersen & Von Meyenburg, 1980):

$$\begin{aligned}\frac{dG6P}{dt} &= \frac{k_{t,Glu}G_{ext}}{G_{ext}+K_{t,Glu}} + \frac{k_{cat,Lac}BL}{L+K_{m,Lac}} + \frac{k_{cat,Allo}BA}{A+K_{m,Allo}} + \frac{k_{cat,Glu}G}{G+K_{m,Glu}} \\ &\quad - \frac{k_{G6P,Rsp}G6P}{G6P+K_{G6P,Rsp}} - \frac{k_{G6P,Frm}G6P^8}{K^8_{G6P,Frm}+G6P^8} - \mu G6P. \end{aligned} \tag{2.8}$$

Transcription of the *lac* operon depends on the cAMP concentration, which again depends inversely on glucose transport by the PTS. This is modeled as

$$\frac{dC}{dt} = k_{syn,cAMP}\frac{K_{syn,cAMP}}{\frac{k_{t,Glu}G_{ext}}{G_{ext}+K_{t,Glu}}+K_{syn,cAMP}} - (\gamma_{cAMP}+\mu)C. \tag{2.9}$$

Furthermore, we need a measure for the amount of energy the cells consume and produce, which determines their growth rate; this is done via ATP. Energy-producing actions are respiration and fermentation. Energy-consuming actions are basal metabolism, growth, cost for *lac* operon activity, and a cost to convert galactose to glucose, which then gives

$$\frac{dATP}{dt} = \frac{Y_{Rsp}k_{G6P,Rsp}G6P}{G6P + K_{G6P,Rsp}} + \frac{2k_{G6P,Frm}G6P^8}{K^8_{G6P,Frm} + G6P^8} - BMC - GC \times \mu$$
$$-PC \times PA(A,C) - \frac{k_{cat,Lac}BL}{L + K_{m,Lac}} - \frac{k_{cat,Allo}BA}{A + K_{m,Allo}}. \quad (2.10)$$

To determine the growth rate of the bacteria, we assume a relationship between the amount of ATP and the growth rate. We assume a sigmoid relationship between energy and growth, such that cell growth becomes

$$\frac{dX}{dt} = \mu_{max}\frac{ATP^4}{ATP^4 + K^4_{ATP}}X. \quad (2.11)$$

2.2.2 The Evolutionary Model

To study the evolution of the *lac* operon we developed an individual-oriented spatial model. A detailed description of this model can be found in the Supplementary Material.

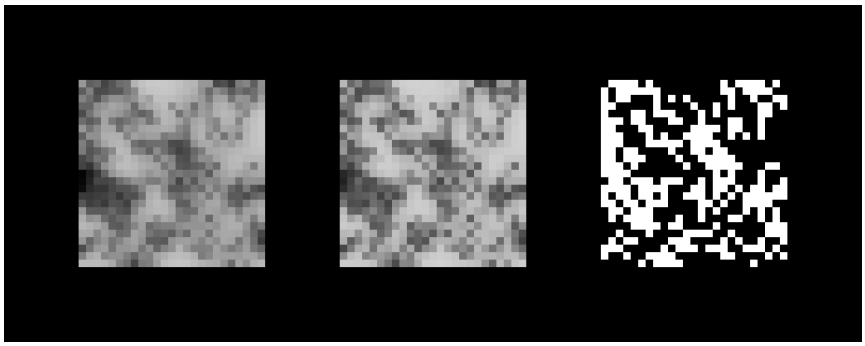
The intracellular dynamics are determined by all 10 differential equations described in the previous section. A population of cells evolves the parameters that determine their promoter activity. These are the parameters of Eq. 2.1. The five different V-parameters consist of seven biological binding parameters (see Supplementary Material). These seven plus k_A, k_C, n, and m are being evolved, while all other parameters are kept constant.

Space has been shown to be crucial for the evolution of metabolic regulation (Pfeiffer *et al.*, 2001). Without space, bacteria would never evolve to an efficient use of metabolites, but rather always consume the amount of metabolites that optimizes their growth rate at that time. Therefore, we choose to develop a spatially explicit model.

We model a population of a few hundred cells, on a square grid of 25 x 25 grid points, growing on a fluctuating environment of glucose and lactose. Because of the small population size, these cells should be interpreted as a colony of identical cells. In Fig. 2.1, a screen-shot of the model is shown.

These cells consume glucose and lactose according to the model described in the previous section. The cells divide (with possible mutation) if they have grown twice their original size, "die" in a density-dependent way (or when their energy, as described in Eq. 2.10 drops below zero), and perform a random walk across the grid.

Because we want the cells to adapt to a fluctuating environment, glucose and lactose influx are modeled in periods with and without glucose and lactose influx, which are independent of each other. The lengths of these periods are determined stochastically, but all have on average an equal length. The duration of the periods with and without glucose or lactose are chosen such that the cells can just adapt their protein levels to the new environment. However, we checked the



A *B* *C*

Figure 2.1: A screen-shot of the grid. (*A*) Glucose concentration. (*B*) Lactose concentration. (*C*) The cells on the grid.

results for longer and shorter periods. The glucose and lactose influx is homogeneously over the grid. Due to metabolism by the cells, the glucose and lactose concentration become heterogeneous (see Fig. 2.1). Glucose and lactose diffuse over the grid.

A list of all parameters used is given in the Supplementary Material (Table 2.1). Quite different parameter values for intracellular dynamics are used in different models of the *lac* operon (Wong *et al.*, 1997; Yildirim & Mackey, 2003; Santillan & Mackey, 2004). We use the parameter values also used in Wong *et al.* (1997). Because the cells can adapt to these parameters, their precise values are less crucial for the outcome. We propose that evolutionary modeling is a good way to deal with the inevitable parameter uncertainty.

We perform three independent simulations for each condition. For each simulation, we trace the last common ancestor. A common ancestor is an individual cell that has all cells at some later time as progeny. Hence the last common ancestor is the last individual cell that has all cells at the end of the simulation as progeny.

To obtain reliable results, while minimizing computational cost, we compete the last common ancestor of each simulation against each other. This is done by initializing a grid with an equal density of two common ancestors. These ancestors are randomly distributed over the grid and the mutation rates are set to zero. If one population dies out, the other population has won the competition. In this way we compete all three last common ancestors pairwise against each other at least 10 times. The simulation yielding the best competitor is chosen to represent that particular condition.

2.3 Results

2.3.1 Approximation of the Differential Equations Describing Bistability

The equations describing bistability from the previous section can be approximated, such that the cusp bifurcation can be analytically found. Note that only Eqs. 2.2-2.6 determine the bistability of the system.

We do not take the cutoff of the transcription rate at $V_{mRNA,max}$ into account. The evolved promoter functions always have maximal transcription rates close to $V_{mRNA,max}$. Furthermore, we will show that bistability with respect to lactose is only determined by the repressed transcription rate.

First we derive an approximation of the equilibrium of Eqs. 2.2-2.6. The first three equations are trivial and for the lactose equilibrium we have the equation

$$\begin{aligned} &\overline{P}\left(\frac{k_{Lac,in}L_{ext}}{K_{Lac,in}+L_{ext}} - \frac{k_{lac,out}\overline{L}}{K_{Lac,out}+\overline{L}}\right) = \\ &(k_{Lac-Allo}+k_{cat,Lac})\overline{B}\frac{\overline{L}}{\overline{L}+K_{m,Lac}} + (\mu+\gamma_L)\overline{L}, \end{aligned} \tag{2.12}$$

where $\overline{B}$, $\overline{P}$ and $\overline{L}$ denote equilibrium values. In the bistable region $\overline{L}$ is small compared to the different k-values, and we can neglect the saturated behavior of the model. Furthermore we use $\overline{B} = \frac{k_B PA(\overline{A},C)}{(\gamma_B+\mu)(\gamma_M+\mu)}$ and $\overline{P} = \frac{k_P PA(\overline{A},C)}{(\gamma_P+\mu)(\gamma_M+\mu)}$ to get

$$\overline{L} = \frac{\xi L_{ext}}{K_{Lac,in}+L_{ext}}\frac{PA(\overline{A},C)}{\zeta PA(\overline{A},C)+1}, \tag{2.13}$$

$$\xi \equiv \frac{k_P k_{Lac,in}}{(\gamma_P+\mu)(\gamma_M+\mu)(\gamma_L+\mu)}, \tag{2.14}$$

$$\zeta \equiv \frac{\left(\frac{k_P k_{Lac,out}}{K_{Lac,out}(\gamma_P+\mu)} + \frac{k_B(k_{cat,Lac}+k_{Lac-Allo})}{K_{m,Lac}(\gamma_B+\mu)}\right)}{(\gamma_L+\mu)(\gamma_M+\mu)}. \tag{2.15}$$

From Eq. 2.6 we approximate the allolactose equilibrium up to first order in $\overline{L}$. To do this we again assume that $\overline{A}$ is small compared to the different k-values and $(\gamma_A+\mu) << \frac{k_{cat,Allo}\overline{B}}{K_{m,Allo}}$. So we assume a linear relationship between the lactose and allolactose concentration and we find

$$\overline{L} = \frac{k_{cat,Allo}K_{m,Lac}}{k_{Lac-Allo}K_{m,Allo}}\overline{A}. \tag{2.16}$$

$\overline{A}$ can now be found by equating Eq. 2.13 and Eq. 2.16

$$PA(\overline{A},C) = \frac{\overline{A}}{\frac{\xi L_{ext}}{K_{Lac,in}+L_{ext}}\frac{k_{Lac-Allo}K_{m,Allo}}{k_{cat,Allo}K_{m,Lac}} - \zeta\overline{A}}. \tag{2.17}$$

This equation gives the equilibrium allolactose concentration for all cAMP concentrations. To study under which conditions bistability occurs, we want to calculate the cusp bifurcation in this model. Writing $x \equiv \overline{A}/k_A$ we get

$$\frac{V_1(1+V_2\mathcal{A}+\frac{V_3}{1+x^m})}{1+V_4\mathcal{A}+\frac{V_5}{1+x^m}} = \frac{x}{\frac{\xi L_{ext}}{K_{Lac,in}+L_{ext}}\frac{k_{Lac-Allo}K_{m,Allo}}{k_{cat,Allo}K_{m,Lac}k_A}-\zeta x}, \tag{2.18}$$

where $A = \frac{(C/k_C)^n}{1+(C/k_C)^n}$. If m were an integer this equation would become a polynomial of power $m+1$, but m is an evolvable parameter, so it has real, noninteger values. Defining $\theta \equiv \frac{\xi L_{ext}}{K_{Lac,in}+L_{ext}}\frac{k_{Lac-Allo}K_{m,Allo}}{k_{cat,Allo}K_{m,Lac}k_A}$ we can rewrite the above to

$$f(x) = ax^{m+1} - bx^m + cx - d = 0, \tag{2.19}$$

where

$$\begin{aligned} a &\equiv 1+V_4\mathcal{A}+\zeta V_1(1+V_2\mathcal{A}),\\ b &\equiv \theta V_1(1+V_2\mathcal{A}),\\ c &\equiv 1+V_4\mathcal{A}+V_5+\zeta V_1(1+V_2\mathcal{A}+V_3),\\ d &\equiv \theta V_1(1+V_2\mathcal{A}+V_3). \end{aligned} \tag{2.20}$$

For the cusp bifurcation to take place, $f(x)$, $f'(x)$, and $f''(x)$ need to be zero for the same x-value. So to calculate the condition for the cusp bifurcation we now have to solve the equations

$$\begin{aligned} f(x) &= ax^{m+1} - bx^m + cx - d = 0,\\ f'(x) &= a(m+1)x^m - mbx^{m-1} + c = 0,\\ f''(x) &= a(m+1)mx^{m-1} - bm(m-1)x^{m-2} = 0 \end{aligned} \tag{2.21}$$

simultaneously. We then find

$$c = b\left(\frac{b(m-1)}{a(m+1)}\right)^{m-1}, d = a\left(\frac{b(m-1)}{a(m+1)}\right)^{m+1}. \tag{2.22}$$

From these two equations, the bifurcation parameter L_{ext} can be eliminated by eliminating θ. This is what we want, because we want to know under which parameter conditions a fold bifurcation occurs for a certain value of L_{ext}. After substituting Eq. 2.20 back, we find, after a long but straightforward calculation, the following condition for the cusp bifurcation, dependent on the cAMP concentration:

$$\lambda(C) \equiv \frac{PA(0,C)}{(m-1)^2}\left(\frac{(m+1)^2}{PA(\infty,C)}+4m\zeta\right) < 1. \tag{2.23}$$

If $\lambda(C) < 1$, a fold bifurcation occurs for a certain value of L_{ext}. After the approximations made for the allolactose and lactose equilibrium, this result is

exact. We checked this formula for many promoter functions, and the predictions whether the promoter function is bistable are very good. Generally, due to the first-order approximation of the allolactose equilibrium, bistability sometimes also occurs if $\lambda(C)$ is slightly larger than one, but if $\lambda(C)$ is larger than two, bistability never occurs. In the evolutionary simulations, the fluctuations in $\lambda(C)$ are much larger than a factor two, and therefore the approximation is accurate.

From Eq. 2.15, we observe that ζ determines the ratio of internal lactose that is degraded or transported by operon activity and internal lactose that is degraded, transported, or diluted independently from the operon, $(\gamma_L + \mu)$. The first term of ζ describes lactose efflux by permease and the second lactose degradation via β-galactosidase. Therefore, the only terms that effect bistability are the lactose efflux and degradation terms, and we indeed see that lactose influx, hence inducer exclusion, is not important for bistability.

Artificial inducers are, in contrast to lactose, not degraded by β-galactosidase. This means that ζ is much smaller for artificial inducers than for lactose. The values $k_{cat,Lac}$ and $k_{Lac-Allo}$ are zero, and $k_{Lac,out}$ and $K_{Lac,out}$ will have different values for artificial inducers. These values are reported for the artificial inducer IPTG[1] (Cheng *et al.*, 2001). The value of γ_L is not known, but because there is no degradation of internal inducer by β-galactosidase, there will probably exist a different pathway to degrade the artificial inducer. But even if γ_L is small, we still find that ζ is very small, and Eq. 2.23 can be approximated by

$$\frac{PA(0,C)}{(m-1)^2}\frac{(m+1)^2}{PA(\infty,C)} < 1. \tag{2.24}$$

If we now define the repression factor ρ as the ratio of maximal and basal activity, this can be rewritten to

$$\rho(C) \equiv \frac{PA(\infty,C)}{PA(0,C)} > \frac{(m+1)^2}{(m-1)^2}. \tag{2.25}$$

This limit for artificial inducers is in agreement with a previous study (Ozbudak *et al.*, 2004), where $\rho > 9$ was derived and a value of $m = 2$ was used. This result is also in agreement with their experiments, which show bistability with artificial inducers for $\rho > 9$, but a continuous response for $\rho \approx 5$. *E. coli* has been reported to have a repression factor > 100 (Ozbudak *et al.*, 2004; Setty *et al.*, 2003); therefore, bistability is indeed predicted for artificial inducers.

For lactose, the natural inducer, however, the situation is quite different. Because internal lactose is degraded by β-galactosidase, the fraction between operon-dependent and -independent lactose degradation, ζ, is much larger and the ζ term dominates Eq. 2.23:

$$\frac{PA(0,C)4m\zeta}{(m-1)^2} < 1. \tag{2.26}$$

[1] In this article we did not take into account that IPTG is also transported into the cell independently from permease, which reduces the bistability. However, we did do this in **chapter 3**. There, induction by TMG was modeled in the same way as we modeled induction by IPTG in this chapter. Therefore, in this chapter, when IPTG is mentioned, TMG should be read.

Now we see that, instead of the repression factor ρ, the absolute transcription rate at zero allolactose, $PA(0, C)$, determines whether bistability occurs.

It is not easy to check whether Eq. 2.26 does or does not hold for the wild-type *lac* operon, because this depends, much more than for artificial inducers, on the detailed parameter values. However, we can compare ζ_{art} and ζ_{lac}, if we use the same value for γ_L, a conservative estimate. We find that $\zeta_{lac} \approx 6000\zeta_{art}$. Therefore, the condition determining bistability for lactose is very difficult to satisfy.

From these findings we conclude that experimental observation of bistability with artificial inducers does not tell much about whether bistability occurs for lactose. Instead of the repression level, the absolute transcription rate at zero allolactose determines whether bistability occurs. But should we expect bistability to occur with lactose as inducer? To study this, we developed the individual-oriented, computational model, explained in the previous section and more detailed in the Supplementary Material, with which we study the *in silico* evolution of the *lac* operon.

2.3.2 In Silico Evolution of the *lac* Operon

Because we expect, from our results in the previous section, that bistability could be difficult to achieve, we start evolution with a bistable population and study whether bistability remains in the population. We choose values of γ_L and γ_A ($\gamma_L = \gamma_A = 0.15/\text{min}$), such that the promoter functions are firmly in the bistable region.

The initial promoter function is plotted in Fig. 2.2 A . Note that initially the induced transcription rate equals $V_{mRNA,max}/2$.

Two bifurcation diagrams, corresponding to the promoter function of Fig. 2.2 A, for minimal ($C_{min} \approx 1.0 \times 10^{-5}$ mM) and maximal ($C_{max} \approx 4.8 \times 10^{-4}$ mM) cAMP, i.e., high and low glucose concentration, respectively, are shown in Fig. 2.2 B. These concentrations can be calculated from Eq. 2.9, using that the minimal and maximal glucose influx are 0 and $k_{t,Glu}$, respectively. For all intermediate cAMP concentrations, the bifurcation diagrams lie between these two extremes. The bifurcation diagrams were numerically calculated using the five equations describing bistability (see Methods) and no approximation was used.

The most important parameter determining the outcome of evolution is the cost for *lac* operon activity. We estimate this cost as follows: the fraction of *lac* operon proteins at full activity of the operon is approximately 3% (Koch, 1983). We assume that the growth rate is also 3% lowered at full activity. We assume a linear relationship between the activity of the operon and the cost. Recently, the cost of the *lac* operon at full activity was measured to be 4-5% (Dekel & Alon, 2005), which compares well with the value we chose. We performed evolutionary simulations for different environments and cost for *lac* operon activity.

Evolution Away From Bistability.

We can use the bifurcation condition $\lambda(C)$ to follow the switching behavior of all cells in the population. In Fig. 2.3 A, we show the population average of the

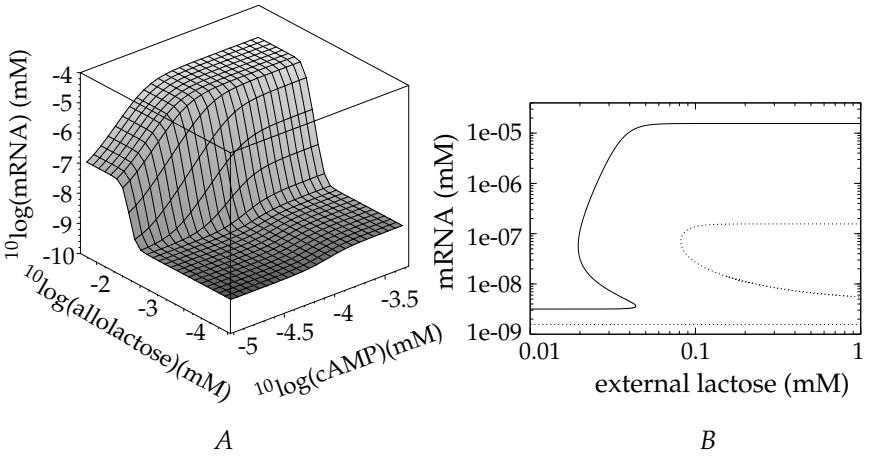


Figure 2.2: (*A*) The initial promoter function of every simulation. (*B*) Two corresponding bifurcation diagrams, for maximal cAMP, i.e. low glucose (*solid line*) and minimal cAMP, i.e. high glucose (*dotted line*). For the bifurcation diagram at high cAMP, low-glucose concentration, a growth rate of $\mu = 0/\text{min}$ is used, while for the bifurcation diagram at low-cAMP, high-glucose a growth rate of $\mu = 0.01/\text{min}$ is used. Bifurcation diagrams however are very similar for different growth rates, because we took $\gamma_L = \gamma_A \gg \mu$.

cusp parameter for minimal and maximal cAMP concentration. Only the population average of $\lambda(C)$ is shown, but the fluctuations in the population are small compared to the fluctuations in the population average and the population is always unimodal. Furthermore note that the fluctuations in $\lambda(C)$ are much larger than the error in the approximation determining $\lambda(C)$. In Fig. 2.3 *B*, the cAMP concentration at which the cusp bifurcation takes place is shown. This cAMP concentration can be calculated using Eq. 2.26. If bistability occurs for all cAMP concentrations, we, by default, assign C_{max} (and likewise, C_{min}, if bistability occurs for no cAMP concentration).

We observe that initially evolution drives $\lambda(C)$ rapidly away from the bistable region. This happens for all cAMP concentrations. This behavior is seen in all three independent simulations. In this particular simulation, after only approximately 3000 days the population is back in the bistable region, but then again rapidly evolves away. Note that the promoter functions that enter the bistable region after approximately 3000 days are very different from the initial promoter function. The most important difference is the location of the shift, which has gone to a much lower allolactose concentration. This causes the cells almost never to reach the bistable region, because such low external lactose concentrations are very seldom reached.

We can count how many promoter functions are bistable for different cAMP concentrations during all three evolutionary simulations. Not taking into account

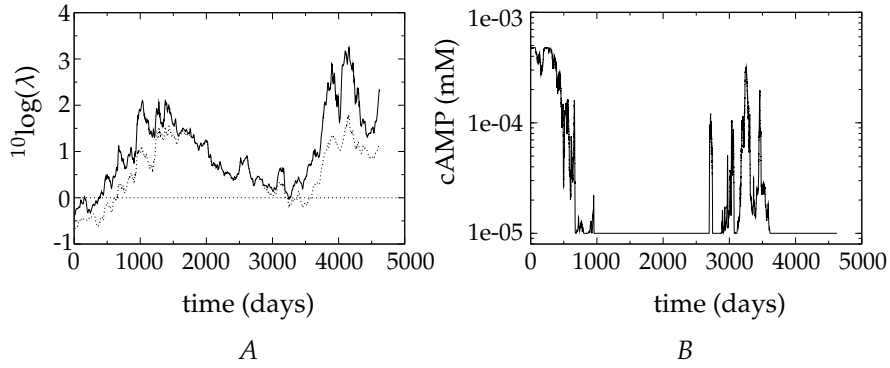


Figure 2.3: (*A*) Population average of the cusp parameter $\lambda(C_{max})$, i.e. low glucose, (*solid line*) and $\lambda(C_{min})$, i.e. high glucose (*dotted line*). (*B*) Population average of the cAMP concentration at which the cusp bifurcation occurs.

the initial bistable period, we find that over all three simulations, on average, 9% of the population is potentially bistable for high cAMP, low glucose and 24% for low cAMP, high glucose. In the simulation yielding the best competitor, these percentages are 1.5% and 9%, respectively.

In Fig. 2.4 we plot the promoter function of the last common ancestor of the evolutionary simulation of Fig. 2.3, as well as the experimentally observed promoter function (Setty *et al.*, 2003). Qualitatively, the two promoter functions look strikingly similar. The shifts are very shallow; both in the experiments as in the evolved promoter functions, an approximately 10-fold increase in inducer concentration is needed to fully induce the operon. Also the four different plateau transcription levels nicely compare with each other. Because in experiments only relative promoter activities are measured, and different experiments yield different quantitative results, a quantitative comparison is not possible.

In Fig. 2.5 we depict bifurcation diagrams corresponding to the evolved promoter function of Fig. 2.4 *A*, both for lactose and IPTG as inducer. For lactose, Fig. 2.5 *A*, no bistability is observed, as was already clear from Fig. 2.3. For IPTG, however, we see that the promoter function acts bistably for all cAMP concentrations. This is because the repression factor of the evolved operon is indeed high enough (approximately 150) to cause bistability for artificial inducers. Note that we scaled the IPTG concentration with respect to the evolved binding parameter of allolactose to LacI (k_A). IPTG probably has a different binding parameter, but since k_A does not enter the bifurcation condition, the switching behavior is not affected.

Recently the phase diagram of the wild-type lactose utilization network has been measured (Ozbudak *et al.*, 2004). This phase diagram shows for which external glucose and TMG concentrations the network is bistable, repressed, or induced. Fig. 2.6 shows this phase diagram for the evolved promoter function of Fig. 2.4 *A*. The locations of the fold bifurcations were measured in bifurcation dia-

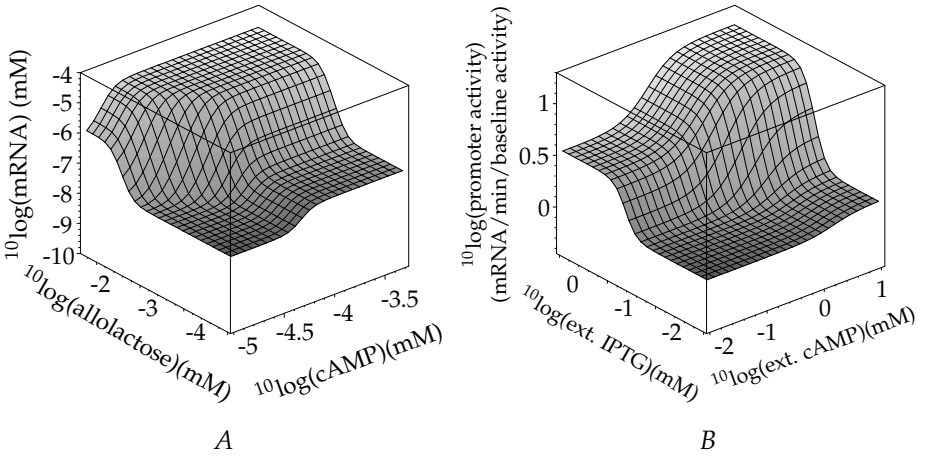


Figure 2.4: (*A*) The promoter function of the last common ancestor with cost of 3%. All evolved promoter functions are only plotted over the cAMP and allolactose concentration the bacterium experiences, which depend on the promoter function itself. (*B*) The experimentally measured promoter function (Setty *et al.*, 2003).

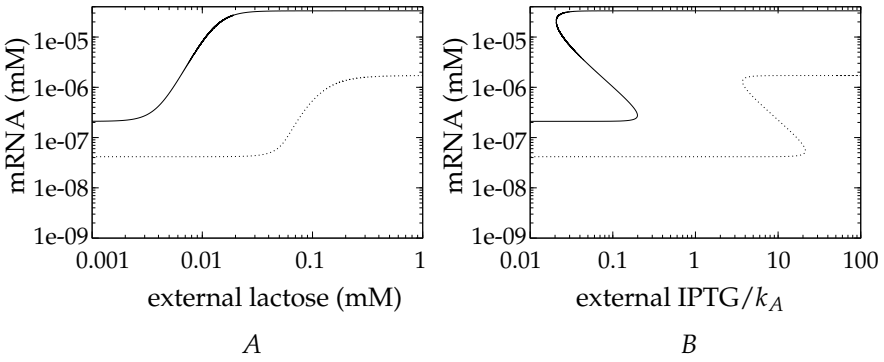


Figure 2.5: Two bifurcation diagrams of the promoter function of Fig. 2.4 *A*, using lactose (*A*) or IPTG (*B*) as inducer. The solid lines represent maximal cAMP concentration and the dotted lines minimal cAMP concentration. The same growth rates were used as in Fig. 2.2.

grams similar to Fig. 2.5 *B*, for different glucose concentrations. Instead of TMG we again used IPTG as artificial inducer.

Qualitatively as well as quantitatively this phase diagram is very similar to the experimentally observed phase diagram. As in Ozbudak *et al.* (2004), we observe that for high glucose concentrations, the operon becomes induced at higher IPTG concentrations, in a sigmoid way. The reason for this behavior is that when

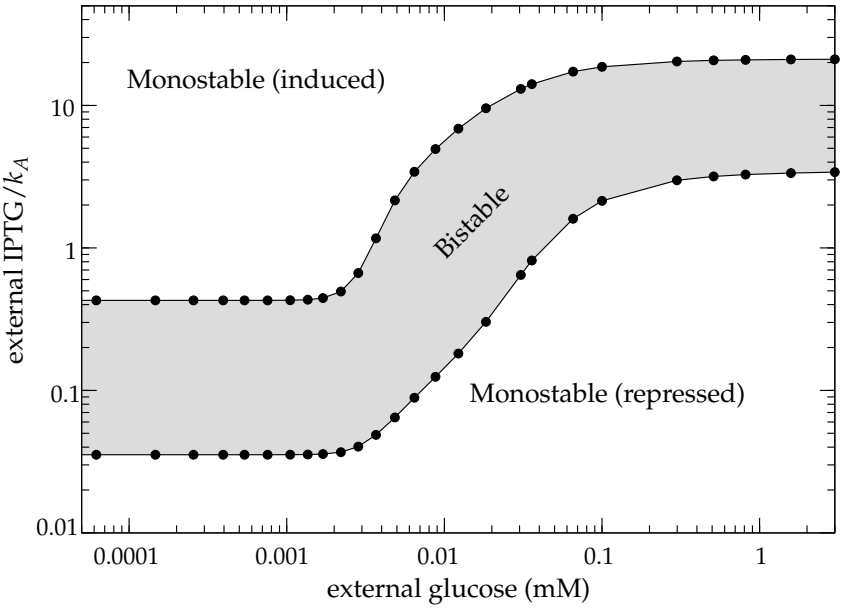


Figure 2.6: Phase diagram corresponding with the promoter function of Fig. 2.4 *A*. For this picture we used a growth rate of $\mu = 0.01/\text{min}$.

glucose is abundant, there is less need to induce the operon. Quantitatively, the location and the width of the shift from low to high glucose and the width of the bistable region of the evolved promoter function are very similar to that experimentally observed.

Because we did not observe bistability with respect to lactose, we tried to find a condition for which we expect bistability to evolve more easily. If the cost for *lac* operon activity would be higher, it becomes more important to repress activity and bistability might evolve more easily. Therefore we performed evolutionary simulations with unrealistically high cost (10 times higher, such that growth is lowered 30% at full activity).

High Cost For Promoter Activity

The evolutionary dynamics of the cusp parameter $\lambda(C)$ of the simulation is shown in Fig. 2.7 *A*. Again we see, for high-cAMP, low-glucose concentrations, that the cusp parameter increases and that the population evolves out of the bistable region. For low-cAMP, high-glucose concentrations, however, the opposite is the case. In Fig. 2.7 *B* the population average of the cAMP concentration at which the cusp takes place is shown.

From Fig. 2.7 *B* we can see that the cAMP concentration at which the cusp takes place decreases over time. Note that $C_{min} \approx 1 \times 10^{-5}$ mM. Therefore, only at very high glucose concentrations is bistability maintained. At high glucose

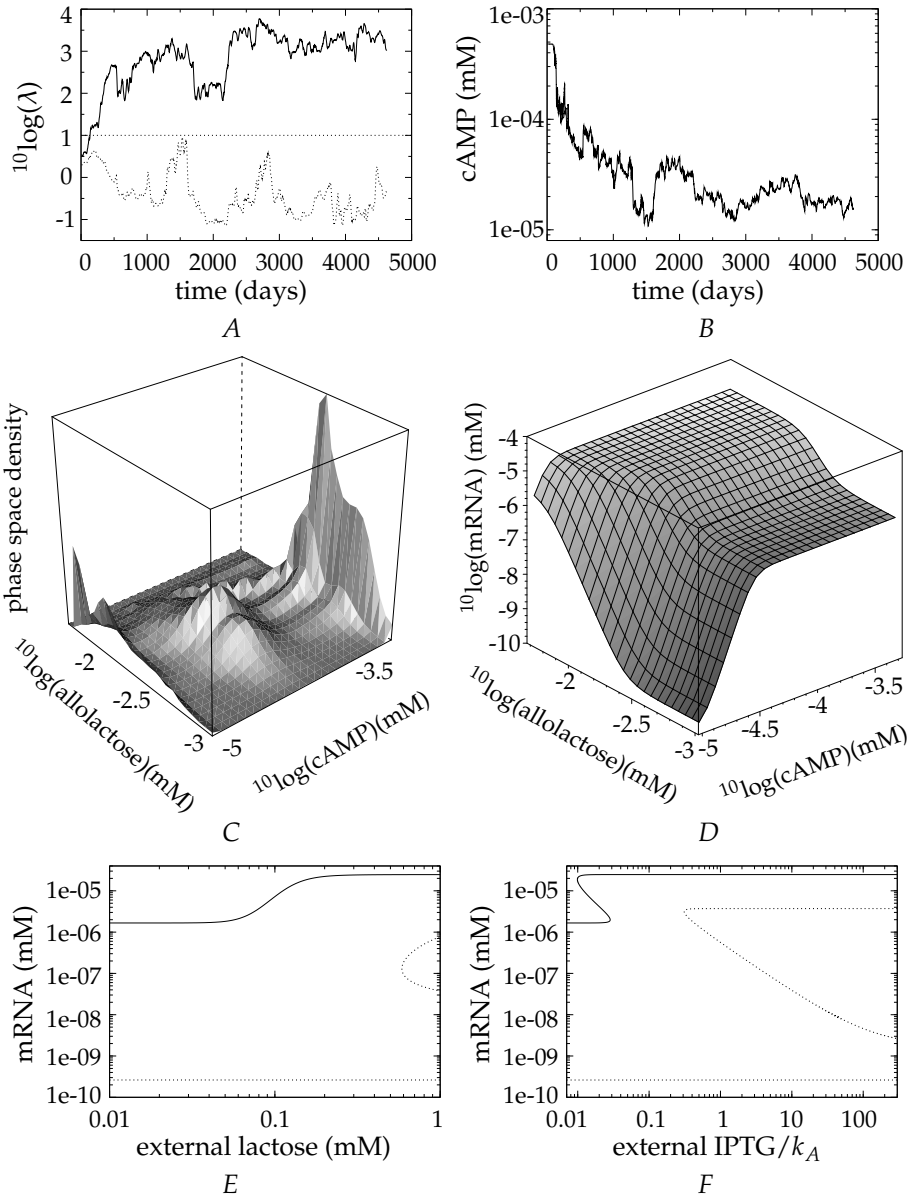


Figure 2.7: Results of evolution with high cost for promoter activity: (*A*) Population average of the cusp parameter $\lambda(C_{max})$ (*solid line*) and $\lambda(C_{min})$ (*dotted line*). (*B*) Population average of the cAMP concentration at which the cusp bifurcation occurs. (*C*) Histogram of the number of visits of the promoter function shown in Fig. 2.7 *D* in cAMP-allolactose space. (*D*) The promoter function of the last common ancestor. (*E*) The bifurcation diagram, corresponding to the promoter function in Fig. 2.7 *D*, when growing on lactose, for maximal-cAMP, low-glucose (*solid line*) and minimal-cAMP, high-glucose (*dotted line*). Again the same growth rates were used as in Fig. 2.2. (*F*) Bifurcation diagrams using IPTG as inducer.

concentrations, the bacteria use glucose before using lactose, and when the glucose concentration then decreases, bistability is lost again. In this way, bistability at high glucose could enforce sequential uptake of glucose and lactose.

However, the cells are too short a time in this part of the phase space to make this bistability functional. For functional bistability, the protein concentrations must be able to adapt to the transcription rate, which, due to the low protein degradation rate, takes hours. In the mean time, glucose is rapidly consumed by the cells and the glucose concentration drops below the glucose concentration required for bistability before the protein concentrations reach their equilibrium values. In Fig. 2.7 *C* we show how often cells are on a certain position in the phase space. We can see that the low-allolactose, low-cAMP corner is indeed rarely visited by the cells. Still, the low promoter activity in this region lowers cost and enlarges the delay in lactose uptake in the presence of glucose.

The promoter function of the last common ancestor of this simulation is shown in Fig. 2.7 *D*, and the corresponding bifurcation diagram for lactose and IPTG, respectively, in Fig. 2.7 *E* and Fig. 2.7 *F*. For artificial inducers, the evolved promoter function again behaves bistably for all glucose concentrations, whereas this only happens at very high cAMP levels when induced by lactose.

For high cost, on average 24% of the population is bistable for low glucose and 73% for high glucose. These averages are again calculated over all three evolutionary simulations. Indeed, for high cost, bistability is more likely, but is still most often avoided for low glucose concentrations.

Another important difference between the two costs is the location of the shift between low and high allolactose concentrations that evolves. For high cost, the operon becomes induced at approximately a 10-times higher allolactose (and hence external lactose) concentration (compare Fig. 2.4 *A* and Fig. 2.7 *D*). This is because when the cost for promoter activity is high, it does not pay to be active at relative low external lactose concentrations, whereas it does when the cost is lower.

Longer and Shorter Periods With and Without Lactose Influx

In different environments, different promoter functions might evolve. Therefore, we checked our results for different durations of periods with and without influx. Increasing the duration (with a factor two) of periods without lactose influx might increase the cost for having a high repressed transcription rate and therefore favor bistability. However, we again found similar results.

Decreasing the duration of the periods, both with and without glucose and lactose (by a factor four), leads to a decrease in regulation. Because there is too little time between periods of lactose influx, the cells choose to stay active at low lactose concentrations. This is seen in three independent simulations.

In one simulation, the population totally lost regulation with respect to lactose at low glucose concentrations. However, the last common ancestor of this simulation was the least competitive. The two other last common ancestors were competitively equal. One of these ancestors is shown in Fig. 2.8. At low-glucose, low-lactose, this ancestor decreases its activity only by approximately factor-two.

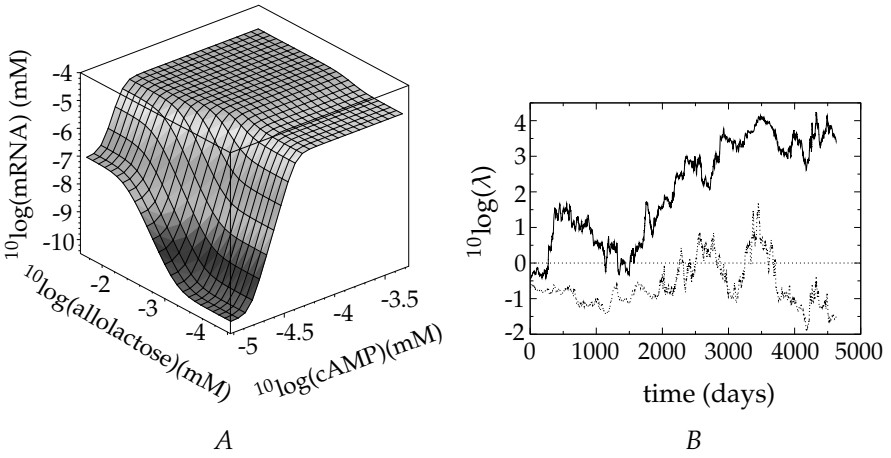


Figure 2.8: (*A*) The promoter function of one of the two most competitive last common ancestors. (*B*) Population average of the cusp parameter $\lambda(C_{max})$ (*solid line*) and $\lambda(C_{min})$ (*dotted line*).

At very high glucose concentrations however, it exhibits bistability, in the same way as the high cost promoter function.

The third promoter function has a very shallow slope relative to lactose: Only at low lactose concentrations appreciable regulation occurs. At very low lactose concentrations the promoter function is bistable, for all glucose concentrations. Such low lactose concentrations are, however, very seldom encountered, due to the frequent switching of the environment, and again the bistability is not functional. If the occurrence of these very low lactose concentrations is enforced by removal of external lactose, the promoter functions evolved away from bistability, again in three independent simulations.

2.3.3 Why to Avoid Bistability?

In Fig. 2.9 we plot the shifting behavior, from a lactose-poor to a lactose-rich environment, of seven different promoter functions, differing in only one parameter (γ, see Supplementary Material). This parameter, which gives the leakiness of the promoter function, changes the transcription rate at zero allolactose, ($PA(0, C)$). The external lactose, mRNA, and permease concentration are shown.

When influx starts after 10 min, we see that the highest four promoters immediately become induced, whereas for the lowest three it takes a while before they become induced. The lower three promoter functions are bistable, and the external lactose concentration has to increase over a certain threshold before they become induced. But we also see for continuous switches that the higher $PA(0, C)$, the faster the switch.

The doubling-time of *E. coli* is approximately 1 h, of the same order of the

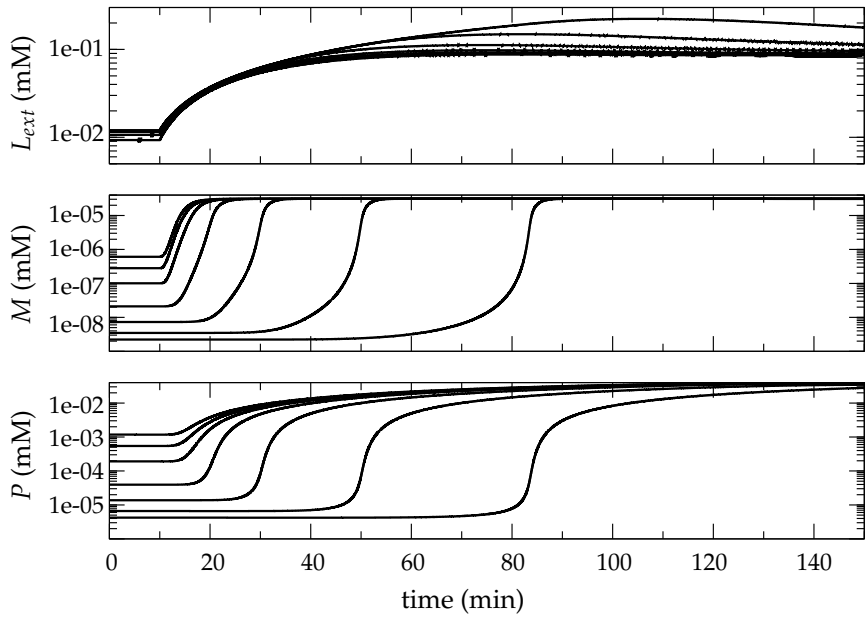


Figure 2.9: Dynamics of L_{ext}, M and P during a shift from a lactose-poor to a lactose-rich environment. The different lines correspond to different promoter functions, which are only different in one parameter, which determines $PA(0, C)$.

delay in lactose uptake. For wild-type *E. coli* it also takes approximately 1 h after induction, before the protein concentrations are maximal. Therefore, decreasing its delay can give a cell a very significant growth advantage. This explains why the population evolves out of the bistable region.

Thattai & Van Oudenaarden (2004) explain advantageousness of bistability using a model with instantaneous switching between equilibria. We now see that the crucial factor determining the (dis)advantageousness of bistability is the fact that switching between the two equilibria is not instantaneous, because of the slow protein dynamics. Because the difference between induced and repressed transcription levels is higher for bistable promoters, switching is slower for bistable promoters than for continuous promoters.

2.4 Discussion

The response of the *lac* operon to artificial inducers has had much attention, but from an evolutionary point of view, only the behavior with respect to lactose is important. Lacking information about the behavior of the operon with respect to lactose, the results for artificial inducers have been extrapolated to lactose.

However, only recently the response of the *lac* operon to lactose has been investigated (Setty *et al.*, 2003; Ozbudak *et al.*, 2004) and no bistability has been

demonstrated (Ozbudak *et al.*, 2004). From our results, this is only to be expected. The analytical expression we derived simultaneously describes the switching behavior of the operon with respect to artificial inducers and to lactose. For artificial inducers we indeed find that the promoter function is in the bistable region. Due to the fact that lactose is metabolized by β-galactosidase, it is much less likely that the operon is bistable with respect to lactose. Direct observation whether the *lac* operon is bistable with respect to lactose is still difficult (Ozbudak *et al.*, 2004). Our analytical result can predict whether or not the operon is bistable with respect to lactose. For this prediction we would need to know how large the operon-independent lactose decay and transport is, and a reliable estimate for the protein concentrations for the repressed operon.

It has already been claimed that bistability with respect to lactose would not be possible (Savageau, 2001). In that model, however, no operon-independent degradation is taken into account. In our model, this is described by the limit $\lim_{\zeta\to\infty}\lambda(C) = \infty$, and therefore, in that limit, bistability can indeed never be found. We find, like Yildirim & Mackey (2003), however, that due to growth or possible mechanisms to degrade internal lactose (γ_L), bistability for lactose is also possible, albeit much more difficult than for artificial inducers.

The evolved promoter function for realistic operon cost strikingly resembles the experimentally observed promoter function (Setty *et al.*, 2003), especially in the simulation yielding the best competitor. High transcription rates at zero allolactose and shallow shifts were experimentally observed, and this is also what we find. Both these properties help to avoid bistability when growing on lactose, which is a strong indication that the operon indeed evolved to avoid bistability. Furthermore, the phase diagram of the lac operon, when growing on an artificial inducer, is almost identical to the experimentally observed phase diagram (Ozbudak *et al.*, 2004).

In different environments or with different parameters, different promoter functions evolve. Interestingly, which promoter function evolves again shapes the environment the cells experience. At high cost, for example, the promoter function evolves such that cells never experience lactose concentrations as low as those evolved at low cost, because the cells evolved at high cost do not deplete lactose to such low levels.

Because cells never experience extremely low lactose levels, bistability at very low lactose concentrations is nonfunctional. We performed simulations in which we enforced very low lactose concentrations by adding an external decay for glucose and lactose, such that bistability is more often functional. Indeed, a population with nonfunctional bistability at these low lactose levels evolves out of the bistable region when lactose decays externally.

We have opted to study evolution of the *lac* operon in a spatially explicit model, because previous work (Pfeiffer *et al.*, 2001; Pagie & Hogeweg, 2000) has shown that a spatial context favors the evolution of regulation, in contrast to evolutionary adaptation (compare Dekel & Alon (2005)). In particular, space appears to be crucial when regulatory states last over many generations (Pagie & Hogeweg, 2000), as is the case for the lac operon.

Furthermore, a spatial model ensures that the environment of the cells

changes over multiple time scales. The glucose and lactose influx period define a long time scale. Due to cell division and movement, cells also experience changes in the environment over a much shorter time scale. In this way, space makes the environment of the cells inherently noisy. Without space, cells "know" that if the lactose concentration starts increasing, it will keep increasing for a long time. Bistability can in theory function as a noise-filter of a system, which makes space an important factor to consider.

The search space in the evolutionary simulations is very high-dimensional and redundant. We found that minimizing the dimensionality of the search space, as is done in the mathematical analysis, by evolving the V-parameters instead of the biological parameters, decreases the search efficiency. This is in agreement with the known role of neutrality in evolution (Huynen *et al.*, 1996).

Our evolutionary simulations clearly point out that bistability is disadvantageous for the cells, due to the increase in delay in lactose uptake it causes, especially in the absence of glucose. Even with unrealistically high cost, bistability is only observed for very high glucose concentrations. This bistability is nonfunctional as we explained above.

We used a deterministic model to describe the intracellular dynamics of each cell. It is known, however (Kierzek *et al.*, 2001), that protein numbers in the repressed *lac* operon are low. Therefore, stochasticity might play an important role in regulation of the *lac* operon.

A bistable cell can, due to stochasticity, switch between both equilibria. In this way the population can become heterogeneous, a phenomenon called bet-hedging, which could be beneficial, because a fraction of the population will always be in the best state.

Recently, there has been a lot of interest in the circumstances under which bet-hedging would be beneficial. It has been shown that bet-hedging strategies can be beneficial when environmental sensors are imperfect or when the cost for sensing the environment is high enough (Wolf *et al.*, 2005; Kussell & Leibler, 2005). Without any sensor imperfection or cost, bet-hedging can still be beneficial in periodic environments, but not in a stochastic environment (Thattai & Van Oudenaarden, 2004).

However, all these studies assume instantaneous switching between intracellular states. We now show that the crucial factor determining the disadvantage of bistability is the increase in delay in switching between the equilibria. Because the repressed transcription rate must be very low for bistability to occur, it takes much time to switch between both equilibria, even when stochasticity allows switching in the bistable region. These crucial considerations are neglected in these previous models.

Our study indicates that a priori we should expect the *lac* operon not to be bistable, because of the mentioned disadvantages. Whether bet-hedging due to stochasticity can overcome these disadvantages remains to be seen. Preliminary simulations, in which stochasticity is incorporated, suggest that bistability is avoided in the same way as described in this article. As stochasticity is a very important issue in this context, we will further investigate the effects of stochasticity in our model.

In any case, our results suggest that no advantages for bistability need to be sought. From our mathematical analysis it is clear that bistability is much more difficult for lactose than for artificial inducers, and that the wild-type promoter probably is not bistable with respect to lactose. Furthermore, our evolutionary results show that bistability is disadvantageous and therefore avoided during evolution. Finally, experimentally, bistability has not been observed for lactose, although it has been for artificial inducers (Ozbudak *et al.*, 2004).

In evaluating the promoter function, we should be aware that there is no such thing as "the" promoter function of *E. coli*. It has been shown that a population of *E. coli* cells can adapt to new environments by changing the operon activity in only a few hundred generations (Dekel & Alon, 2005), and that prolonged absence of lactose can even destroy regulation altogether. In our model, prolonged absence of lactose can push the promoter function into the bistable region. Our results suggest that these so-evolved individuals will lose out when lactose becomes an essential metabolite once again. All in all, we conclude that there is now ample evidence that bistability in the *lac* operon of *E. coli* is an artifact of using artificial inducers and it has not evolved for lactose.

Acknowledgments

We thank Athanasius Marée for helpful discussion.

2.5 Supplementary Material

We will give here an entire description of the model we used to study the evolution of the *lac* operon in space.

2.5.1 Intracellular Dynamics

An overview of all cellular processes that are taken into account is given in Fig. 2.10.

Gene Regulation

The activity of the *lac* operon is determined by an activator protein, cAMP receptor protein (CRP) and a repressor protein, LacI. CRP is activated by cAMP whereas LacI is repressed by allolactose. Allolactose is produced from lactose by β-galactosidase.

Setty *et al.* (2003) derived an equation for the transcription rate of the *lac* operon. They find the following equation

$$PA(A, C) = V_1 \frac{1 + V_2\mathcal{A} + V_3\mathcal{R}}{1 + V_4\mathcal{A} + V_5\mathcal{R}}, \tag{2.27}$$

where A stands for the allolactose concentration and C for the cAMP concentration and $\mathcal{A}$ and $\mathcal{R}$ are the fraction of active CRP and repressed LacI, respectively.

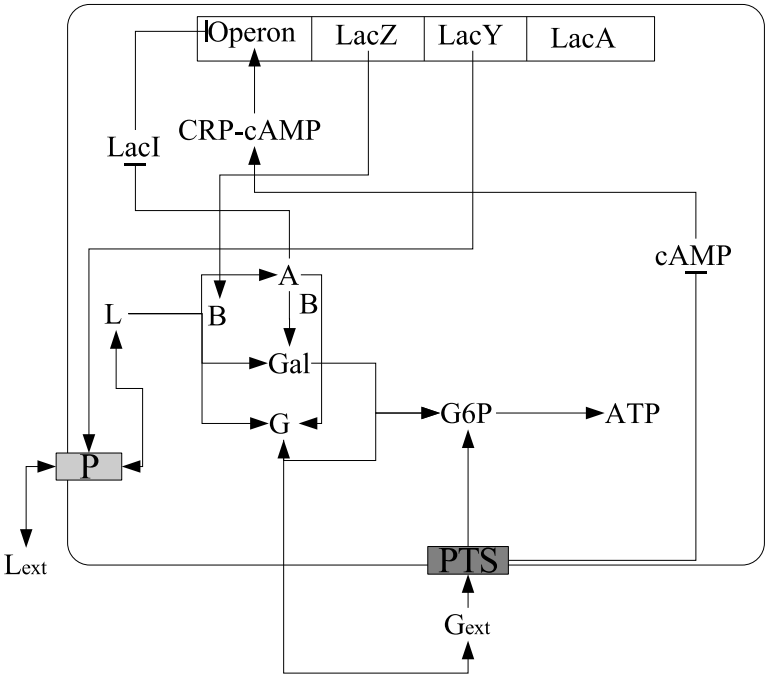


Figure 2.10: An overview of our model of the *lac* operon.

$$\mathcal{A} = \frac{(C/k_C)^n}{1 + (C/k_C)^n} \tag{2.28}$$

$$\mathcal{R} = \frac{1}{1 + (A/k_A)^m}, \tag{2.29}$$

where n and m are the Hill-coefficients of cAMP binding to CRP and allolactose binding to LacI. k_C and k_A are the dissociation constants for these reactions. Furthermore we have defined

$$\begin{aligned} V_1 &= (a\alpha + \gamma)/(1 + a) \\ V_2 &= d(b\beta + \gamma)/(a\alpha + \gamma) \\ V_3 &= \gamma c/(a\alpha + \gamma) \\ V_4 &= d(b + 1)/(a + 1) \\ V_5 &= c/(a + 1) \end{aligned} \tag{2.30}$$

and
$a = RNAP/k_{RNAP}$, RNA-polymerase in units of its dissociation constant for

binding to a free site.
$b = RNAP/k_{RNACP}$, RNA-polymerase in units of its dissociation constant for binding to a site with bound CRP.
$c = LACI_T/k_{LACI}$, the total LacI concentration in units of its dissociation constant for binding to its site.
$d = CRP_T/k_{CRP}$, the total CRP concentration in units of its dissociation constant for binding to its site.
α, the transcription rate when RNA Polymerase is bound to the DNA, but CRP and LacI are not.
β, the transcription rate when both RNA Polymerase and CRP are bound, but LacI is not bound to the DNA.
γ, the "leakiness", the transcription rate when RNA Polymerase is not bound to the DNA.

This function, with the right parameters, accurately describes the measurements done by Setty *et al.* (2003). We model the mRNA dynamics using this equation.

$$\frac{dM}{dt} = PA(A, C) - (\gamma_M + \mu)M \tag{2.31}$$

We assume first order degradation, as we do for all chemicals. γ_M is the mRNA degradation rate and μ is the growth rate of the cell. This formula is different from the formula used by Wong *et al.* (1997). We use it because it is a flexible and general formula, which can describe a large class of regulatory functions.

Translation

Translation is modeled in precisely the same way as Wong *et al.* (1997). For the β-galactosidase activity we have

$$\frac{dB}{dt} = k_B M - (\gamma_B + \mu)B \tag{2.32}$$

and for the permease activity

$$\frac{dP}{dt} = k_P M - (\gamma_P + \mu)P. \tag{2.33}$$

Lactose Uptake and Metabolism

Regulation of lactose uptake by glucose is mediated by the phosphoenolpyruvate: carbohydrate phosphotransferase system (PTS). The PTS is reviewed by Postma *et al.* (1993) Glucose uptake causes dephosphorylation of enzyme IIA^{glc} via the PTS. This enzyme is involved in different regulatory mechanisms. Phosphorylated IIA^{glc} activates adenylate cyclase and therefore the production of cAMP, this is called catabolite repression. Dephosphorylated IIA^{glc} is believed to inhibit lactose uptake by binding the lactose permease. This process is called inducer exclusion. We did not take inducer exclusion into account in this model, as is explained in the main text. For lactose transport we now have

$$V_{t,Lac} = P(\frac{k_{Lac,in}L_{ext}}{K_{Lac,in} + L_{ext}} - \frac{k_{Lac,out}L}{K_{Lac,out} + L}). \tag{2.34}$$

The influx of lactose is assumed to be proportional to the permease concentration, like Wong *et al.* (1997) assume. This automatically means that the lactose influx is proportional to the volume of the cell instead of the surface area. We therefore assume that the area of the cell is large enough, so that all membrane permease can be situated in the cell membrane.

After transport into the cell lactose is degraded by β-galactosidase into either allolactose or glucose and galactose. We assume, like Wong *et al.* (1997) and Yildirim & Mackey (2003) a constant ratio between these two pathways. Huber *et al.* (1976) found that 46% of the internal lactose is transformed to allolactose, while 54% is hydrolysed. This then gives

$$V_{Lac-Allo} = k_{Lac-Allo}B\frac{L}{L + K_{m,Lac}} \tag{2.35}$$

and

$$V_{cat,Lac} = k_{cat,Lac}B\frac{L}{L + K_{m,Lac}}. \tag{2.36}$$

We also take operon independent lactose and allolactose degradation into account, in order to have the possibility of bistability also when $\mu = 0$. The equation for lactose dynamics now becomes

$$\frac{dL}{dt} = V_{t,Lac} - V_{Lac-Allo} - V_{cat,Lac} - (\gamma_L + \mu)L. \tag{2.37}$$

The allolactose dynamics is governed by two processes, production from lactose and hydrolyzation by β-galactosidase. The hydrolyzation rate of allolactose is given by

$$V_{cat,Allo} = k_{cat,Allo}B\frac{A}{A + K_{m,Allo}}. \tag{2.38}$$

We neglect the binding of allolactose to the repressor in the dynamics of allolactose, so we get for the allolactose dynamics

$$\frac{dA}{dt} = V_{Lac-Allo} - V_{cat,Allo} - (\gamma_A + \mu)A. \tag{2.39}$$

Wong *et al.* (1997) assume that the glucose, which is formed, can diffuse out of the cell or can be phosphorylated by hexokinase. They test both possibilities separately. Interestingly, Hogema *et al.* (1999) observed both phenomena. Mutant strains with very high lactose uptake rate excrete glucose, whereas the wild type strain does not. The internal glucose concentration in the mutant strains was observed to be much higher than in the wild type strain. This observation suggests that the hexokinase activity is saturated and if the hexokinase activity is

saturated, the internal glucose is excreted into the medium. We now model the glucose degradation by

$$V_{cat,Glu} = k_{cat,Glu} \frac{G}{G + K_{m,Glu}} \tag{2.40}$$

and glucose excretion by

$$V_{t,Glu} = k_{Glu,out}(G - G_{ext}), \tag{2.41}$$

where we assume that this process is reversible. Note that this passive glucose influx is very small compared to the active glucose influx via the PTS. The internal glucose dynamics now becomes

$$\frac{dG}{dt} = V_{cat,Lac} + V_{cat,Allo} - V_{cat,Glu} - V_{t,Glu} - \mu G. \tag{2.42}$$

Glucose Metabolism

Glucose metabolism is also incorporated in the model. Glucose is transported into the cell and phosphorylated at the same time by the PTS. We again assume, like Wong *et al.* (1997) that the number of PTS complexes in the cell membrane is proportional to the biomass and we find

$$V_{t,G6P} = k_{t,Glu} \frac{G_{ext}}{G_{ext} + K_{t,Glu}}. \tag{2.43}$$

The galactose formed by the hydrolysis of lactose and allolactose is converted to glucose-6-phosphate via the Leloir pathway. We assume, like Wong *et al.* (1997), that this pathway is instantaneous. The glucose-6-phosphate is metabolized via the glycolysis and possibly respiration. When *E. coli* has electron acceptors like oxygen, nitrate or fumarate at its disposal it chooses to respire, when there are no electron acceptors available, it ferments. The environment of *E. coli* in the gut is anaerobic. But near the epithelial cells aerobic respiration may be important (Clarke, 1977). When oxygen is absent nitrate is the most important electron acceptor (Cole, 1996). The ATP yield of nitrate reduction is unknown, but lies somewhere between the ATP yield of aerobic respiration and fermentation (Gennis & Stewart, 1996). Andersen & Von Meyenburg (1980) observed that growth rates of *E. coli* are limited by respiration. When the glucose uptake rate is maximal, the respiratory pathway is saturated and excess glucose (approximately 24%) is fermented. Therefore we modeled respiration as a quickly saturating function and fermentation as a slowly saturating function with a much higher plateau. We tuned our parameters such that also in our model, when the glucose uptake is maximal, approximately 24% of the glucose is fermented. Respiration and fermentation are now modeled as

$$V_{G6P,Rsp} = k_{G6P,Rsp} \frac{G6P}{G6P + K_{G6P,Rsp}} \tag{2.44}$$

and

$$V_{G6P,Frm} = k_{G6P,Frm} \frac{G6P^8}{K_{G6P,Frm}^8 + G6P^8}. \tag{2.45}$$

This all gives us

$$\begin{aligned} \frac{dG6P}{dt} &= V_{t,G6P} + V_{cat,Glu} + V_{cat,Lac} + V_{cat,Allo} - V_{G6P,Rsp} - \\ & V_{G6P,Frm} - \mu G6P. \end{aligned} \tag{2.46}$$

cAMP Dynamics

As mentioned above, cAMP production is thought to be dependent on dephosphorylation of enzyme IIA^{glc}. Dephosphorylation of IIA^{glc} by glucose is dependent on the glucose transport and therefore we model cAMP as

$$\frac{dC}{dt} = k_{syn,cAMP} \frac{K_{syn,cAMP}}{V_{t,G6P} + K_{syn,cAMP}} - (\gamma_{cAMP} + \mu)C. \tag{2.47}$$

Like for allolactose we neglect binding of cAMP to CRP in the dynamics of cAMP.

Energy and Growth

Growth of *E. coli* is dependent on the respiratory and fermentative fluxes . But also on energy consuming actions. We can vary the energy yield of respiration, Y_{Rsp}, because as mentioned this is not known for anaerobic respiration. We simply choose the energy yield of aerobic respiration, but this parameter is not crucial for the outcome. There are several energy costs taken into account. First, Carlson & Srienc (2004) measured costs for basal metabolism, BMC, for *E. coli*. They also determined the amount of glucose, which is used for biomass and the growth cost, GC, for *E. coli*. In our model we interpret biomass production as growth cost, because the amount of biomass needed also grows linearly with the growth rate. Furthermore, transformation of galactose into glucose costs 1 ATP. The costs for protein production, PC, are taken to be proportional to the mRNA production rate. This gives us

$$\begin{aligned} \frac{dATP}{dt} &= Y_{Rsp} V_{G6P,Rsp} + 2V_{G6P,Frm} - BMC - GC \times \mu - \\ & PC \times PA(A, C) - V_{cat,Lac} - V_{cat,Allo}. \end{aligned} \tag{2.48}$$

We do not claim that the values of ATP concentrations are realistic. We only want a measure for the amount of energy of the cell and we use the amount of ATP for this. We assume a relation between the "ATP-concentration" of the cell and the growth rate

$$\mu = \mu_{max} \frac{ATP^4}{ATP^4 + K_{ATP}^4}. \tag{2.49}$$

This equation states that cells will not invest in cell growth if the energy of the cell is low. The energy level that is needed to start cell growth depends on K_{ATP}. The growth of the bacterium now becomes

$$\frac{dX}{dt} = \mu X. \tag{2.50}$$

2.5.2 Extracellular Dynamics

The above differential equations describe how the cells, with a given promoter function behave in a certain environment of glucose and lactose. We will now describe the spatial and population aspects of the evolutionary model in detail. Simulations are performed using a square grid of 25×25 points. We integrate the intracellular differential equations using a time step of 0.2 seconds.

Glucose and Lactose Influx

Glucose and lactose influx are modeled in periods with and periods without influx. These periods for glucose and lactose are independent. The periods of influx are modeled such that the total influx of carbon is on average equal for each period. This means that short periods have high fluxes, while long periods have low fluxes and that lactose fluxes are twice as low as glucose fluxes, because lactose is a disaccharide and glucose a monosaccharide.

On average the influx periods have a length of 11 hours, with a standard deviation of 2.8 hours. These influx lengths are randomly chosen from a normal distribution. The total amount of influx during one period equals on average 1.66 mM and 0.83 mM, with a standard deviation of 0.5 and 0.25 for glucose and lactose respectively. All grid points have the same amount of influx, so influx is homogeneous over space.

The periods without influx for glucose and lactose are of equal length, also on average 11 hours, but picked randomly from a uniform distribution between 0 and 22 hours, with a standard deviation of 9 hours.

In 10% of the cases, the duration of a period with influx equals one time-step (0.2 seconds), such that there is a large instantaneous increase in food concentration of on average 1.66 mM for glucose or 0.83 mM for lactose. Therefore, the total amount of influx is still the same as during other periods, but instead all food is given at once. In 25% of these cases, there is an instantaneous influx of glucose AND lactose. This is done to make sure that high concentrations of glucose, lactose and glucose AND lactose and lactose are experienced once in a while.

The period of 11 hours is chosen arbitrarily, but such that in some instances the cell can and in some instances cannot adapt to this new environment. As described in the main text, also runs with different period lengths are used.

Extracellular Concentrations

Besides influx, the external glucose and lactose concentrations at each site decreases due to consumption.

$$\frac{dG_{ext}}{dt} = -QX(V_{t,Glu} + V_{t,G6P}) \tag{2.51}$$

$$\frac{dL_{ext}}{dt} = -QXV_{t,Lac} \tag{2.52}$$

In these equations, Q is a fixed constant, which represents the fraction of intracellular volume and extracellular volume at a site. Hence a decrease in extracellular concentration of 1 μM corresponds with an increase in intracellular concentration of 2.9 mM. Furthermore, extracellular glucose and lactose diffuse over the grid with diffusion constant $2D$ and D respectively, because glucose is a two times smaller molecule than lactose. Diffusion is the most important source of spatial heterogeneity in the model.

Reproduction, Cell Loss and Movement

The initial size of a cell is, in an arbitrary unit, 1. The cell reproduces if its size exceeds 2 and there is a neighboring empty grid point. If there is not, the cell continues to grow, until its size exceeds 3.5. If this happens, the cell goes into a rest-state until there is space to reproduce. When a cell reproduces, its size is divided by two, while all concentrations remain equal. There are two ways for the cells to disappear. The first is density dependent cell loss. The cells have a probability of $0.001 + 0.01\rho$ per minute to "die", where ρ is the fraction of occupied neighboring grid points. The second way is when the energy of a cell, as given in Eq. 2.48, drops below zero. Finally, there is some cell movement incorporated in the model, by a random walk of the cells. Every cell has a probability of $0.005(1 - \rho)$ per minute to move to a random neighboring empty grid cell.

Evolution

Because we want to study the evolution of the *lac* operon, we only allow parameters that determine the shape of the promoter function to mutate. All other parameters, which describe the intracellular dynamics are fixed, because we want to know how the promoter function adapts, given the intracellular dynamics. We evolve the biological relevant parameters a, b, c, d, α, β, γ, k_{cAMP}, k_A, n and m.

When a cell reproduces, it has a possibility to mutate. The mutation rate for each parameter equals 0.01 per reproduction. In order to change a promoter function significantly, the biological parameters have to change orders of magnitude. Therefore we found it more suitable to mutate the biological parameters multiplicatively in stead of additively. Only n and m, which we did not want to vary over several orders of magnitude, are mutated additively.

Mutation steps were chosen from the following distribution: $10^{normal(0,\Delta)}$ and from a normal distribution for n and m. In order to prevent unrealistically high mRNA production rates, we impose a maximal mRNA production rate $V_{mRNA,max} = 2.2 \times 10^{-5}$ mM/min, such that the maximal lactose influx has realistic values. We can compare this value, with the values as mentioned by Malan *et al.* (1984). They observe an *in vitro* transcription initiation rate of approximately

0.1 RNA chain/minute. They mention however that *in vivo*, values ranging from 1 to 10 RNA/chains per minute are observed. Using a volume of 8×10^{-16} liters, our maximal transcription rate corresponds to approximately 11 molecules/cell, which is in the right order of magnitude.

We start evolution with a monomorphic population of cells with a bistable switch. The initial values of the evolvable parameters are listed in Table 1. We start with a population size of approximately 60 cells, randomly distributed over the grid (every site has a probability of 10% to be occupied by a cell). We start with a period of glucose and lactose influx, of unequal length. Initially, the glucose and lactose concentrations are zero.

Parameters

Because our model is based on detailed models of the dynamics of the *lac* operon, there are many parameters. Most parameter values we use are taken from literature, some are assumed. We noted that there is quite some variation between reported parameter values in literature (Wong *et al.*, 1997; Yildirim & Mackey, 2003). In most cases we choose the value that was also chosen by Wong *et al.* (1997). A full list of all model parameters can be found in Table 2.1.

Table 2.1: All model parameters with their values.

parameter	equation	value	comments
k_C	Eq. 2.28	evolvable, mM	initial value: 1.0×10^{-3} mM
n	Eq. 2.28	evolvable	initial value: 4.0
k_A	Eq. 2.29	evolvable, mM	initial value: 5.5×10^{-4} mM
m	Eq. 2.29	evolvable	initial value: 8.0
a	Eq. 2.30	evolvable	initial value: 1.0
b	Eq. 2.30	evolvable	initial value: 1.0
c	Eq. 2.30	evolvable	initial value: 1.0×10^6
d	Eq. 2.30	evolvable	initial value: 50
α	Eq. 2.30	evolvable, mM/min	initial value: 1.1×10^{-7} mM/min
β	Eq. 2.30	evolvable, mM/min	initial value: 2.2×10^{-5} mM/min
γ	Eq. 2.30	evolvable, mM/min	initial value 1.1×10^{-9} mM/min
γ_M	Eq. 2.31	0.693/min	Wong *et al.* (1997)
k_B	Eq. 2.32	9.4 mM enzyme/ (mM mRNA min)	Wong *et al.* (1997)
γ_B	Eq. 2.32	0.01/min	Wong *et al.* (1997)
k_P	Eq. 2.33	18.8 mM enzyme/ (mM mRNA min)	Wong *et al.* (1997)
γ_P	Eq. 2.33	0.01/min	Wong *et al.* (1997)
$k_{Lac,in}$	Eq. 2.34	2148 mmol lactose/ (mmol permease min)	Wong *et al.* (1997)
$K_{Lac,in}$	Eq. 2.34	0.26 mM	Wong *et al.* (1997)
$k_{Lac,out}$	Eq. 2.34	2148 mmol lactose/ (mmol permease min)	Wong *et al.* (1997)
$K_{Lac,out}$	Eq. 2.34	0.26 mM	unlike Wong *et al.* (1997), intracellular concentrations are in mM
$k_{Lac-Allo}$	Eq. 2.35	8460/min	Wong *et al.* (1997)
$K_{m,Lac}$	Eq. 2.35, Eq. 2.36	1.4 mM	Martinez-Bilbao *et al.* (1991), referred to by Wong *et al.* (1997)
$k_{cat,Lac}$	Eq. 2.36	9540/min	Wong *et al.* (1997)
γ_L	Eq. 2.37	0.15/min	assumed, to get a significant bistable region, compare Yildirim & Mackey (2003)

Table 2.1: continued.

parameter	equation	value	comments
$k_{cat,Allo}$	Eq. 2.38	18000/min	Wong *et al.* (1997)
$K_{m,Allo}$	Eq. 2.38	0.28 mM	Wong *et al.* (1997)
γ_A	Eq. 2.39	0.15/min	assumed, to get a significant bistable region, compare Yildirim & Mackey (2003)
$k_{cat,Glu}$	Eq. 2.40	11.5 mM/min	fitted with data of Hogema *et al.* (1999)
$K_{m,Glu}$	Eq. 2.40	0.45 mM	fitted with data of Hogema *et al.* (1999)
$k_{Glu,out}$	Eq. 2.41	0.093/min	fitted with data of Hogema *et al.* (1999)
$k_{t,Glu}$	Eq. 2.41	45 mM/min	Wong *et al.* (1997), Carlson & Srienc (2004)
$K_{t,Glu}$	Eq. 2.41	0.015 mM	Wong *et al.* (1997)
$k_{G6P,Rsp}$	Eq. 2.44	34 mM/min	assumed, saturated respiratory flux assumed for maximal glucose influx. Andersen & Von Meyenburg (1980)
$K_{G6P,Rsp}$	Eq. 2.44	0.5 mM	idem. Andersen & Von Meyenburg (1980)
$k_{G6P,Frm}$	Eq. 2.45	200 mM/min	assumed, maximal fermentative flux is much larger than maximal respiratory flux. Andersen & Von Meyenburg (1980)
$K_{G6P,Frm}$	Eq. 2.45	20 mM	assumed, fermentation saturates much slower than respiration. Andersen & Von Meyenburg (1980)
$k_{syn,cAMP}$	Eq. 2.47	0.001 mM/min	Wong *et al.* (1997)
$K_{syn,cAMP}$	Eq. 2.47	1.0 mM/min	assumed, to have a large range of possible cAMP concentrations.
γ_{cAMP}	Eq. 2.47	2.1/min	Wong *et al.* (1997)
Y_{RsP}	Eq. 2.48	32 mM ATP/mM glucose-6-phosphate	assumed equal to the ATP-yield of aerobic respiration.
BMC	Eq. 2.48	23.5 mM/min	Carlson & Srienc (2004)
GC	Eq. 2.48	7.28×10^5 mM	estimated with data of Carlson & Srienc (2004)

Table 2.1: continued.

parameter	equation	value	comments
PC	Eq. 2.48	2.36×10^6 mM ATP/ mM mRNA	calculated assuming 3% growth cost at maximal activity, Koch (1983). (for high cost a value ten times higher is used)
μ_{max}	Eq. 2.49	0.0233/min	Wong *et al.* (1997)
Q	Eq. 2.51, Eq. 2.52	0.00035	assumed
D		0.0020 $gridsize^2/min$	assumed, scalable
Δ_a		0.075	assumed
Δ_b		0.075	assumed
Δ_c		0.15	assumed
Δ_d		0.15	assumed
Δ_α		0.075	assumed
Δ_β		0.075	assumed
Δ_γ		0.075	assumed
Δ_{k_A}		0.15	assumed
Δ_{k_C}		0.05	assumed
Δ_n		0.5	assumed
Δ_m		0.5	assumed
$V_{mRNA,max}$		2.2×10^{-5} mM/min	assumed, to have a realistic maximal lactose uptake rate.

3

The Effect of Stochasticity on the *lac* Operon: An Evolutionary Perspective

M.J.A. van Hoek and P. Hogeweg
Theoretical Biology/Bioinformatics Group, Utrecht University
Padualaan 8, 3584 CH Utrecht, The Netherlands.

PLoS Comput Biol. **3(6)**: e111 (2007 June)

Abstract

The role of stochasticity on gene expression is widely discussed. Both potential advantages and disadvantages have been revealed. In some systems, noise in gene expression has been quantified, in among others the *lac* operon of *Escherichia coli*. Whether stochastic gene expression in this system is detrimental or beneficial for the cells is, however, still unclear. We are interested in the effects of stochasticity from an evolutionary point of view. We study this question in the *lac* operon, taking a computational approach: using a detailed, quantitative, spatial model, we evolve through a mutation-selection process the shape of the promoter function and therewith the effective amount of stochasticity. We find that noise values for lactose, the natural inducer, are much lower than for artificial, non-metabolizable inducers, because these artificial inducers experience a stronger positive feedback. In the evolved promoter functions, noise due to stochasticity in gene expression, when induced by lactose, only plays a very minor role in short-term physiological adaptation, because other sources of population heterogeneity dominate. Finally, promoter functions evolved in the stochastic model evolve to higher repressed transcription rates than those evolved in a deterministic version of the model. This causes these promoter functions to experience less stochasticity in gene expression. We show that a high repression rate and hence high stochasticity increases the delay in lactose uptake in a variable environment. We conclude that the *lac* operon evolved such that the impact of stochastic gene expression is minor in its natural environment, but happens to respond with much stronger stochasticity when confronted with artificial inducers. In this particular system, we have shown that stochasticity is detrimental. Moreover, we demonstrate that *in silico* evolution in a quantitative model, by mutating the parameters of interest, is a promising way to unravel the functional properties of biological systems.

3.1 Introduction

Noise in gene expression, i.e., the variation in gene expression in an isogenic population in a homogeneous environment, has drawn much attention in recent years. When two isogenic cells vary in gene expression, this can be due to variation in factors determining gene expression in these cells, such as transcription factors, the concentration of RNA polymerase, the cell cycle, etcetera, which is called extrinsic noise. When, however, all extrinsic noise is absent, gene expression between these cells would still be different, because gene expression is inherently stochastic, due to the low numbers of molecules involved. The latter is called intrinsic noise.

Indeed, it has been clearly shown that gene expression can be stochastic (Ozbudak *et al.*, 2002; Elowitz *et al.*, 2002; Raser & O'Shea, 2004). The implications of stochastic gene expression are, however, much less clear. Stochasticity has been proposed to be deleterious in some systems (Fraser *et al.*, 2004), while being advantageous (Thattai & Van Oudenaarden, 2004; Blake *et al.*, 2006) in others. However, there is very little known about the consequences of stochasticity on particular systems.

Maybe the best-known system for genetic regulation is the *lac* operon of *E. coli*. The *lac* operon codes for three genes, two of which have a function in lactose uptake and metabolism. It codes for a permease protein that transports lactose into the cell and β-galactosidase, which degrades lactose. Gene expression in the *lac* operon has been experimentally shown (Elowitz *et al.*, 2002; Cai *et al.*, 2006) to be stochastic. The *lac* operon is regulated via a positive feedback loop. The operon is induced by allolactose (which is formed by β-galactosidase, from lactose). Induction of the operon again leads to higher permease and β-galactosidase concentrations and hence to higher allolactose concentrations. This positive feedback loop can cause bistability, which means that for certain extracellular inducer concentrations, two stable equilibria exist for the operon, induced and repressed. In a bistable system, stochastic gene expression can cause switching between these equilibria and hence give rise to a bimodal population. Such a bimodal population can be advantageous for the population (Thattai & Van Oudenaarden, 2004), a phenomenon called bet-hedging.

Bistability for the *lac* operon has been demonstrated using thiomethyl β-D-galactoside (TMG) (Novick & Weiner, 1957). Recently, these experiments were repeated and bistability for TMG was confirmed (Ozbudak *et al.*, 2004). In this paper, bistability for isopropyl β-D-thiogalactopyranoside (IPTG) and the natural inducer lactose was also tested. Although it is known that IPTG also enters the cell independently from the operon, it still behaved bistably. For lactose, no bistability was found. In **chapter 2** we showed that this difference in behavior is because artificial inducers are not degraded by β-galactosidase, while lactose is. Therefore, the positive feedback loop is much stronger for these artificial inducers.

Furthermore, we showed that, using a deterministic evolutionary model of the *lac* operon in a fluctuating environment, cells adapt their promoter function such that the response with respect to lactose becomes continuous instead of

bistable (see **chapter 2**). This can be explained by the increase in delay in lactose uptake that bistability unavoidably causes. These *in silico* evolved promoter functions, however, still behaved bistably with respect to artificial inducers. Indeed, the *in silico* evolved promoter function resembled the experimentally measured promoter function. Here we study how noise in gene expression influences the adaptation of the *lac* operon promoter function, both on a physiological and an evolutionary time scale.

We use a computational approach to tackle this problem. We modified the previously developed deterministic model (see **chapter 2**) for the evolution of the *lac* operon to include stochasticity in gene expression. This model is spatially explicit and consists of cells that grow on glucose and lactose, divide, and die. The intracellular model consists of detailed differential equations describing *lac* operon transcription, translation, and metabolism, with parameters taken from literature. The cells evolve the parameters which determine the *lac* operon promoter function and in this way adapt to the (fluctuating) environment. Importantly, the cells can in this way also adapt to the constraints that are imposed by the fixation of the other parameters. In our view, this is a good way to cope with the inevitable parameter uncertainty. See Materials and Methods for a more detailed description of the model.

We added stochasticity in gene expression on the protein level. We assumed that protein production occurs in bursts, as is experimentally observed (Cai *et al.*, 2006). The amount of protein produced per burst (i.e., the burst size) was shown to be geometrically distributed with a mean of five proteins (Cai *et al.*, 2006). This observation suffices to make our deterministic model stochastic, as is explained in the Materials and Methods section. Protein degradation is modeled binomially. When a cell divides, the number of proteins is divided between the two cells in a binomial way. In this way we added stochasticity without introducing any unknown parameter.

By comparing the deterministic and stochastic models, we can directly observe the consequences of stochasticity on the *lac* operon, which is experimentally difficult, because the *lac* operon is inherently stochastic. We compared the amount of noise in gene expression in our model with experimentally observed values of noise for the *lac* operon (Elowitz *et al.*, 2002) and found that noise in gene expression in our model is comparable to the experimentally observed noise.

These noise values were measured using IPTG. IPTG is not degraded by β-galactosidase and therefore behaves very differently than the natural inducer, lactose. The positive feedback loop is much weaker for lactose than for artificial inducers such as IPTG. Accordingly, we find that noise values for lactose are much lower than for IPTG. Therefore, the effect of stochasticity on evolution of the *lac* operon might be lower than what would be expected from these experiments.

In experiments where stochasticity in gene expression is measured, isogenic populations in well-mixed systems are considered in order to exclude all other sources of population heterogeneity. When we want to investigate the importance of stochasticity in gene expression in natural circumstances, we should, however, also take these other sources of population heterogeneity into account.

Therefore, by using a spatially explicit model of cells that evolve their *lac* operon promoter function, spatial and genetic heterogeneity are automatically taken into account. We find that both genetic and spatial heterogeneity contribute more to population heterogeneity than stochasticity in gene expression.

To explore the effect of stochasticity on evolution of the *lac* operon, we compared the results of the evolutionary simulations between the stochastic and the deterministic models (see **chapter 2**). We found that in the stochastic simulations, cells evolve a higher repressed transcription rate than in the deterministic model. Therefore, the promoter functions that evolved in the deterministic model experience more stochasticity when placed in the stochastic model than the promoter functions that evolved in the stochastic model. We show that this causes a reduction in fitness compared with the promoter functions evolved in the stochastic model, due to an increase in delay in lactose uptake. We conclude that in the stochastic model, the promoter functions evolve to minimize stochasticity in gene expression.

Indeed, stochasticity, when growing on lactose, is relatively unimportant for the dynamics of these evolved promoter functions, except sometimes at high glucose, low lactose concentrations. The dynamics when growing on lactose can well be described using a deterministic model. When modeling the dynamics of induction with artificial inducers, stochasticity, however, is much more important, due to the stronger positive feedback, and should be incorporated.

3.2 Results

3.2.1 Effects of Stochasticity on Noise in Gene Expression

First we studied how stochasticity in gene expression influences noise in gene expression. The promoter function of the cell in large part determines the amount of stochasticity a cell experiences, because stochasticity is higher at low protein levels. Therefore, promoter functions with low repressed transcription rates experience more noise than promoter functions with high repressed transcription rates.

It has been shown that noise in gene expression can be split up into two orthogonal components, intrinsic and extrinsic noise, such that $\eta_{int}^2 + \eta_{ext}^2 = \eta_{tot}^2$ (Swain *et al.*, 2002). Here the total noise η_{tot} is defined as the standard deviation divided by the mean of the population. Extrinsic noise is all noise that would affect two identical, independent copies of one gene in a single cell in exactly the same way, and intrinsic noise is noise that causes differences in expression levels of identical copies in a single cell.

The amount of intrinsic and extrinsic noise of *lac*-repressible promoters in different *E. coli* strains has been measured (Elowitz *et al.*, 2002). This was done by placing two genes, coding for two fluorescent proteins, which are controlled by identical promoters, in an *E. coli* strain. The intrinsic and extrinsic noise can now be simultaneously measured, by measuring how the protein levels fluctuate.

We performed similar simulations to validate the stochastic model with these

experiments. We initiated a population of 100 induced or repressed cells (solid and dotted lines, respectively, in Fig. 3.1) in a homogeneous environment with a certain extracellular inducer concentration. As in the experiments, spatial and genetic heterogeneity are absent. We waited for 41 hours, more than enough time for cells to go to equilibrium in the deterministic model. Then we calculated the noise in the protein number in the population. Noise levels were obtained in the same way as in the experiments. We kept track of the activity of two independent, identical *lac*-repressible genes that do not have a function in lactose uptake.

Because we did not introduce any free parameter in the model when introducing stochasticity, noise levels only depend on the promoter function used. We used a promoter function that has the same repression rate as the one used in the experiments (≈110, see Fig. 3.1 *A*). As a comparison, for the wild-type lac promoter function, the repression rate has been reported to be 170 (Ozbudak *et al.*, 2004). Furthermore, we used a Hill-coefficient of 4.0 (Setty *et al.*, 2003). For the growth rate, we took 0.01/min in these simulations. The induced transcription rate of the operon in these simulations is equal to the maximal transcription rate we imposed during the evolutionary experiments.

The experiments were done using IPTG, an artificial inducer of the *lac* operon. In contrast to lactose, IPTG is not degraded by β-galactosidase. This changes the dynamics, hence the noise, of the system considerably. Our model describes operon dynamics with lactose as inducer. Using parameters for IPTG, as used in Cheng *et al.* (2001), and not allowing for degradation of IPTG by β-galactosidase, our model can also describe *lac* operon dynamics when induced by IPTG. It is known that IPTG also enters the cell in the absence of permease. This is not taken into account in our model describing lactose dynamics and hence we add a permease-independent influx term for IPTG (see Materials and Methods). The amount of permease independent influx we chose is such that the maximum in the total noise is similar to that in the experiments.

In Fig. 3.1 *A* the results for IPTG are shown, using a permease-independent influx, k_{IPTG}, of 1.35/min. The intrinsic noise found in our simulations is almost identical to the experimentally observed intrinsic noise. Intrinsic noise levels in the model only depend on the number of proteins. We found that the data can almost perfectly be fitted by the function $\eta_{int} = C * P^a$, with $a = -0.505$, and $C = 2.35$ with an R^2-value of 0.99996. At steady state the theoretical prediction (Thattai & Van Oudenaarden, 2001) is $a = -0.5$ and $C = \sqrt{\frac{b}{1+\zeta} + 1} \approx 2.41$, where b equals the average burst size and ζ is the ratio of mRNA to protein lifetimes. This figure also confirms that the induced transcription rate we used has the right order of magnitude. It has been reported that the maximal *in vitro* transcription rate is approximately 0.18/min, but that the maximal *in vivo* transcription rate is approximately 1-10/min (Malan *et al.*, 1984). In contrast to Santillan *et al.* (2007), who used the maximal *in vitro* transcription rate, we used a maximal transcription rate of 11/min.

Qualitatively, the extrinsic noise corresponds with the experimentally observed noise. The maximum in the extrinsic noise is caused by the positive feedback loop in the *lac* operon. Fluctuations due to the intrinsic noise are amplified by the pos-

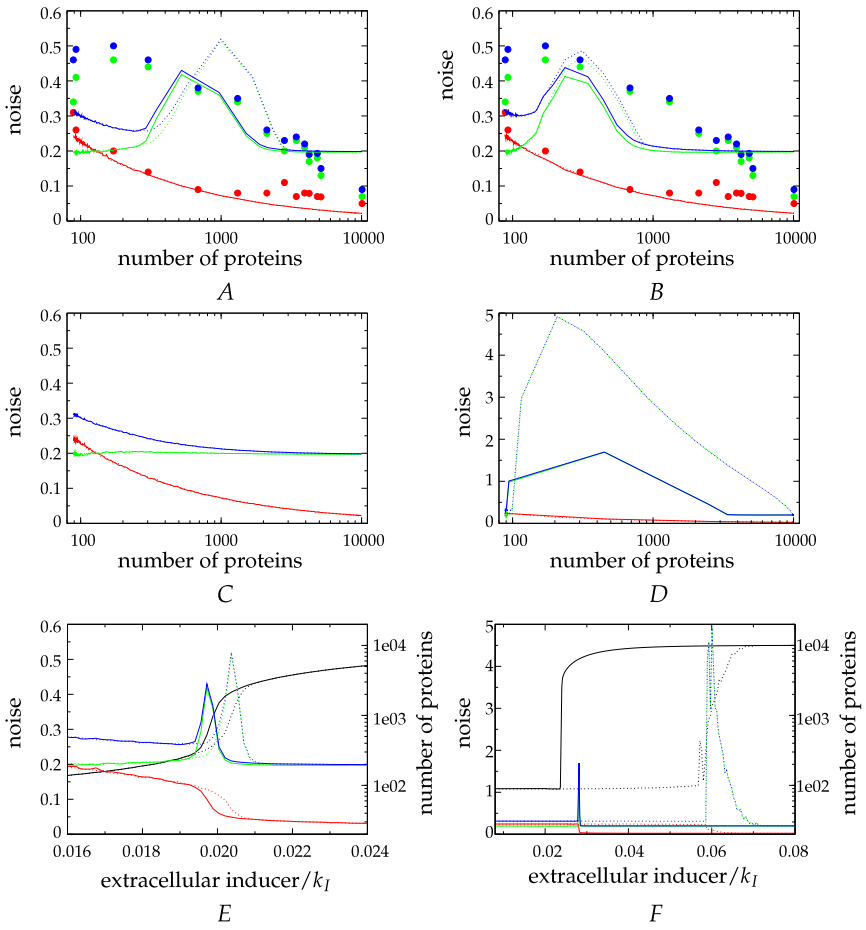


Figure 3.1: Noise levels in the model compared with the experimentally observed noise. (*A*) Noise as a function of the average number of proteins, for IPTG (k_{IPTG}=1.35/min). The red lines indicate intrinsic noise, the green lines extrinsic noise and the blue lines total noise. The filled dots indicate the experimentally observed noise levels, adapted from Elowitz *et al.* (2002). Solid lines indicate that cells are initially induced, dotted lines that cells are initially repressed. (*B*) Noise levels in the model using the higher value for the protein dependent inducer efflux and k_{IPTG}=0.10/min. The maximum in the extrinsic noise is shifted to lower protein numbers. (*C*) Noise levels in the model using lactose as inducer (default model), leading to much lower extrinsic noise. (*D*) Noise levels in the model using TMG as inducer (k_{TMG}=0/min), which leads to an increase of the extrinsic noise. (*E*) Noise as a function of extracellular IPTG concentrations. Inducer concentrations are scaled relative to the binding affinity of the inducer with LacI, the repressor protein of the operon. The black curves indicate the mean protein number. (*F*) Noise as a function of extracellular TMG concentrations. See also the color plate on page 146.

itive feedback, but only at intermediate inducer (hence protein) concentrations, where the promoter function is steepest. Note that the cell cycle and the intracellular inducer concentration are the only extrinsic noise sources we included in the model.

The maximum in the extrinsic noise is located at lower protein concentrations in the data than in our model. This is because the extrinsic noise is high if the positive feedback is strong. Sufficient positive feedback can only be accomplished if the protein-dependent and protein-independent inducer influx are of the same order of magnitude. Therefore, when the protein-independent inducer influx is high, the maximum in the extrinsic noise will be located at high protein concentrations. We can reproduce the experimental data better if we assume a lower growth rate or a higher protein-dependent inducer efflux. Both these changes diminish the positive feedback and therefore we need a smaller protein-independent influx to fit the maximum amount of extrinsic noise. Therefore, the maximum in extrinsic noise will also be shifted to lower protein concentrations. When we use, for example, a protein-dependent inducer efflux of 300 mM/(mM permease min) instead of 49.35, the value reported by Cheng *et al.* (2001), we indeed find that the maximum in the extrinsic noise shifts to lower protein numbers (Fig. 3.1 *B*.) We obtained these curves for a protein-independent inducer influx (k_{IPTG}) of 0.1/min.

For lactose as inducer, the picture is very different (Fig. 3.1 *C*). The intrinsic noise remained unchanged, but extrinsic noise changed considerably. There is still a maximum in extrinsic noise, but this is barely visible. The reason is that the positive feedback loop is much weaker for lactose than for IPTG, due to the degradation of lactose by β-galactosidase. Indeed, the operon used is monostable for lactose. For lactose, we found that extrinsic noise is almost completely determined by the cell cycle for all protein numbers, instead of only for high protein numbers as we found for IPTG. We conclude that the results of the experiments can only be understood when we realize that IPTG was used as an inducer. When lactose as inducer is used, extrinsic noise levels as high as in the experiments can in our model only be observed when the repressed transcription rate is considerably lower.

Finally, we tried to simulate noise in gene expression for TMG, another artificial inducer. As IPTG, TMG is not degraded by β-galactosidase, but in contrast to IPTG, there is no permease-independent influx. Therefore, the positive feedback loop is stronger for TMG. We simulate TMG by using the same parameters as for IPTG, except the permease influx rate is changed from 1.35/min to 0/min. We observe that the extrinsic noise increases drastically (Fig. 3.1 *D*), while the intrinsic noise again remains unchanged. Indeed, the stronger positive feedback loop increases extrinsic noise.

Experimentally, it has been observed that the wild-type *lac* operon behaves bistably for both IPTG and TMG (Ozbudak *et al.*, 2004). For IPTG, however, bistability is expected to be much less severe (Ozbudak *et al.*, 2004). We find that the experimentally observed noise can best be described by a promoter function that is just bistable for IPTG. This can be seen when we compare the amount of noise for initially induced and repressed cells. Although all individual cells are in equi-

librium, the results are different when we start with an initially induced population or an initially repressed population. For TMG, we also find that the promoter function is much more bistable.

To observe the effect of bistability better, we also show how noise levels depend on the extracellular inducer concentration, for both IPTG and TMG (Fig. 3.1 *E* and *F*). The hysteresis-loop, indicating that for certain extracellular inducer concentrations the amount of proteins is dependent on the history of the cells, is clearly visible. Indeed, for TMG this loop is much larger than for IPTG. Furthermore, we see that when inducer concentrations are low, transitions from the induced to the repressed equilibrium are more likely, and vice versa.

3.2.2 Effects of Stochasticity on Population Heterogeneity

In the previous section, we showed that stochasticity in gene expression can cause significant amounts of noise when the operon is repressed. The amount of noise at intermediate inducer concentrations was, however, much larger for IPTG and TMG than for lactose.

These experiments and simulations were done using cells with identical promoters in a well-mixed environment. In natural and evolving populations, different cells will have slightly different promoter functions, due to genetic variation, and the environment will not be well-mixed (Dekel & Alon, 2005; Clarke, 1977). Both these factors can cause population heterogeneity, but are neglected in almost all experiments. This makes sense if we want to study whether stochasticity in gene expression occurs at all. Stochasticity in gene expression can only be proven when gene expression is different in two identical cells in an identical environment. For the importance of stochasticity in natural circumstances, however, other factors causing population heterogeneity need to be considered.

We embedded the intracellular stochastic model of *lac* operon dynamics in an evolutionary, spatial context, as described in the Materials and Methods section and in more detail in **chapter 2**. We used this stochastic model and the deterministic model from **chapter 2** to unravel the contributions of the various factors on population heterogeneity.

To this end, we performed evolutionary simulations in both a spatial and a well-mixed environment, for both the stochastic model and the deterministic model. In these simulations, cells evolve their promoter function in an environment in which the glucose and lactose concentrations fluctuate. To understand the effect of genetic diversity, we also performed simulations using a genetically identical population. We used the last common ancestor (the last individual cell that had the whole population at the end of a simulation as offspring) of each evolutionary simulation for this. In this way, we can compare the effects when genetic variation, spatial variation, both, or none are incorporated, in both the deterministic model (Fig. 3.2 *A*) and the stochastic model (Fig. 3.2 *B*), yielding eight different simulations.

Of the four sources of population heterogeneity in our model, spatial, genetic, stochastic, and cell cycle related, we can exclude all, except for the cell cycle. When all these three noise sources were excluded, we still observed some

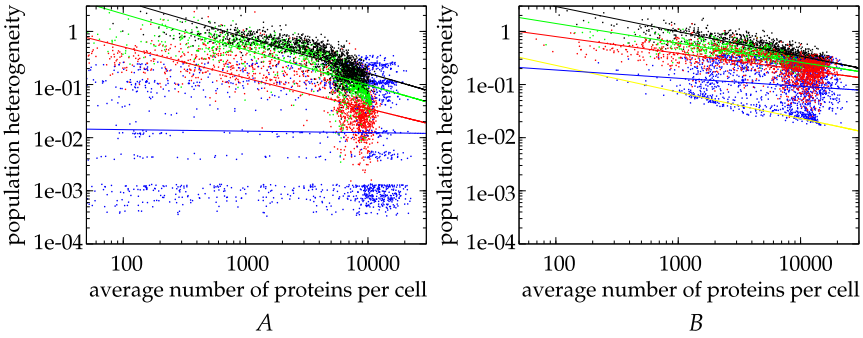


Figure 3.2: Population heterogeneity as a function of protein number, both for the deterministic and stochastic simulations. (*A*) Population heterogeneity in the deterministic simulations, expressed as standard deviation/mean of the number of β-galactosidase proteins per cell. Different colors represent different simulations. All dots indicate the population heterogeneity at certain equally spaced timepoints. Solid lines give the result of a power-law regression between these dots. Black: an evolutionary simulation in a spatial environment, both genetic and spatial heterogeneity, evolutionary time scale. Red: an evolutionary simulation in a well-mixed environment, genetic heterogeneity, but no spatial heterogeneity, evolutionary time scale. Green: a simulation with one clone (the last common ancestor of the "black" simulation), spatial heterogeneity, but no genetic heterogeneity, physiological time scale. Blue: a simulation with one clone (the last common ancestor of the "red" simulation), no spatial heterogeneity, no genetic heterogeneity, physiological time scale. (*B*) Population heterogeneity in the stochastic simulations. All colors are as in Fig. 3.2 *A*, except yellow: regression curve of intrinsic noise versus average protein number (from Fig. 3.1). See also the color plate on page 147.

population heterogeneity, but it was completely independent of protein number (Fig. 3.2 *A*, blue dots). The population heterogeneity varied wildly over time, due to the partial synchronization of the cells. Sometimes all cells have just divided, and population heterogeneity is very low. When only half of the population has recently divided, population heterogeneity is maximal. This explains the extremely broad distribution of blue dots in Fig. 3.2 *A*.

When genetic or spatial heterogeneity was present, very low values of population heterogeneity did not occur anymore (Fig. 3.2 *A*, red, green, and black dots). When these sources of population heterogeneity were present, population heterogeneity became inversely correlated with protein number. It is clear that in our model, space has a larger effect on population heterogeneity than genetic variation (see Fig. 3.2 *A*, red and green dots).

When we compare population heterogeneity in the deterministic model (Fig. 3.2 *A*) with the population heterogeneity in the stochastic model (Fig. 3.2 *B*), we find, as expected, the largest difference when spatial and genetic heterogen-

eity are both absent. Intrinsic noise (Fig. 3.2 *B*, yellow line) gives a lower boundary to the population heterogeneity in the stochastic model. The mean population heterogeneity is increased considerably by stochastic gene expression. When genetic or spatial heterogeneity was present in the stochastic model (Fig. 3.2 *B*, red and green), we observed, however, that much of the difference in population heterogeneity between the deterministic model and the stochastic model disappeared. Population heterogeneity due to stochastic gene expression, therefore, apparently is small compared with population heterogeneity due to spatial and genetic variation.

Intrinsic noise, as shown in the previous section, follows power law behavior with respect to protein number, with a coefficient of 0.5, just like Poisson noise. Surprisingly, both in the deterministic model (Fig. 3.2 *A*) and the stochastic model (Fig. 3.2 *B*) we see that if spatial or genetic variation is present, the data can still reasonably be described using a power law. In Table 3.1, we give the regression and correlation coefficient of a power law regression of the data.

Table 3.1: Regression and correlation coefficients.

Model	Source of Heterogeneity	Regression	Correlation
Stochastic	Black: spatial+genetic	-0.47	-0.85
	Green: spatial	-0.37	-0.86
	Red: genetic	-0.32	-0.58
	Blue: no	-0.16	-0.16
	Yellow: intrinsic noise	-0.51	-1.00
Deterministic	Black: spatial+genetic	-0.68	-0.81
	Green: spatial	-0.67	-0.78
	Red: genetic	-0.59	-0.71
	Blue: no	-0.029	-0.025

The regression coefficient indicates how strong population heterogeneity and protein number are correlated. The more sources causing population heterogeneity, the higher the regression coefficient is. Genetic heterogeneity correlates with protein number because the selection pressure on the induced operon is higher than on the repressed operon. The reason for this is that the cost for promoter activity is much more important for the induced operon than for the repressed operon. This appears to be a reasonable assumption that is likely to hold in the natural environment of *E. coli*.

Population heterogeneity due to space is largest at intermediate extracellular inducer concentrations. At very high inducer concentrations, noise in the inducer concentrations does not cause noise in gene expression, because of the sigmoid shape of the promoter function. This also holds for very low extracellular inducer concentrations. For intermediate inducer concentration, gene expression is very much influenced by spatial heterogeneity. We observed, however, a monotonic relationship between protein number and noise due to spatial heterogeneity. This

is because when the promoter activity of cells becomes low, cells stop lactose consumption, and therefore cells never experience very low lactose concentrations. This effect is likely to play a role in the natural environment, but in the natural environment, lactose degradation is probably not only due to *E. coli*, and we might expect noise due to spatial heterogeneity not to decrease monotonically with protein number.

From all this we expect that only in a monomorphic population, living in a well-mixed environment, does stochasticity in gene expression play an important role. Whether stochasticity influences evolution, however, is therefore doubtful. To study whether stochasticity does or does not play an important role in evolution, in the next section we compare evolution of the *lac* operon in the stochastic and deterministic models. We do this in a well-mixed environment, because there stochasticity is expected to have the largest influence.

3.2.3 Effects of Stochasticity on Evolution in a Well-Mixed Environment

We performed six independent evolutionary simulations in well-mixed environments, in both the deterministic model and the stochastic model. For all 12 last common ancestors, we plotted the promoter activity if no glucose is present (Fig. 3.3). Most of the changes between the initial promoter function (dotted line) and the evolved promoter functions are consistent between all simulations. The induced transcription rate is increased, to approximately the same value in all 12 simulations. The steep part of the promoter function evolved to approximately 10-fold lower allolactose concentrations. Finally, the repressed transcription rate increased very significantly. As proven in **chapter 2**, the repressed transcription rate in large part determines whether a promoter function is bistable. The initial promoter function was chosen to be bistable. Of the 12 last common ancestors, only one was bistable. This is indeed the promoter function with the lowest repressed transcription rate. There appears to be a trend that promoter functions, which evolved in the stochastic simulations, have higher repressed transcription rates than those evolved in the deterministic simulations.

To check this more precisely, we calculated the time average of the repressed transcription rate at zero glucose for all 12 simulations. There is considerable variation in this quantity over time during the whole evolutionary simulation.

The first quarter of the evolutionary simulation was not taken into account, to give the cells time to adapt somewhat to the environment. We found that on average in the stochastic simulations the repressed transcription rate was 5.5-fold higher than in the deterministic simulations. In the stochastic simulations, we found an average repression rate of approximately 45, while for the deterministic model this was approximately 250. Previously, a repression rate of approximately 170 was experimentally found (Ozbudak *et al.*, 2004).

Whether this difference is significant, we checked by permuting the average repressed transcription rates over the different simulations and calculating the difference between the stochastic and deterministic simulations for all possible permutations. In less than 3% of the permutations, we found a larger average

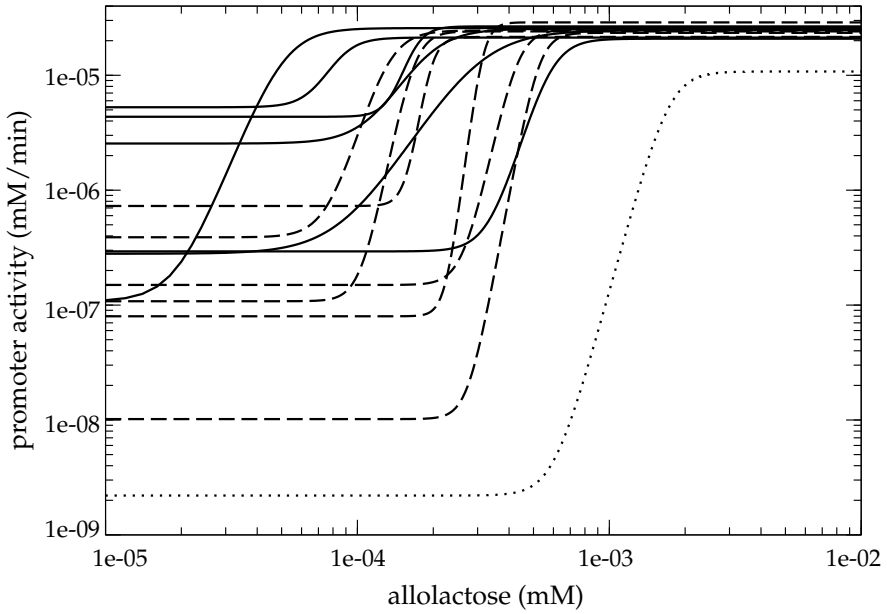


Figure 3.3: Promoter functions at zero glucose, evolution in a well-mixed environment. Dotted line, initial promoter function. Solid lines, promoter functions of the six last common ancestors of the stochastic evolutionary simulations. Dashed lines, promoter functions of the six last common ancestors of deterministic evolutionary simulations.

difference than observed, which gives a measure of the significance of this result.

Next, we performed competition experiments between two promoter functions, one evolved in the stochastic model, the other in the deterministic model. We used the same environment as was used for the evolutionary simulations. For clarity we chose the promoter function with the highest repressed transcription rate and the promoter function with the lowest repressed transcription rate in Fig. 3.3. Both these promoter functions perform well in comparison with the five other evolved promoter functions.

We placed these two promoter functions in the stochastic and deterministic model. When one population died out, we stopped the simulation and scored which promoter function had died out. No mutation was allowed during these competition experiments. Both for the stochastic model and for the deterministic model, we performed 100 of these competition experiments. In the deterministic model, we found that in 61 cases the promoter function with the low repressed transcription rate won, while in the stochastic model in 60 cases the promoter function with the high repressed transcription rate won. Using a two-tailed binomial test, this corresponds to a p-value of 0.057 and 0.035, respectively. Combined, this gives a p-value of 0.0020. This confirms that the promoter with the low

repressed transcription rate performs better in the deterministic model, while the promoter with the high repressed transcription rate performs better in the stochastic model. Due to the small population sizes during the simulations (on average approximately 200), there is a lot of drift. This causes the most competitive promoter function to not always win the competition experiment.

To understand this result, we studied the dynamics of these two promoter functions in both the deterministic model and the stochastic model. For both models, we again initiated a heterogeneous population and followed the dynamics of two individual cells, one with a high and one with a low repression rate. In both models, we used an identical environment, such that we can compare the dynamics. An example of the dynamics during a period of lactose influx is shown in Fig. 3.4. Fig. 3.4 *A* shows the dynamics of both promoter functions in the stochastic model, while Fig. 3.4 *B* shows the dynamics in the deterministic model, in the same environment.

The promoter function with the high repressed transcription rate behaves almost identically, whether placed in the stochastic or the deterministic model. The only observable difference is caused by the cell cycle. For the promoter function with the low repressed transcription rate, the picture is very different. This promoter function is slightly bistable. Only when the external lactose concentration exceeds approximately 0.02 mM, does the operon switch on. This causes a slight delay in protein production, as can be seen in Fig. 3.4 *B*. In the stochastic model, the delay is, however, much larger. Only after approximately five hours, does the operon switch on. Because protein production occurs in bursts, and not in a gradual way, the cell has to wait for a sufficiently large burst to become induced. It is important to note that for the lactose concentrations the cells experience after approximately 4.5 hours, the operon is not bistable in the deterministic model, but, nevertheless, in the stochastic model the operon is not able to switch to the induced state.

Fig. 3.4 only shows the dynamics during one lactose pulse. In the evolutionary and competition experiments, many different pulses, of different height and length, are experienced. Also, the periods without influx have different lengths. Therefore, for every pulse the situation is somewhat different. During the periods without lactose influx, the promoter with the low repressed transcription rate has a higher growth rate than the promoter with the high repressed transcription rate, because its cost for operon activity is lower. When lactose influx starts, the operon with the high repressed transcription rate has a higher growth rate, because it can start lactose uptake earlier.

Therefore, depending on the state of the environment (whether there is lactose influx or not), the growth rate difference between the two promoters will be positive or negative. In an environment with the same parameters as the environment used for the evolutionary and competition experiments, we kept track of the growth rates during a period of time approximately equal to the time that is needed for one competition experiment. The result is shown in Fig. 3.5. The dotted line indicates that the promoter function with the high repressed transcription rate on average has a lower growth rate in the deterministic model than the promoter function with the low repressed transcription rate, although the growth

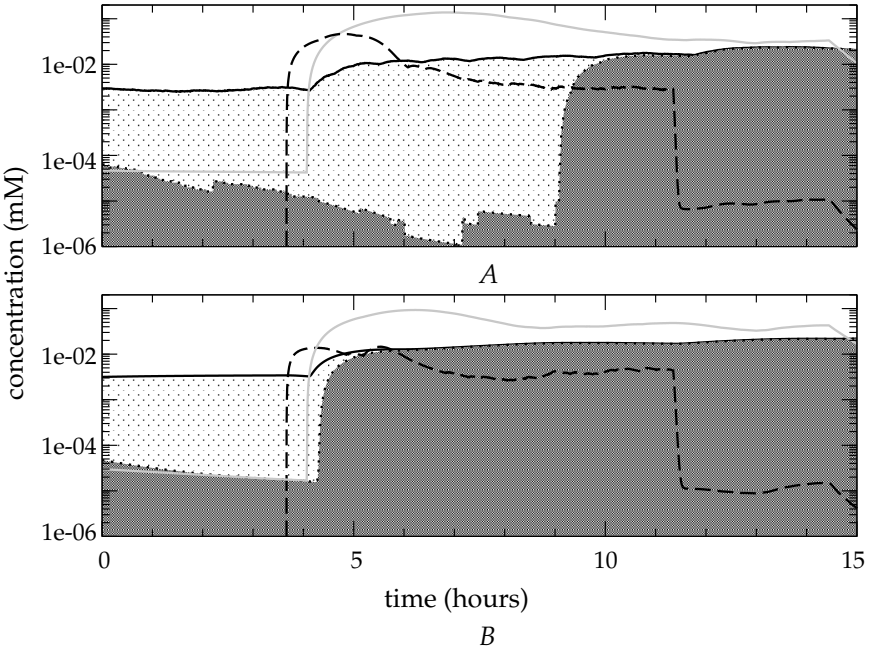


Figure 3.4: Dynamics of two promoter functions. (*A*) The stochastic model: the gray line indicates external lactose concentration; the dashed line external glucose concentration; and the solid and dotted lines the β-galactosidase concentration (which is converted to mM to compare it with the deterministic model, a concentration of 2e-06 mM corresponds with 1 molecule per cell, just after cell division). The solid line indicates the β-galactosidase concentration for the promoter function with the high repressed transcription rate, while the dotted line corresponds to the promoter function with the low repressed transcription rate. (*B*) Same as (*A*), for the deterministic model. Note that, because cells behave differently in both models, the extracellular glucose and lactose concentrations are not identical (but very similar) in both (*A*) and (*B*).

rate difference indeed fluctuates, according to whether lactose is present in the environment or not. Again this proves that the promoter function with the low repressed transcription rate performs better in the deterministic model. When we compare both promoter functions in the stochastic model (the solid and the dashed line in Fig. 3.5) we see, however, that the average growth rate of the promoter function with the high repressed transcription rate is highest.

This all shows that promoter functions evolved in the stochastic simulations evolve to a higher repressed transcription rate, which make them outcompete promoter functions with very low repressed transcription rates. In the deterministic simulations, the situation is the opposite, and these high repressed transcription rates are not optimal.

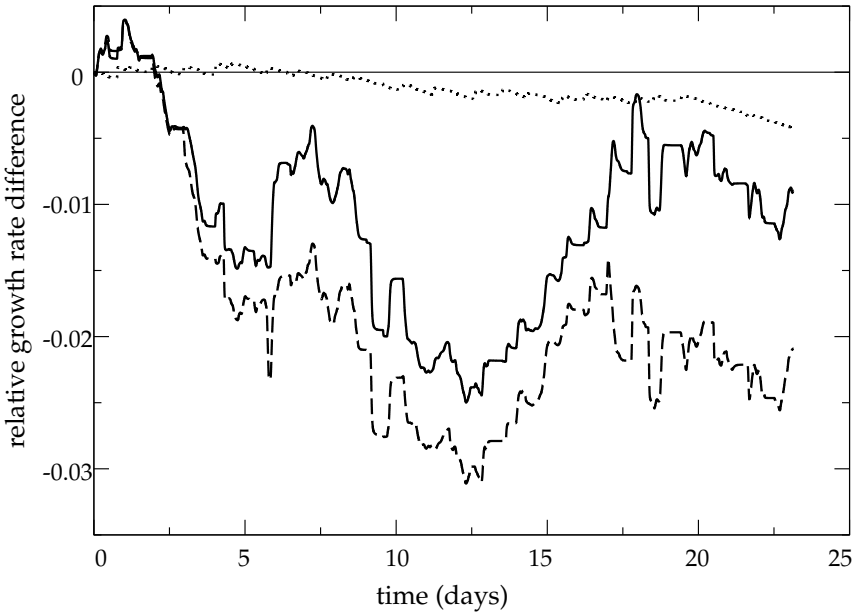


Figure 3.5: Growth rate of the two promoter functions in both models. Here we report the growth rate difference compared with the promoter function with the low repressed transcription rate in the deterministic model. Solid line, promoter function with the high repressed transcription rate in the stochastic model. Dashed line, promoter function with the low repressed transcription rate in the stochastic model. Dotted line, promoter function with the high repressed transcription rate in the deterministic model.

3.2.4 The Importance of Nonequilibrium Conditions

Whether the cell is in equilibrium has considerable effect on the noise levels. In this section, we discuss two examples in which nonequilibrium dynamics is crucial for the amount of noise.

Bistability in a deterministic system causes hysteresis. Depending on the history of the system, the system will be in one of the two equilibria. This hysteretic behavior disappears when stochasticity is added. In relatively short time scales, the system remains hysteretic, i.e., when cells were induced, they remained induced, while when cells were repressed, they remained repressed. In longer time scales, transitions between the two equilibria are possible. Therefore, the probability distribution of the state of cells goes to a stable equilibrium, with some cells in the repressed and others in the induced state, depending not on the history but on the transition probability between the two equilibria. Therefore, if we would have waited long enough, the difference in noise levels between initially repressed and induced cells as shown in Fig. 3.1 would have disappeared. Most of the cells would have gone to the induced state, because the transition probab-

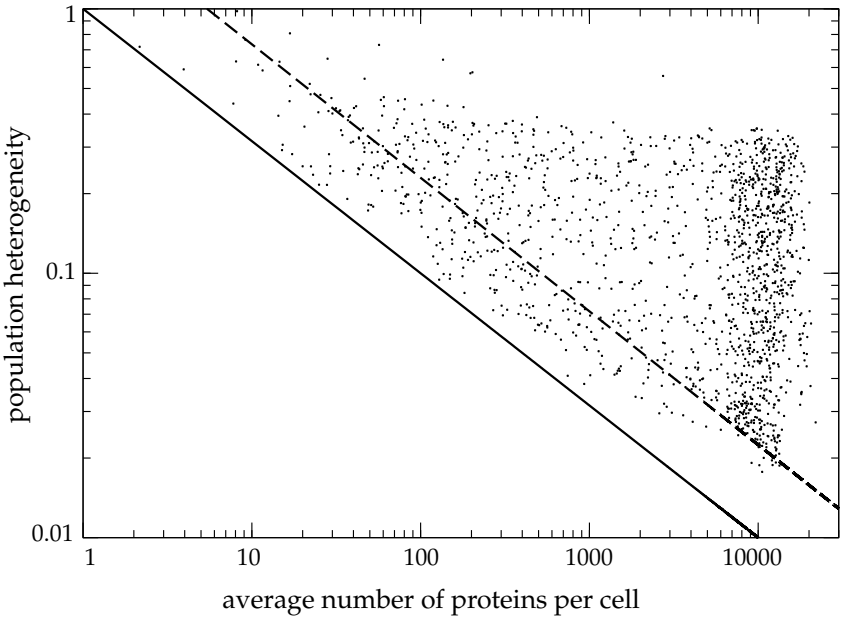


Figure 3.6: Population heterogeneity can be lower than the intrinsic noise. One promoter function is placed in a well-mixed environment; therefore both genetic and spatial heterogeneity are absent. Dashed line, intrinsic noise (from Fig. 3.1): $\sqrt{b+1}/\sqrt{P}$, with P the number of proteins in the cell. Solid line: Poisson noise: $1/\sqrt{P}$.

ility to go from the induced to the repressed state is lower than vice versa, due to the lower noise for the induced operon.

A second example is shown in Fig. 3.6. Here we show the population heterogeneity of a promoter function belonging to a last common ancestor of an evolutionary simulation. Both genetic and spatial heterogeneity are absent. We observe that the population heterogeneity is frequently lower than the intrinsic noise. Intrinsic noise is in general seen as inherent and therefore unavoidable for a cell. We would expect then that noise cannot be lower than the intrinsic noise. Here we see, however, that this is not necessarily so.

This striking observation can be understood when we realize that the intrinsic noise was measured in cells that were in equilibrium. During our simulations, cells were very often not in equilibrium. Intuitively, we might expect that cells that are not in equilibrium have even higher population heterogeneity, but this is apparently not the case. If cells have a much higher protein concentration than the equilibrium value (if, for example, external lactose has just been depleted), the protein concentration decreases. In such a situation, transcription can be neglected and the dynamics are purely determined by protein degradation and dilution. Noise caused by protein degradation and dilution is, however, much

lower than transcriptional noise, because degradation does not occur in bursts. Noise due to degradation and dilution can be described as $\eta_{deg} = 1/\sqrt{P}$, Poisson noise (solid line in Fig. 3.6). Translational noise is, however, much broader than Poisson: $\eta_{trans} \approx \sqrt{b+1}/\sqrt{P}$ (dashed line in Fig. 3.6), where b is the burst size. Indeed, we observe that Poisson noise does give a lower boundary to the population heterogeneity, whereas the intrinsic noise does not.

3.3 Discussion

We study the influence of stochasticity in gene expression on evolutionary adaptation of the *lac* operon of *E. coli*. To this end, we used a detailed quantitative model of the *lac* operon in which stochasticity is incorporated on the protein level. This approach has the advantage that only one (experimentally known) parameter needs to be added to the model to make it stochastic, namely the average burst size of protein translation. We find good agreement between noise levels in our model and experimental noise measurements (Elowitz *et al.*, 2002).

The experimentally observed noise levels, however, can only be explained when we realize that IPTG, which is not degraded by β-galactosidase, is used as inducer. We find that induction by IPTG leads to very different dynamics than induction by the natural inducer lactose (see also **chapter 2**). When the operon is induced by lactose, stochasticity in gene expression is strongly reduced, and the total noise is mostly determined by the cell cycle. This is due to the fact that degradation of lactose by β-galactosidase reduces the strength of the positive feedback loop. Induction by TMG, however, leads to even higher noise values, because in contrast to IPTG, there is no protein-independent TMG influx. It would be very interesting to measure noise in gene expression of the *lac* operon, for TMG, IPTG, and lactose, to validate these results.

In literature, different values for the Hill-coefficient are reported. For example, in Ozbudak *et al.* (2004) a Hill-coefficient of 2 is used, while in Setty *et al.* (2003) a value of 4.0 was measured. When the Hill-coefficient is high, the positive feedback is strong and noise values are high. For IPTG, however, the noise curves would be very similar, because we chose the protein-independent inducer influx such that the maximal amount of noise corresponded to the experimentally measured noise values. For lactose we found that, all other parameters being equal, the promoter function only becomes bistable when the Hill-coefficient is larger than 52, which is clearly unrealistically high. During the evolutionary experiments, the Hill-coefficient can be mutated and it mostly varies between 2 and 10.

To investigate the effect of different noise sources on population heterogeneity, we compared the amount of population heterogeneity in simulations with and without space, mutations, and stochasticity in gene expression. In these simulations the operons were induced by lactose, instead of by artificial inducers. We observed that only in the well-mixed simulations, without mutation, was the amount of population heterogeneity much larger in the stochastic than in the deterministic simulation. If spatial or genetic heterogeneity was added, stochasticity hardly influenced population heterogeneity. Surprisingly, we found that both

genetic and spatial heterogeneity decrease with protein number, more or less in the same way as population heterogeneity by stochastic gene expression. Especially for genetic heterogeneity, we expect this also to be the case in nature.

In nature, it seems likely that the spatial heterogeneity is very large. The gut is a highly diverse ecosystem (Clarke, 1977), and not at all well-mixed. It has been shown that *E. coli* is able to entirely change its *lac* operon promoter function in a few hundred generations (Dekel & Alon, 2005). This suggests that in nature the genetic diversity is also high. Therefore, we believe that the large genetic and spatial differences in our model are biologically realistic. However, we did check our results for a ten times lower mutation rate and ten times higher diffusion constant, which determine genetic and spatial heterogeneity, respectively. Even when both the mutation rate and the diffusion constant are modified, we observe that noise due to stochastic gene expression is only comparable to spatial and genetic heterogeneity at very low protein numbers (unpublished data).

Finally, we directly compared the promoter functions evolved in the deterministic and the stochastic evolutionary simulations. We observed that in the stochastic simulations cells evolve to higher repressed transcription rates and thus prevent stochasticity in gene expression. We show that this is because in the stochastic model low transcription rates cause longer delays than in the deterministic model.

This is in striking contrast with the result found in Blake *et al.* (2006). There it was found that in *Saccharomyces cerevisiae*, bursts in gene expression enable a more rapid cell response. When we initialize cells in an environment with a fixed inducer concentration, for which the operon is bistable, we also see that larger bursts enable a more rapid response to this inducer concentration than smaller bursts. Indeed, in the deterministic model, cells would stay indefinitely in the "wrong" equilibrium. Because in our model, however, the inducer concentration varies over time, the inducer concentration quickly increases over the point at which the operon is still bistable, and in the deterministic model the cells then respond very fast. In the stochastic model, cells still need to wait for a sufficiently large burst to occur even when the inducer concentration has increased over the concentrations for which the cells are bistable. Even when the average burst size is large, this takes a long time, because then the frequency of the bursts is lower. This explains why the net effect of stochasticity in our model is negative for such promoters (compare the growth rate of the promoter function with the low repressed transcription rate in the deterministic and the stochastic models; Fig. 3.5, dashed line). Furthermore, having a somewhat higher repressed transcription rate ensures that *all* cells have a rapid response.

Both in the deterministic and the stochastic simulations, bistability with respect to lactose is most often avoided (except at high glucose concentrations, which are very rare). Although the promoter function evolved in the deterministic model, which we used in Fig. 3.4, is slightly bistable, this has little influence on the behavior of this promoter function in the deterministic model, while the behavior in the stochastic model is changed drastically by the bistability.

In **chapter 2** we already showed that when using a 10 times higher cost for *lac* operon activity, or a different environment (with longer or shorter periods with

and without lactose), bistability was also avoided. In the stochastic model, we also performed simulations in different environments, but again the results did not change essentially.

We conclude that stochasticity cannot avoid the inherent disadvantages of bistability (namely longer delays in protein dynamics (see **chapter 2**)). Even more, we showed that bistability is even more deleterious in the presence of stochastic gene expression than in a deterministic system. These conclusions are in line with Fraser *et al.* (2004), who have shown that essential genes in *S. cerevisiae* have evolved to lower noise values than nonessential genes and thus that stochasticity for many genes appears to be detrimental during evolution.

3.4 Materials and Methods

In this section we discuss the computational model that is used in this paper. We used both a deterministic and a stochastic version of the model. The deterministic version is explained in detail in **chapter 2**, but here we describe the major points of the deterministic model. For the stochastic version of the model, we modified this deterministic model in a few ways. These modifications we describe in detail.

3.4.1 Model for Simulating *lac* Operon Evolution.

The deterministic model is a spatially explicit, computational model of *E. coli* cells, growing on glucose and lactose while evolving their *lac* operon promoter function. It consists of an intracellular and an extracellular part.

Intracellular Dynamics

The intracellular dynamics is modeled using ten differential equations, following Wong *et al.* (1997). The following intracellular variables are incorporated: mRNA (M) β-galactosidase (B), permease (P), lactose (L), allolactose (A), glucose (G), glucose-6-phosphate ($G6P$), cAMP (C), ATP (ATP), and cell size (X). Transcription of the *lac* operon, translation, lactose and glucose uptake and metabolism, cAMP and energy production, and cell growth are all modeled in detail. When possible, parameter values are taken from literature. All ten differential equations are integrated using a timestep of 0.2 seconds. Here we shortly discuss all differential equations. A list of all parameters is given in Table 2.1.

Transcription is modeled as a two-dimensional Hill-function, dependent on the cAMP and allolactose concentration (see Setty *et al.* (2003)). In this way, glucose, via cAMP, represses the operon, while lactose, via allolactose, induces the operon. This two-dimensional Hill-function depends on 11 biochemical parameters, such as the k-value and Hill-coefficient of allolactose binding to the repressor. In the evolutionary simulations, these are the only parameters that can mutate, all other parameters are fixed, because we are interested in the evolution of the promoter function, given realistic boundary conditions (which are the other parameters in the intracellular model).

$$\frac{dM}{dt} = PA(A,C) - (\gamma_M + \mu)M, \tag{3.1}$$

$$PA(A,C) = max\Big(V_{mRNA,max}, \frac{a\alpha + \gamma + \frac{d(b\beta+\gamma)(C/k_C)^n}{1+(C/k_C)^n} + \frac{\gamma c}{1+(A/k_A)^m}}{1 + a + \frac{d(b+1)(C/k_C)^n}{1+(C/k_C)^n} + \frac{c}{1+(A/k_A)^m}}\Big) \tag{3.2}$$

$a = RNAP/k_{RNAP}$, RNA-polymerase in units of its dissociation constant for binding to a free site.
$b = RNAP/k_{RNACP}$, RNA-polymerase in units of its dissociation constant for binding to a site with bound CRP.
$c = LACI_T/k_{LACI}$, the total LacI concentration in units of its dissociation constant for binding to its site.
$d = CRP_T/k_{CRP}$, the total CRP concentration in units of its dissociation constant for binding to its site.
α, the transcription rate when RNA Polymerase is bound to the DNA, but CRP and LacI are not.
β, the transcription rate when both RNA Polymerase and CRP are bound, but LacI is not bound to the DNA.
γ, the "leakiness", the transcription rate when RNA Polymerase is not bound to the DNA.
k_A, k-value for allolactose binding to LacI.
m, Hill-coefficient describing cooperativity in binding of allolactose to LacI.
k_C, k-value for cAMP binding to CRP.
n, Hill-coefficient describing cooperativity in binding of cAMP to CRP.

Protein production (β-galactosidase and permease) depends on the mRNA concentration, and proteins are slowly degraded.

$$\frac{dB}{dt} = k_B M - (\gamma_B + \mu)B \tag{3.3}$$

$$\frac{dP}{dt} = k_P M - (\gamma_P + \mu)P. \tag{3.4}$$

Lactose influx is permease-dependent, while lactose degradation and conversion to allolactose is dependent on β-galactosidase. Small protein-independent lactose and allolactose degradation terms also are added.

$$\frac{dL}{dt} = P\left(\frac{k_{Lac,in}L_{ext}}{K_{Lac,in} + L_{ext}} - \frac{k_{Lac,out}L}{K_{Lac,out} + L}\right) - B\frac{(k_{cat,Lac} + k_{Lac-Allo})L}{L + K_{m,Lac}} - (\gamma_L + \mu)L \tag{3.5}$$

$$\frac{dA}{dt} = B\frac{k_{Lac-Allo}L}{L + K_{m,Lac}} - B\frac{k_{cat,Allo}A}{A + K_{m,Allo}} - (\gamma_A + \mu)A. \tag{3.6}$$

Lactose is converted to glucose by β-galactosidase. Glucose uptake from the medium is also incorporated. Glucose metabolism (via glycolysis and TCA-cycle) produces ATP on which the cells grow. ATP production depends on the glycolytic fluxes and the fluxes through the TCA-cycle. ATP is consumed by basal metabolism, cell growth, and *lac* operon activity. The cAMP concentration is assumed to be dependent on the glucose influx rate (see Wong *et al.* (1997)).

$$\frac{dG}{dt} = B\frac{(k_{cat,Lac} + k_{Lac-Allo})L}{L + K_{m,Lac}} - \frac{k_{cat,Glu}G}{G + K_{m,Glu}} - k_{Glu,out}(G - G_{ext}) - \mu G. \quad (3.7)$$

$$\begin{aligned}\frac{dG6P}{dt} &= \frac{k_{t,Glu}G_{ext}}{G_{ext} + K_{t,Glu}} + \frac{k_{cat,Glu}G}{G + K_{m,Glu}} + B\frac{(k_{cat,Lac} + k_{Lac-Allo})L}{L + K_{m,Lac}} \\ &\quad - \frac{k_{G6P,Rsp}G6P}{G6P + K_{G6P,Rsp}} - \frac{k_{G6P,Frm}G6P^8}{K^8_{G6P,Frm} + G6P^8} - \mu G6P \end{aligned} \quad (3.8)$$

$$\frac{dC}{dt} = k_{syn,cAMP}\frac{K_{syn,cAMP}}{\frac{k_{t,Glu}G_{ext}}{G_{ext}+K_{t,Glu}} + K_{syn,cAMP}} - (\gamma_{cAMP} + \mu)C \quad (3.9)$$

$$\begin{aligned}\frac{dATP}{dt} &= \frac{Y_{Rsp}k_{G6P,Rsp}G6P}{G6P + K_{G6P,Rsp}} + \frac{2k_{G6P,Frm}G6P^8}{K^8_{G6P,Frm} + G6P^8} - BMC - GC \times \mu \\ &\quad - PC \times PA(A,C) - B\frac{(k_{cat,Lac} + k_{Lac-Allo})L}{L + K_{m,Lac}} \end{aligned} \quad (3.10)$$

$$\frac{dX}{dt} = \mu_{max}\frac{ATP^4}{ATP^4 + K^4_{ATP}}X. \quad (3.11)$$

The Spatial Model

Cells, of which the dynamics are determined by the above-described intracellular model, are placed on a square grid of 25 $\times$ 25 grid points. These cells grow on extracellular glucose and lactose. The extracellular glucose and lactose concentrations are determined by a fluctuating influx of glucose and lactose into the grid, consumption of glucose and lactose by the cells, and diffusion over the grid. In the well-mixed simulations, we assume infinite diffusion, such that the glucose and lactose concentrations are equal over the whole grid. Furthermore, the cells are shuffled randomly over the grid.

Glucose and lactose influx into the grid is modeled in pulses, independently of each other. Pulses of glucose and lactose influx both have an average duration of 11 hours. The total amount of carbon influx is on average equal for each pulse, such that short pulses have high influx rates, and vice versa. Every pulse has a probability of 10% of being instantaneous, such that very high glucose or lactose concentrations also sometimes occur. In 25% of these cases, simultaneous glucose

AND lactose influx occurs, in order to enforce simultaneous high glucose AND lactose concentrations. These pulses are followed by a period without glucose or lactose influx, also on average 11 hours.

This environment is chosen for two reasons. First, all combinations of glucose, lactose, glucose AND lactose, and neither one occur repeatedly. Second, the average length of the periods is chosen such that cells can just adapt their protein concentrations to the environment.

The cells consume glucose and lactose, grow (as the growth rate is given by the intracellular model), divide (if their size has doubled and there is space to reproduce), and move randomly over the grid. The cells are diluted in a density-dependent way, and cells "die" if they have zero energy (although this happens very rarely). Therefore, fitness differences are mainly caused by differences in growth rates between cells (which are again dependent on glucose and lactose uptake) instead of by differences in death rates.

In the evolutionary experiments, we always start with a monomorphic population of approximately 60 cells that are bistable with respect to lactose and have a diauxic shift (such that glucose and lactose are consumed sequentially). After each cell division, the daughter cell has a probability of 0.01 to mutate one of the 11 biochemical parameters that determine the shape of the *lac* operon promoter function. All other parameters are kept constant. This means that cells can adapt the precise shape of the promoter function to the (fluctuating) environment. We let the cells evolve for approximately 2,000 days, which is sufficient to adapt to the environment.

3.4.2 Modification of the Model to Incorporate Stochasticity

It has been predicted (McAdams & Arkin, 1997) and recently experimentally verified (Cai *et al.*, 2006) that protein production occurs in bursts and that the size of these bursts, i.e., the number of proteins that are produced per burst, is geometrically distributed. The reason for this is that mRNA molecules are translated several times, before they are degraded. This leads to the rapid production of several proteins once an mRNA molecule is produced, and protein production ends when the mRNA molecule is degraded. When an mRNA molecule is translated, it cannot be degraded. Therefore, after each translation, the mRNA molecule has a probability p to be translated again and a probability $1 - p$ to be degraded. From this it follows that protein production occurs in bursts with a burst size that is geometrically distributed. The experimentally observed mean burst size is five β-galactosidase proteins (Cai *et al.*, 2006).

Using only this experimental observation, we changed the intracellular part of the above-described model to incorporate stochasticity in gene expression. This was done in the following way. We model the mRNA concentration in the cell in the same way as in the deterministic model, i.e., it only depends on the transcription and decay rate of mRNA. We now, however, interpret the mRNA concentration (if this concentration is smaller than the concentration corresponding to one molecule) as the probability that an mRNA molecule is present in the cell. We then use this probability to directly infer the frequency that a translational burst

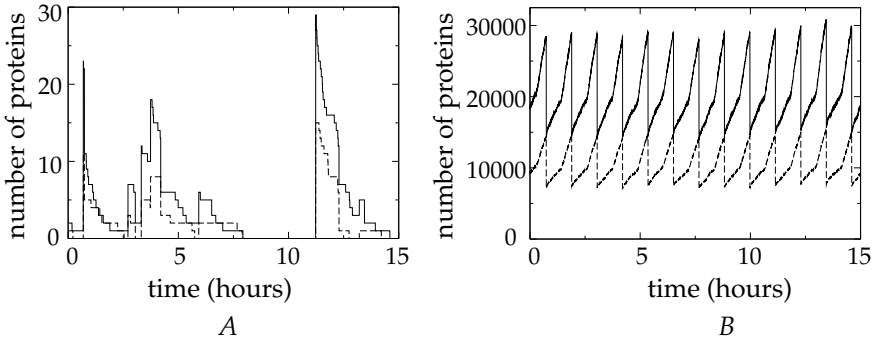


Figure 3.7: Protein dynamics of the stochastic model. (*A*) Protein dynamics for a repressed operon. Solid lines are the number of permease molecules, dashed line the number of β-galactosidase molecules. (*B*) Protein dynamics for an induced operon.

occurs, because we also know the average translation rate from the deterministic model. For example, when a certain mRNA concentration leads to a translation rate of five proteins per minute in the deterministic model and given that the average burst size equals five proteins, this now leads to a burst frequency of one burst per minute. In this way, although stochasticity is mostly determined on the mRNA level, we can still model the mRNA concentration deterministically.

Furthermore, we assume that bursts of permease and β-galactosidase are correlated, but that the number of permease molecules per burst is twice the number of β-galactosidase molecules, because the translation rate of permease is twice as large as that of β-galactosidase (Kennell & Riezman, 1977). The unit of protein levels is now the number of molecules, and we use a minimum cell size of 8×10^{-16} l (Yildirim & Mackey, 2003). The units of all other variables are still mM.

Protein degradation is modeled binomially. When a cell divides, the proteins are divided randomly between the cells. Furthermore, we assume instant DNA replication, when a cell has an (arbitrarily chosen) size of 1.5 times the minimum cell size, while the cell divides at a size two times the minimum cell size. Such cell division dynamics was not taken into account in the deterministic model. In this way, we have incorporated stochasticity in gene expression with adding only one experimentally known parameter, namely the average translational burst size.

The dynamics of this model is very similar to the dynamics of previously developed models of stochastic gene expression in *E. coli* (Kierzek *et al.*, 2001; Swain *et al.*, 2002). The "steady state" dynamics for a repressed and an induced operon in our model is shown in Fig. 3.7.

Table 3.2: Parameters used for lactose.

Parameter	Value
k_C	1.0×10^{-4} mM
n	4.0
k_A	5.5×10^{-4} mM
m	4.0
a	1.0
b	1.0
c	1.0×10^{6}
d	50
α	1.1×10^{-7} mM/min
β	4.4×10^{-5} mM/min
γ	1.89×10^{-7} mM/min

3.4.3 Model for Studying Noise Levels for Lactose, IPTG, and TMG

Lactose

For the simulations where we studied the amount of noise in the *lac* operon, we only considered Eqs. 3.1-3.6 and 3.9. For Eq. 3.3 and Eq. 3.4, we used the stochastic counterparts. We assumed a fixed growth rate of 0.01/min and (for the cAMP dynamics) zero glucose influx. The parameters we used are listed in Table 3.2 (if not mentioned, the parameters are equal to those listed in Table 2.1).

IPTG

IPTG is not degraded by β-galactosidase and binds directly to LacI. Therefore, the equation for promoter activity becomes

$$PA(I,C) = max\Big(V_{mRNA,max}, \frac{a\alpha + \gamma + \frac{d(b\beta+\gamma)(C/k_C)^n}{1+(C/k_C)^n} + \frac{\gamma c}{1+(A/k_I)^m}}{1 + a + \frac{d(b+1)(C/k_C)^n}{1+(C/k_C)^n} + \frac{c}{1+(A/k_I)^m}}\Big) \tag{3.12}$$

Furthermore, we add a protein-independent inducer influx term, which gives us for the IPTG dynamics

$$\frac{dI}{dt} = P(\frac{k_{IPTG,in} I_{ext}}{K_{IPTG,in} + I_{ext}} - \frac{k_{IPTG,out} I}{K_{IPTG,out} + I}) - \mu I + k_{IPTG} I_{ext}. \tag{3.13}$$

The parameters we used for IPTG are listed in Table 3.3 (see Cheng *et al.* (2001)).

We found that, because there is no IPTG degradation and very little efflux, the external IPTG concentrations, for which the promoter function becomes induced

Table 3.3: Parameters used for IPTG.

Parameter	Value
$k_{IPTG,in}$	495 mmol IPTG/(mmol permease min)
$K_{IPTG,in}$	0.42 mM
$k_{IPTG,out}$	49.35 mmol IPTG/(mmol permease min)
$K_{IPTG,out}$	21 mM
k_{IPTG}	1.35/min
k_I	5×10^{-3} mM

when using the same value for k_I as for induction by lactose, become unrealistically low. Therefore, we used a different value for k_I than for k_A. However, changing k_I only shifts the promoter function to different inducer concentrations, and the noise dependence with respect to protein number remains unchanged.

TMG

For TMG we used exactly the same parameter values as for IPTG, except that there is no protein-independent inducer influx.

$$\frac{dT}{dt} = P\left(\frac{k_{TMG,in}T_{ext}}{K_{TMG,in} + T_{ext}} - \frac{k_{TMG,out}T}{K_{TMG,out} + T}\right) - \mu T. \tag{3.14}$$

Acknowledgments

This work has been supported by the Faculty of Biology at Utrecht University.

Part II

Evolution of *Saccharomyces cerevisiae* After its Whole Genome Duplication.

4

The Role of Mutational Dynamics in Genome Shrinkage

M.J.A. van Hoek and P. Hogeweg
Theoretical Biology/Bioinformatics Group, Utrecht University
Padualaan 8, 3584 CH Utrecht, The Netherlands.

Mol. Biol. Evol. **24(11)**: 2485-94 (2007 Nov)

Abstract

Genome shrinkage occurs after whole genome duplications (WGDs) and in the evolution of parasitic or symbiotic species. The dynamics of this process, whether it occurs by single gene deletions or also by larger deletions are however unknown. In yeast, genome shrinkage has occurred after a WGD. Using a computational model of genome evolution, we show that in a random genome single gene deletions cannot explain the observed pattern of gene loss in yeast. The distribution of genes deleted per event can be very well described by a geometric distribution, with a mean of 1.1 genes per event. In terms of deletions of a stretch of base pairs, we find that a geometric distribution with an average of 500-600 base pairs per event describes the data very well. Moreover, in the model, as in the data, gene pairs that have a small intergenic distance are more likely to be both deleted. This proves that simultaneous deletion of multiple genes causes the observed pattern of gene deletions, rather than deletion of functionally clustered genes by selection. Furthermore, we found that in the bacterium *Buchnera aphidicola* larger deletions than in yeast are necessary to explain the clustering of deleted genes. We show that the excess clustering of deleted genes in *B. aphidicola* can be explained by the clustering of genes in operons. Therefore, we show that selection has little effect on the clustering of deleted genes after the WGD in yeast, while it has during genome shrinkage in *B. aphidicola*.

4.1 Introduction

Massive gene loss has been a major evolutionary driving force in the evolution of many different species. After a whole genome duplication (WGD), most duplicate genes are lost and only a relatively small percentage of genes is kept in duplicate. WGD has been shown in the ancestry of yeast (Wolfe & Shields, 1997; Kellis *et al.*, 2004), plants (e.g., in *Arabidopsis thaliana* up to 3 WGDs can be distinguished, (Initiative, 2000; Bowers *et al.*, 2003)), vertebrates (Dehal & Boore, 2005) (although this is still debated), the teleost fishes (Amores *et al.*, 1998; Taylor *et al.*, 2001), and the ciliate *Paramecium tetraurelia* (Aury *et al.*, 2006). Massive gene loss occurred in all these lineages.

Genome shrinkage has first been studied in *mycoplasma*, prokaryotic parasites that lack a cell wall and have very small genomes. Contrary to initial expectations, these bacteria evolved from free-living bacteria with larger genomes (Maniloff, 1983; Woese, 1987). Because these bacteria are intracellular parasites, many metabolic functions can be performed by their hosts, which renders many genes nonfunctional. This caused the dramatic shrinkage of their genomes. This process of genome shrinkage also occurred in, for example, *Buchnera aphidicola*, an endosymbiont of aphids (Moran & Mira, 2001).

An open question concerning the mutational dynamics of genome shrinkage is whether genes are lost one by one, via point mutations and subsequent pseudogenization, or that multiple neighboring genes are lost in one event, for example, via unequal crossing over. Here we study whether the pattern of gene deletions, as observed in different yeast species and *B. aphidicola* can be explained by single gene deletions or whether simultaneous deletion of neighboring genes is needed.

When large stretches of subsequent deleted genes are observed in an alignment between genomes, from which one experienced massive gene loss, this could be due to simultaneous deletion of several neighboring genes or to subsequent deletions of single genes. We constructed a computational model describing genome shrinkage. We use different probability distributions of deletion sizes (in number of genes) to simulate genome shrinkage. We find that nor in yeast, nor in *B. aphidicola*, single gene deletions in a nonstructured genome can explain the observed pattern of gene deletions. We find that when we allow for deletions of multiple adjacent genes, we can explain the pattern of gene loss satisfactorily.

An important issue in genome shrinkage is whether selection, neutral evolution, and/or the mutational dynamics determine which and how many genes are deleted. For endosymbionts, it has been proposed that gene deletion is much more frequent than gene duplication and that this causes the reduction of genome size in these bacteria (Mira *et al.*, 2001) and selection does not play a large role. However, another study reveals that selection is also important (Delmotte *et al.*, 2006). Gene evolution and genome reduction after WGD (in particular which genes are kept in duplicate and which are deleted) are most often explained by selection (Blomme *et al.*, 2006; Lin *et al.*, 2006; Thomas *et al.*, 2006). In any case, it seems obvious that genome shrinkage is not a completely neutral process because some genes will be more important to retain (in duplicate) than others. Here we

study whether the pattern in gene deletions both in yeast and *B. aphidicola* can be explained by the mutational mechanism alone or that selection is needed to explain this pattern.

Because gene order is not random, selection could cause clustering of gene deletions. This nonrandom gene order is very well known for prokaryotes, in which genes are clustered in operons. Also for eukaryotes it has been shown that genes are functionally clustered (Hurst *et al.*, 2004). To distinguish between these two explanations, we constructed a second model for genome shrinkage. In this model, deletions are on the basis of base pairs, so instead of genes, a stretch of base pairs is deleted. We find that we can explain the pattern of gene loss satisfactorily when large deletions (in the order of 100-1000 bp) are frequent enough, both for yeast and *B. aphidicola*. This model predicts that genes which have a short intergenic distance are more likely to be simultaneously deleted. Furthermore, small genes are more likely to be deleted simultaneously with other genes than large genes. We show that both these predictions hold in yeast and therefore find strong evidence that deletion of stretches of base pairs underlie gene deletions in yeast.

We checked this result by studying gaps of base pairs in pseudogenes in *Saccharomyces cerevisiae*, in a similar way as has previously been done for *B. aphidicola* (Gomez-Valero *et al.*, 2007). We used the distribution of gap sizes (in base pairs) that we observed in these pseudogenes in our model. We find that, using this distribution, we can satisfactorily explain the amount of clustering of deleted genes in *S. cerevisiae*. This all shows that the clustering of gene deletions after the WGD in *S. cerevisiae* is caused by the mutational dynamics and not by the clustering of functionally related genes.

For *B. aphidicola*, we observed that on average longer deletions in terms of base pairs were needed ($\approx$1000 vs 500-600, which corresponds to on average 1.6 gene vs. 1.1 gene in *S. cerevisiae*). We found that these longer gaps are most likely due to the operon structure that is present in prokaryotes, which causes deleted genes to be clustered in the genome.

4.2 Materials and Methods

4.2.1 Data

We obtained the distribution of gap sizes (in genes) for a number of species. For the yeast species, we used the Yeast Gene Order Browser (YGOB), version 1.0. (Byrne & Wolfe, 2005). This is an alignment between 3 yeast species that underwent a WGD (*S. cerevisiae*, *Saccharomyces castellii*, and *Candida glabrata*) and 4 that did not (*Ashbya gossypii*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Saccharomyces kluyveri*).

The alignment in YGOB is local. The data in the YGOB consists of "pillars". A pillar represents one ancestral gene, which is duplicated in the WGD. For every post-WGD species, 0, 1, or 2 duplicates of this gene may be conserved. Every pre-WGD species may or may not have this gene. Therefore, a pillar consists of

10 slots that may or may not be filled. If only one slot in a pillar is filled, we discarded that pillar because then the most parsimonious explanation is that this gene is a recent duplication instead of already present at the WGD.

We downloaded the alignment in a region of 50 pillars around every gene of every pre-WGD species. As mentioned by Byrne & Wolfe (2006), focusing on pre-WGD species is the best way to view the syntenic context. We looked which post-WGD genes were in the same pillar as the in-focus pre-WGD genes. Then we counted the gap size left and right of these post-WGD genes. In this way of course gaps are counted twice, and we divided the resulting distribution by 2.

Focusing on different pre-WGD species sometimes gives a different result (as is noted by Byrne & Wolfe (2005)). Therefore, we focused on every pre-WGD species (except *S. kluyveri* because it is only sequenced to $4\times$ coverage and it is clear from YGOB that many genes are missing) and averaged the outcome. However, the 3 resulting distributions were very similar, which confirms the consistency of our results.

B. aphidicola is an endosymbiont of aphids. It is a relative of the Enterobacteriaceae, like *Escherichia coli*. *B. aphidicola*'s genome is practically a subset of the genome of *E. coli* (Moran & Mira, 2001). It has 550 genes, whereas *E. coli* has 4488 genes. It is believed that *B. aphidicola* lost all these genes after becoming an endosymbiont. Delmotte *et al.* (2006) constructed an alignment of the genome of *B. aphidicola* with respect to the reconstructed last common ancestor of endosymbionts and their free-living relatives. They also measured gap sizes of deleted genes, which we use here.

Lengths of genes and intergenic regions are downloaded from National Center for Biotechnology Information, except for the gene lengths and intergenic lengths of *A. gossypii*, which we downloaded from the "Ashbya Genome Database" (Hermida *et al.*, 2005). For finding gaps in pseudogenes, we used a previously published list of identified pseudogenes in Lafontaine *et al.* (2004). As was done in this article, we aligned the pseudogenes with their homologous open reading frame (ORF) using DIALIGN 2.2.1 (Morgenstern, 2004) and subsequently we counted the gaps in the alignment.

4.2.2 Model

In all species we studied, we observed that gap sizes of deleted genes are not geometrically distributed, which would be the case if gene loss occurred due to single gene deletions in a random way. To study how the pattern of gap sizes is influenced by the sizes of allowed deletions, we constructed a computational model to simulate massive gene loss. Only gene deletion is taken into account. A genome is represented by a list of genes. We impose a certain probability distribution of deletion sizes. We can, for example, only allow for single gene deletions or also for larger deletions (of more genes) using a geometrical distribution. A simulation is terminated when the number of deleted genes equals the number of deleted genes in the data set we are interested in. The number of genes a simulation starts with is simply the number of conserved genes plus the observed number of deleted genes in the data set.

When reconstructing genome evolution after a WGD, we take both duplicates of the genome into account. Thus, the starting genome then consists of 2 homologous "genomes". We assume selection against deletion of both genes of an ohnolog pair (an ohnolog is a paralog that arose through a WGD). To account for the fact that in yeast, in a relatively small number of cases both ohnologs are deleted, we assume a certain probability r that a certain ohnolog pair can be deleted, such that the actual number of double deletions corresponds to the actually observed number. For *S. cerevisiae*, we used $r = 0.12$, for *S. castellii* and *C. glabrata* $r = 0.17$.

At the end of a simulation, we count the distribution of gap sizes in the evolved genome. We do this 1000 times and take the average of these 1000 simulations. We now treat this average distribution as a theoretical distribution, to which we can compare the observed distribution. We use a χ^2 test to determine the probability that a certain computationally obtained distribution can explain the observed distribution.

We also constructed a more detailed model of genome shrinkage. In this model, we delete base pairs, instead of genes. The genome now consists of genes and intergenic regions of certain lengths. In the case of yeast, the length of the genes and intergenic regions are randomly drawn from the observed genic and intergenic length distributions in the pre-WGD species *K. lactis*, whereas for *B. aphidicola*, we use lengths of *E. coli*, a close relative in which genome shrinkage did not occur.

This model works identically as the previously explained model. Randomly a base pair is picked from the genome and a deletion length from a certain distribution and, if possible, the deletion is carried out. If a gene has lost some base pairs due to a deletion, we assume that gene becomes nonfunctional and is degraded instantly and all base pairs from that gene are deleted. If, however, a part of an intergenic region is deleted, the rest of the intergenic region stays intact. This model is consistent with the observation that very few pseudogenes belonging to deletions after the WGD in *S. cerevisiae* can still be identified (Lafontaine *et al.*, 2004). We also tried a rule in which a gene, of which a certain part is deleted, becomes a pseudogene (intergenic region). This model gave almost identical results to the instant gene degradation model.

4.3 Results

4.3.1 Yeast

First we look at the distribution of gap sizes in *S. cerevisiae*. The observed distribution is shown by the gray histogram in Fig. 4.1 *A*. First we simulated genome shrinkage when only single gene deletions are allowed. This is done first with deletions occurring on the gene level. We then find the distribution indicated by the solid line. This distribution is simply a geometric distribution $f(n) = P(1-P)^n$, where P equals the probability that a gene is conserved (the number of conserved genes divided by the total number of conserved genes plus the total number of

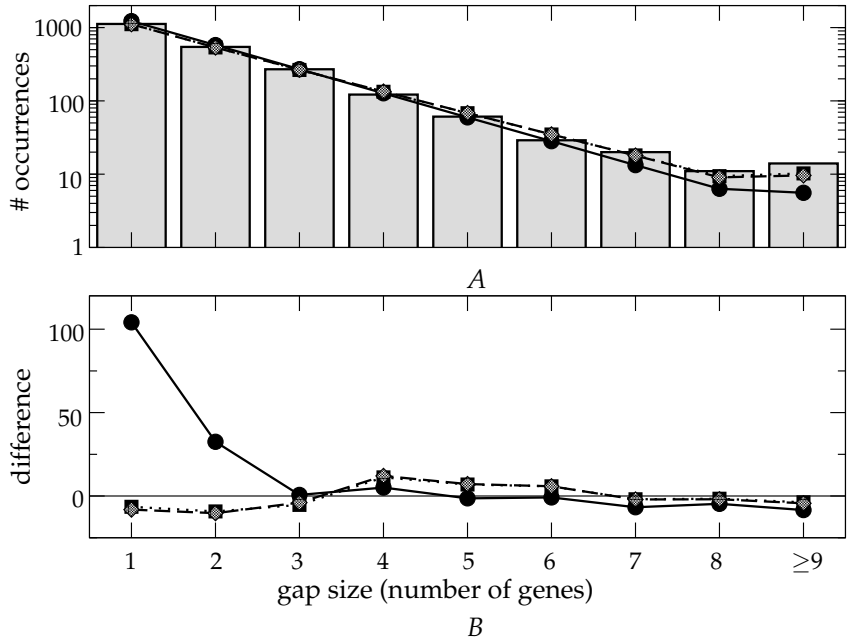


Figure 4.1: The observed and computationally obtained gap size distributions for *Saccharomyces cerevisiae*. (*A*) Gray bars: the observed distribution of gap sizes in YGOB. Solid line, circles: computationally obtained distribution, if only single gene deletions are allowed. Dashed line, squares: computationally obtained distribution, deletions are gene based. Deletion sizes are geometrically distributed ($q = 0.92$). Dotted line, diamonds: computationally obtained distribution when deletions occur on the basis of base pairs. Base pair deletions are geometrically distributed ($q = 0.0019$). (*B*) The difference between the computationally obtained and the observed distribution.

deleted genes) and n the observed gap size.

From Fig. 4.1 *A*, it is clear that there are too many large gaps in the observed distribution to be described by single gene deletions alone. In Fig. 4.1 *B*, the difference between the observed and the computationally obtained distribution is shown. From this picture, it is clear that there are also too few small gaps. If we compare the distribution caused by single gene deletion and the experimentally observed distribution, we find that $\chi^2 = 31$, which gives a P value of $P = 0.0001$. Note that in the last bin in Fig. 4.1, we added all gaps larger than 9 genes. Also when calculating χ^2, this binning was used, to avoid very small numbers of observations per category, which is necessary for a χ^2 test. The exact location for this cut-off, however, does not essentially change the results.

Allowing for larger deletions may explain the data better. A natural assumption for a distribution of deletions is a geometric distribution. If a deletion event has a fixed probability to terminate after every gene, the resulting distribution of

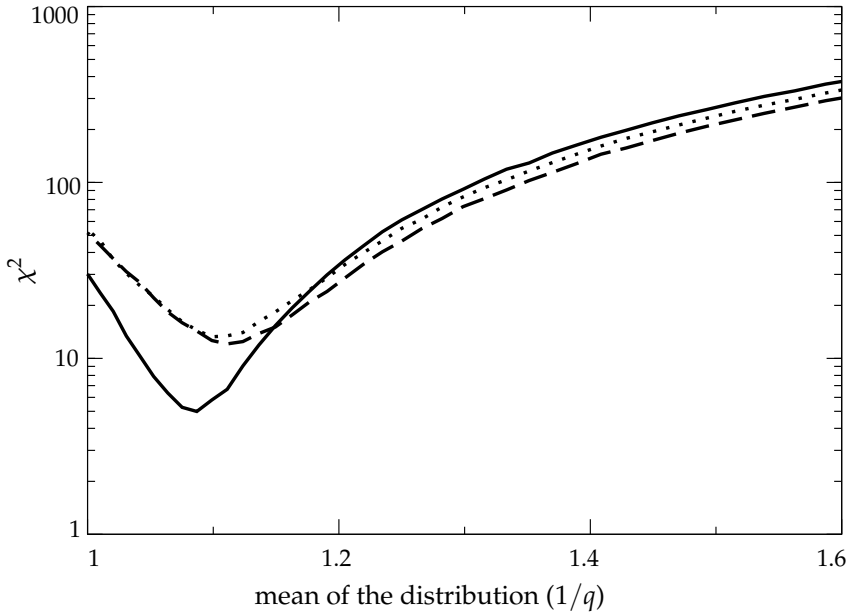


Figure 4.2: χ^2 as a function of the mean of the distribution (mean $= 1/q$). Solid line, *Saccharomyces cerevisiae*; dashed line, *Candida glabrata*; and dotted line, *Saccharomyces castellii*.

sizes of deletion events is geometric. Note the important difference between the distribution of deletion events and the distribution of observed gap sizes. Assuming only single gene deletion leads to a geometric distribution of observed gap sizes. Assuming a geometric distribution of deletion events leads to a different distribution of observed gap sizes, due to clumping of deletion events into larger gaps.

So we assume that sizes of deletion events are geometrically distributed with parameter q. This parameter indicates the probability that a deletion event is a single gene deletion. We can try to fit the observed distribution by varying q and calculating χ^2 between the observed and the simulated distribution. In Fig. 4.2, we show how χ^2 depends on the mean of the distribution used (which is equal to $1/q$). For the interpretation of the results, it is crucial that χ^2 has a single, definite minimum when changing q, which is indeed the case. The best P value is found for $q = 0.92$. Indeed this distribution describes the data very well, with $\chi^2 = 4.8$, $P = 0.78$ (dashed line in Fig. 4.1).

To check in a reverse way whether the single gene deletion model can describe the data, we fitted all 1000 distributions, obtained from the simulations using only single gene deletions, against q, thus treating these as observed distributions. In 0.7% of the cases, the best-fit value of q was 0.95, in all other cases, the resulting q was higher. A q value of 0.92 was therefore never found, and it is therefore very

Table 4.1: Statistics of Fit for All Yeast Species

	q=1		Geometric Gene Deletion			Geometric bp Deletion		
Species	χ^2	P value	q	χ^2	P value	q	χ^2	P value
S. cerevisiae	31	0.0001	0.92	4.8	0.78	0.0019	5.8	0.67
S. castellii	56	<0.0001	0.90	13	0.11	0.0017	16	0.05
C. glabrata	58	<0.0001	0.90	12	0.15	0.0016	14	0.09

improbable that such a mechanism can lead to a distribution which is best fitted by $q = 0.92$.

For the other 2 post-WGD yeast species available, *S. castellii* and *C. glabrata*, we find very similar results. The results of all 3 post-WGD yeast species are summarized in Table 4.1.

Interestingly, we find almost equal q values for the different species, which indicates the consistency of the result. However, the fits are always best for *S. cerevisiae*, maybe due to the better annotation of genes in this species.

Deletion of Base Pairs

From the previous section, we conclude that gene deletions in *S. cerevisiae* are clustered within the genome. However, from this model, we cannot conclude that the mutational dynamics are responsible for this clustering. The results found in the previous section could just as well reflect the fact that genes are functionally clustered in the genome of *S. cerevisiae* (Hurst *et al.*, 2004). When a certain gene is deleted, the chance of neighboring genes to be deleted might increase and this effect would lead to the clustering of deleted genes. When this explanation would be correct, selection would cause the clustering of deleted genes, instead of the mutational dynamics.

Therefore, we constructed a model in which the mutational dynamics are incorporated more realistically. Because mutations obviously act on the level of base pairs instead of on the level of genes, we modified the gene deletion model to a base pair deletion model. In this model, adjacent genes can be deleted simultaneously due to a deletion of a long stretch of base pairs. In this way, we hope to find out whether the mutational dynamics or the clustering of functionally related genes cause the observed clustering of deleted genes.

We tested whether this model, using base pair deletions, could also cause the observed pattern in gap sizes. Again we assume a geometric distribution of base pair deletions, corresponding to a fixed probability of termination of a deletion event after each base pair.

Again we fitted the resulting distribution of gap sizes (still in genes) to the observed distribution, by varying q, the probability that a deletion has a length of only one base pair and therewith the mean length of the deletions. We found that for *S. cerevisiae*, a mean length of 530 bp fits the data best. The obtained

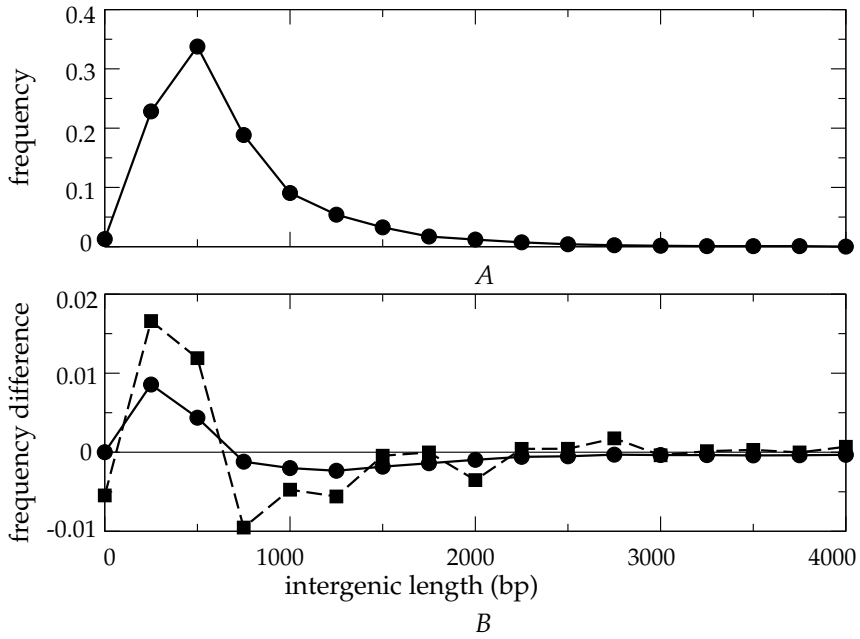


Figure 4.3: Intergenic length between genes of *Kluyveromyces lactis*. (*A*) Frequency distribution of all intergenic regions of *K. lactis* in YGOB. (*B*) Dashed line, squares: difference between the distribution of intergenic lengths between 2 genes in *K. lactis*, which are both deleted in a post- WGD species and the distribution of all intergenic regions (the distribution in Fig. 4.3 *A*). Solid line, circles: the resulting curve of the simulations.

distribution is also shown in Fig. 4.1. For *S. castellii* and *C. glabrata*, we found a mean deletion length of 600 and 620 bp, respectively.

As expected, a geometric distribution of base pair deletion also describes the data nicely. This gives us a mechanistic explanation for the observed gap sizes. More importantly, this model also provides predictions about which genes are more likely to be deleted than others. Therefore, we can check whether these predictions hold in the data and in this way prove whether or not the above mechanism is responsible for the large gaps. For example, neighboring genes that have a short intergenic region in between are more likely to be both deleted than genes that have long intergenic regions in between because a deletion has to reach the next gene in order to delete both genes simultaneously.

In Fig. 4.3 *A*, the frequency distribution of intergenic regions in *K. lactis* is shown. We are interested whether small intergenic regions are overrepresented if both neighboring genes are deleted after the WGD. Therefore, we calculated the difference between the frequency distribution for intergenic regions for which both neighboring genes are deleted and the frequency distribution of all intergenic regions. We did this for each post-WGD species and averaged the 3 curves.

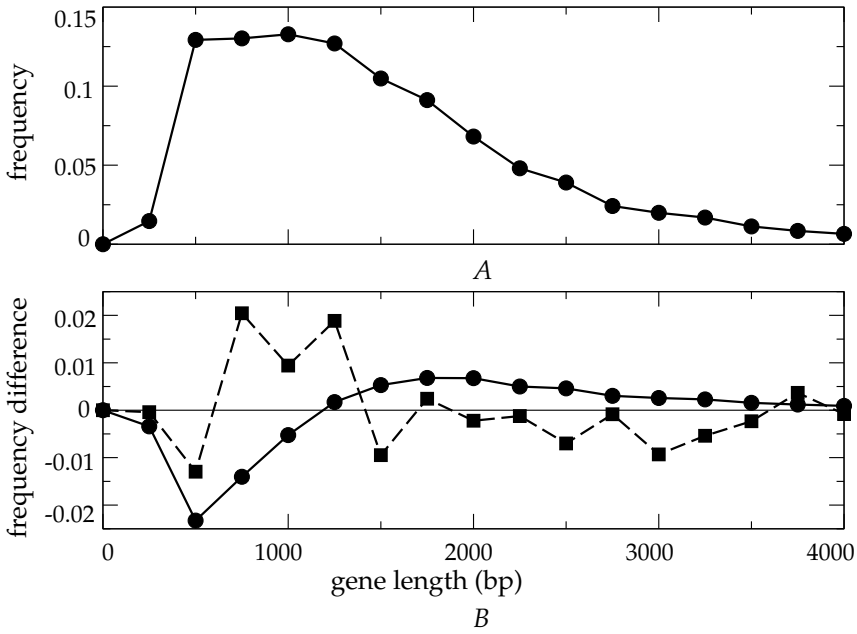


Figure 4.4: Length of genes in *Kluyveromyces lactis*. (*A*) Frequency distribution of gene lengths in *K. lactis*. (*B*) Dashed line, squares: the difference between the gene length of genes that are deleted in gaps of size one and genes that are deleted in gaps of a larger size. Solid line, circles: the resulting curve of the simulations.

This gives the dashed curve in Fig. 4.3 *B*. This curve indicates the over or underrepresentation of genes of a certain length if both neighboring genes are deleted. For the simulations, we performed exactly the same procedure, except that we averaged over 1000 simulations, for each post-WGD species, which gives the solid curve in Fig. 4.3 *B*.

We find that the simulations predict the observed outcome surprisingly well. Both the overrepresentation of small intergenic regions (<500 bp) and the underrepresentation of intergenic regions of intermediate length are captured well by the model.

A second prediction of the model is that when we compare genes in gaps that consist of only one gene with genes in gaps that consist of more genes, small genes will be relatively overrepresented in large gaps compared with gaps consisting of only one gene. This is because when a large gene is picked to be deleted, the probability is higher that the whole base pair deletion will not reach the end of the gene. Therefore, large genes will be relatively more often deleted on their own.

In the simulations, we clearly observe this phenomenon (see Fig. 4.4 *B*). These curves are made in a similar way as the curves in Fig. 4.3 *B*. Small genes are underrepresented in the set of genes in gaps of length one, whereas genes of

intermediate length are overrepresented. Although the data are very noisy, the underrepresentation of small genes (<500 kbp) is precisely in correspondence with the model. The genes that are overrepresented in gaps of size one in the data are, however, smaller than we expect from the model (see below).

For both these observations to be meaningful, it is necessary that intergenic length and gene length are conserved between pre-WGD and post-WGD species. We checked this, and we indeed found good correlation for genic and intergenic length between *S. cerevisiae* and *K. lactis* (for intergenic length, we found a correlation coefficient of 0.65, and for gene length, a correlation coefficient of 0.98).

Influence of Gene Length on Deletion Probability

Inspired by the above result, we studied the effect of gene size on the probability to be deleted after a WGD in more detail. Previously, it has been found that genes which are duplicated in the human genome are on average shorter (Nembaware *et al.*, 2002). The authors believe that this is likely a mutational effect.

In the same article, it is also mentioned that in yeast, this relationship between gene length and the probability of being kept as duplicate after a WGD was not found. On the contrary, the length of genes belonging to an ohnolog pair was even a bit larger than the average gene length.

We more precisely studied the relationship between gene length and being kept in duplicate after a WGD (see Fig. 4.5). We divided all genes of *S. cerevisiae* in 3 groups, genes that are kept in duplicate after WGD, genes that are conserved on their own after WGD, and recent duplications (genes that have no homolog in a pre-WGD species). On average, ohnologs are slightly larger. Furthermore, it is very clear that recent duplications are on average shorter, just as was found for humans (Nembaware *et al.*, 2002). Therefore, we find that there is a correlation between gene length and the probability of being kept as a duplicate after a WGD, but it is reversed with respect to single gene duplications. These results also hold for genes of *C. glabrata*.

The fact that ohnologs are longer than on average indicates that shorter genes might be more often deleted after a WGD than longer genes. To check whether this is correct, we looked at the length of pre-WGD species, instead of post-WGD. For each gene in *K. lactis*, we counted how many homologs, from 0 to 6, are present in *S. cerevisiae*, *S. castellii*, and *C. glabrata*. Fig. 4.6 clearly shows that genes that are often deleted after the WGD are on average smaller than genes that are seldom deleted. For *A. gossypii*, we found a very similar pattern.

All this confirms that indeed, as found by Nembaware *et al.* (2002), short genes are more often duplicated, but in contrast, after a WGD, where all genes are duplicated, small genes are also more often lost. Note that the use of a multiple alignment allows us to distinguish between gene deletions and duplications.

Although this may appear contradictory, it is very reasonable that genes that are more often duplicated also need to be more often deleted. Otherwise, the genome would gradually evolve to contain smaller and smaller genes. Therefore, a possible explanation is that because gene duplications are mutationally more likely for shorter genes, deletions of short genes are favored by selection. In con-

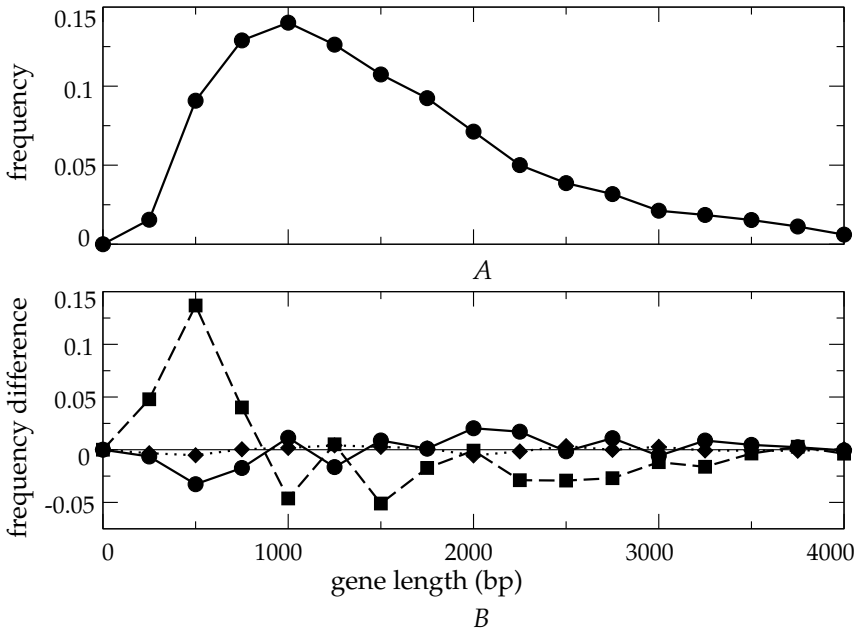


Figure 4.5: Length of genes in *Saccharomyces cerevisiae*. (*A*) Frequency distribution of gene lengths of all genes of *S. cerevisiae* in YGOB. (*B*) Solid line, circles: difference in gene length distribution between genes that are conserved both (ohnologs) and the gene length distribution of all genes (Fig. 4.5*A*). Dotted line, diamonds: idem for genes that are conserved once. Dashed line, squares: recent duplications.

trast, in our neutral model, we observe that longer genes are more often deleted because the base pair where a deletion starts is randomly picked from the whole genome. Therefore, we believe that selection causes short genes to be more often deleted.

In Fig. 4.4, we observed that in the data, the strongest overrepresentation in gene length in the data coincides with the maximum in the length distribution (the curve in Fig. 4.4 *A*). Because long genes in our model have a larger probability to be deleted, the maximum in the model prediction (solid line in Fig. 4.4 *B*) is shifted to longer gene length, which explains the inconsistency between the model prediction and the data.

Gaps in Pseudogenes of *Saccharomyces cerevisiae*

We found that in our model, base pair deletions of 100-1000 bp are responsible for the clustering of deleted genes in *S. cerevisiae*. We wondered whether we could find evidence that deletions of this size occur frequently enough during evolution of *S. cerevisiae*. Inspired by Gomez-Valero *et al.* (2007), we looked at pseudogenes

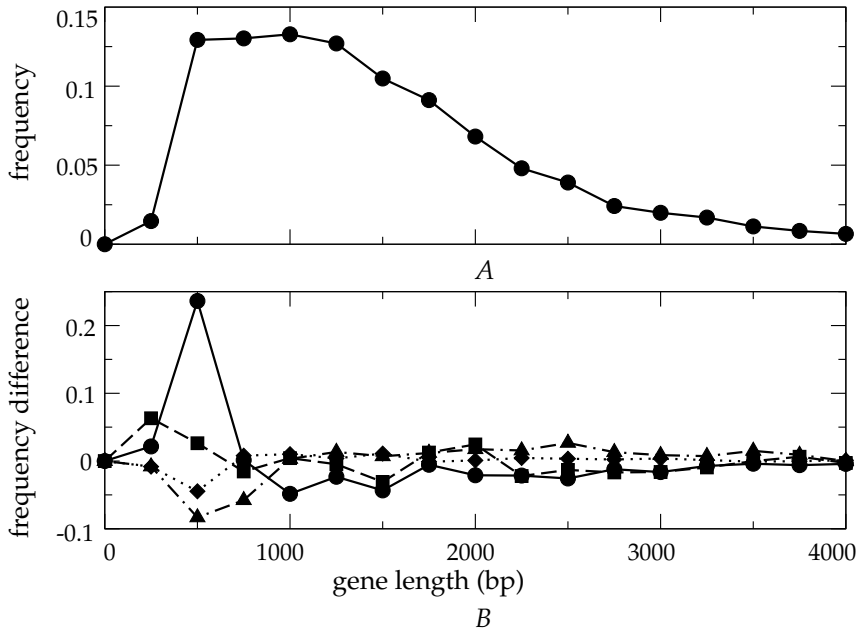


Figure 4.6: Conserved genes in post-WGD species are on average longer in *Kluyveromyces lactis*. (*A*) Frequency distribution of gene lengths of all genes of *K. lactis* in YGOB. (*B*) Solid line, circles: difference in gene length distribution between genes with 0 homologs in the post-WGD species and all genes in YGOB (Fig. 4.6*A*). Dashed line, squares: idem for genes with 1 or 2 homologs in the post-WGD species. Dotted line, diamonds: idem for genes with 3 or 4 homologs in the post-WGD species. Dot-dashed line, triangles: idem for genes with 5 or 6 homologs in the post-WGD species.

in *S. cerevisiae*. We aligned pseudogenes in *S. cerevisiae* with their homologous ORFs, which are identified in Lafontaine *et al.* (2004) and counted the gaps in the alignments. In 230 relics, we observed 744 gaps. Also gaps at the beginning or the end of a pseudogene are counted. However, these gaps will in reality have been larger and therefore some gaps are underestimated in size. In Fig. 4.7, the distribution of gap sizes is shown.

We observe that the distribution of gap sizes is not geometric. It looks more like a power-law distribution, although it is statistically different. We can see from Fig. 4.7 that gaps with sizes as large as 500 bp do occur, although not very often.We wondered whether we could explain the clustering of deleted genes in *S. cerevisiae*, using precisely this distribution of base pair deletions, without any fitting. The result is shown in Fig 4.8.

Indeed, the computational obtained distribution fits the data remarkably well, considered there is no fitting involved. We find a χ^2 value of 9.69, which corresponds to a P value of 0.29. We only overestimate the number of single gene

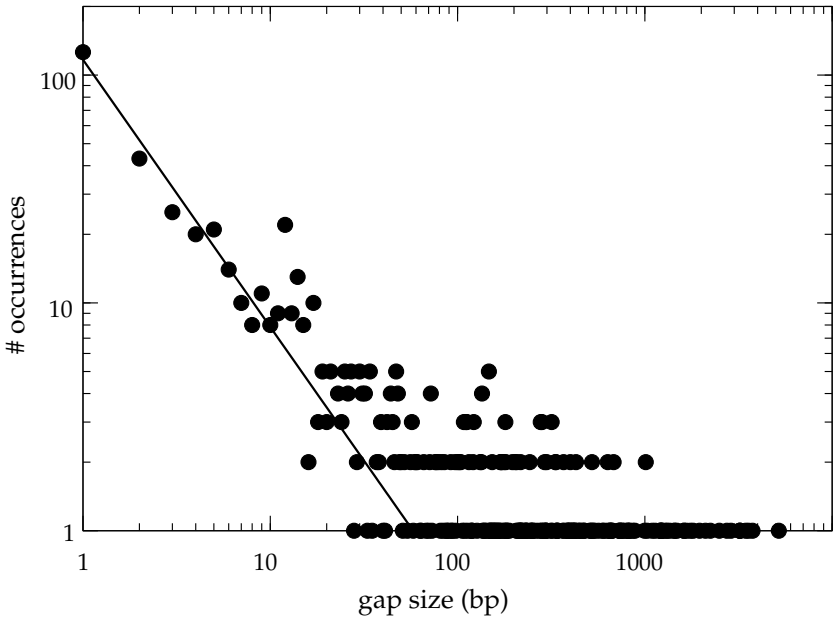


Figure 4.7: Distribution of gap sizes as observed in intergenic relics in *Saccharomyces cerevisiae*. The black lines give a power-law fit, using a maximum likelihood estimation (Goldstein *et al.*, 2004).

deletions, although considerably less than when only single gene deletions are allowed. It must be noted, however, that the distribution of gap sizes as shown in Fig. 4.7 strongly depends on the alignment algorithm used. We also tried ClustalW (Thompson *et al.*, 1994), using the default settings, and we obtained a different distribution, with less large gaps. The fit with the data was also less good than when using DIALIGN 2.2.1 (Morgenstern, 2004). However, it remains that base pair deletions of 100-1000 bp are relatively frequent and can therefore be responsible for simultaneous deletion of neighboring genes. We also performed all other simulations in the previous section using the distribution from Fig. 4.7. Quantitatively there were differences, but qualitatively the results remained unchanged.

4.3.2 *Buchnera aphidicola*

We also used this approach to study gene loss in *B. aphidicola*. The computational models we used are the same as the ones used to study the evolution of the yeast genome, except that every gene is only present once, instead of twice (WGD did not occur in *B. aphidicola*).

Again we find that single gene deletions cannot explain the data (see Fig. 4.9), as was previously noted (Delmotte *et al.*, 2006). However, when deletion sizes (gene based) are geometrically distributed, the fit again becomes very good for

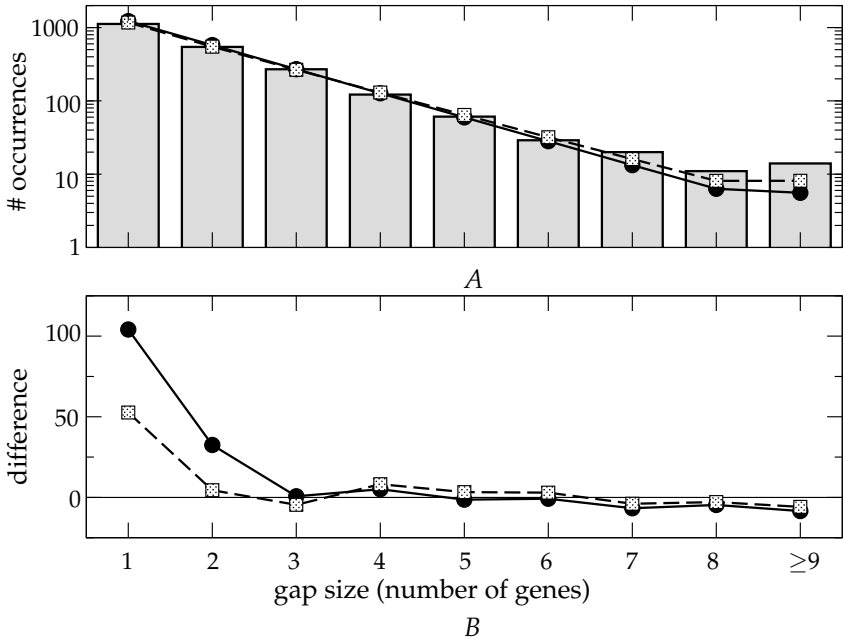


Figure 4.8: The computationally obtained gap size distributions for *Saccharomyces cerevisiae*, when we use the observed gap distribution in relics to delete base pairs. (*A*) Gray bars: the observed distribution of gap sizes in YGOB. Solid line, circles: computationally obtained distribution using only single gene deletions (from Fig. 4.1 *A*). Dashed line, squares: computationally obtained distribution when base pair deletions are according to Fig. 4.7. (*B*) The difference between the computationally obtained and the observed distribution.

an average deletion length of 1.6 gene, which is much larger than for yeast (1.1 gene).

We used gene lengths and intergenic lengths from *E. coli* as initial condition for the base pair based simulations. Both the genes and the intergenic regions of *E. coli* are much shorter than for yeast. This, however, only partly explains the larger gaps found in *B. aphidicola*. The mean deletion length that fits the data best for *B. aphidicola* is 1010 bp. Again, the fit is very well, even better than for the gene-based simulations. Unfortunately, we cannot study the influence of intergenic length and gene length on the deletion probability for *B. aphidicola*, as we did for yeast. For this, we would need an exact alignment between *E. coli* and *B. aphidicola*, which is not available.

Operon Structure

The average number of genes deleted per event is higher for *B. aphidicola* than for *S. cerevisiae*. Correspondingly, on the level of base pairs, larger deletions are

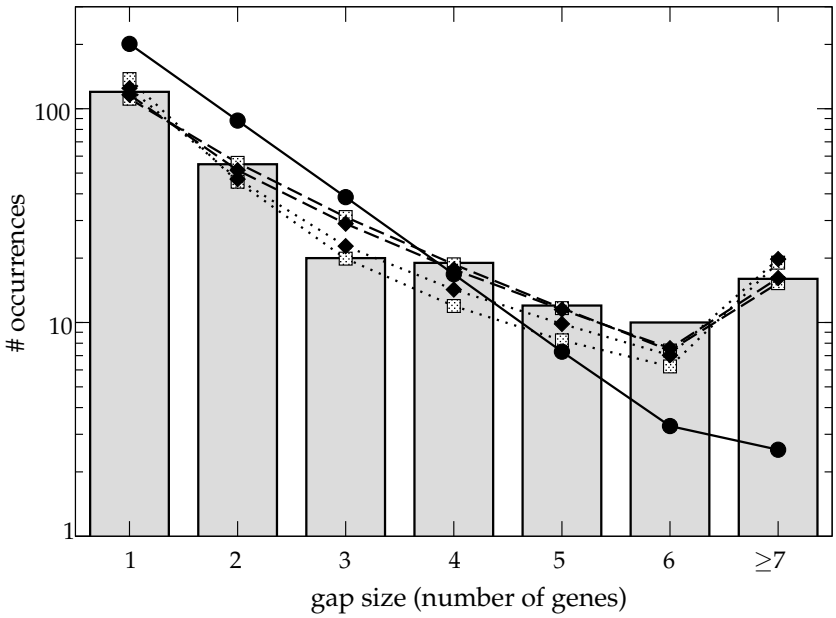


Figure 4.9: Histograms of observed and computationally obtained gap sizes for *Buchnera aphidicola*. Gray bars: observed distribution of gap sizes (Delmotte *et al.*, 2006). Solid line, circles: computationally obtained distribution, if only single gene deletions are allowed ($\chi^2 = 142$, $P < 0.0001$). Dashed line, squares: computationally obtained distribution, deletions are gene based, deletion sizes are geometrically distributed ($q = 0.64$, $\chi^2 = 5.6$, $P = 0.47$). Dashed line, diamonds: computationally obtained distribution, deletions are base pair based and geometrically distributed ($q = 0.0010$, $\chi^2 = 4.0$, $P = 0.67$). Dotted line, squares: computationally obtained distribution, assuming operon structure and a probability of deleting operons as a whole ($q = 0.5$, $\chi^2 = 13$, $P = 0.04$). Dotted line, diamonds: computationally obtained distribution, assuming operon structure and a probability of deleting all genes downstream in that operon ($q = 0.80$, $\chi^2 = 6.0$, $P = 0.43$).

required to fit the data (≈1000 vs. 500-600 bp). Gap sizes in pseudogenes have also been measured for bacteria, in *Rickettsia* (Andersson & Andersson, 1999a,b) and *B. aphidicola* (Gomez-Valero *et al.*, 2007). As we found for yeast, many small gaps (<10 bp) and some large gaps (up to 1500 bp) were observed. However, too few gaps were observed for us to use as a distribution in our model. It, however, appears that the gaps in pseudogenes are somewhat smaller than those we found in *S. cerevisiae* (Andersson & Andersson, 1999b). This is surprising because it has been argued that these larger gaps in *B. aphidicola* are because *B. aphidicola* lacks recA, a DNA repair protein (Gomez-Valero *et al.*, 2007). It appears, therefore, that larger base pair deletions are not a plausible explanation for the difference we observe.

However, it is known that in bacteria, the genome is organized in operons. Therefore, the clustering of functional genes is much more pronounced in bacteria than in eukaryotes. We wondered whether the clustering of functional genes in bacteria could explain the larger gaps we observed. It is very well conceivable that operons are deleted more or less as a whole, instead of gene by gene. When a certain gene from an operon is deleted, the pathway in which this gene functions may become nonfunctional and the deletion of the rest of the operon becomes very likely. This is somewhat similar to the "Domino Theory" of gene death (Dagan *et al.*, 2006), which claims that genome reduction in endosymbionts is gradual at first, but when a crucial gene renders a pathway nonfunctional, the entire pathway will be deleted very fast.

To study whether this effect can also explain the observed large gaps, we counted how many operons of a certain length are known in *E. coli*. These data are available from the RegulonDB (version 5.5) database (Salgado *et al.*, 2006). Also "operons" consisting of only one gene are considered. We now assume that operons in our starting genome have the same length distribution (in genes) as in the operon database. We use our gene-based model and furthermore assume that if a certain gene is deleted, there is a certain probability q that the whole operon is deleted. If this is not so, then we assume that this operon will also not be deleted as a whole by following gene deletions in that operon. Furthermore, the deletion will then always be a single gene deletion.

This gives us, for $q = 0.5$, the best fit. This distribution fits the data reasonably well, much better than the single gene assumption, but also less well than the geometric deletion size assumption (both gene based and base pair based).

If we, however, assume that, when a gene is deleted from an operon, only the genes downstream are deleted, the fit becomes as good as the (gene based) geometric distribution (using $q = 0.80$). Here we assume that the orientation of the operons in the genome is random but fixed. This scenario assumes that genes are ordered in the operon and that sometimes only a part of the operon will be nonfunctional if one gene is deleted. So we conclude that the amount of genome organization in operons suffices to explain the excess of large gaps in *B. aphidicola*.

4.4 Discussion

We have shown that single gene deletions in a nonstructured genome cannot explain the pattern of gene deletions in yeast, nor in *B. aphidicola*, because larger gaps are observed than can be explained by this random model. There are 2 possible explanations for these larger gaps. The first is that the mutational dynamics are responsible for the large gaps. The second possible explanation is selection: if functional genes are clustered in the genome, this can cause clustering of gene deletions.

Genome organization by operons is well known for bacteria, but for eukaryotes, it has been assumed that gene order is random. In recent years, it has become clear that this assumption is false (Hurst *et al.*, 2004). Operons have been identified in some eukaryotes (mostly in *Caenorhabditis elegans* (Blumenthal,

2004)). Furthermore, coexpressed genes in the cell cycle (Cho *et al.*, 1998) or stress-related genes (Burhans *et al.*, 2006) are shown to be clustered in the genome of yeast. The amount of gene clustering in eukaryotes is, however, not well known, although it is clearly less than in prokaryotes and it is hard to predict whether it is sufficient to explain for the amount of clustering of deletions that is observed.

For yeast, we found convincing evidence that the mutational dynamics, instead of clustering of single gene deletions, are responsible for the observed large gaps. Firstly, we showed that genes which have a short intergenic region in between are more likely to be both deleted. However, having a short neighboring intergenic region in itself does not increase the probability for a gene to be deleted, which excludes the possibility that this phenomenon is caused by selection. Secondly, small genes are more often found in large gaps. Thirdly, a prediction of our model is that when 2 neighboring pre-WGD genes are deleted, the probability is higher that they are deleted on the same chromosome. We find that this is also the case in the data. For *S. cerevisiae*, *S. castellii*, and *C. glabrata*, we find that, respectively, 53.0%, 53.5%, and 54.3% of the cases genes are deleted in parallel. In our models, both the gene based and base pair based, we find very similar percentages. The base pair-based deletion model gives 53.9%, 55.0%, and 55.2%, respectively, whereas the gene-based model gives 53.7%, 54.8%, and 54.5%. Finally, we confirmed that the size of base pair deletions that we need to fit the data compares well with the size of base pair deletions we observed in pseudogenes in *S. cerevisiae*.

Because short genes have a higher probability to be deleted after WGD, a natural explanation for the occurrence of large gaps would then be that short genes are clustered in the genome. We checked this, but only very limited clustering of short genes can be observed in *K. lactis* and the amount of clustering does not appear to be sufficient to account for the amount of large gaps.

In *A. thaliana*, it has been observed that after a WGD, genes were preferentially lost from one chromosome (Thomas *et al.*, 2006). This phenomenon has been called biased fractionation. Biased fractionation could also result in large gap sizes because in the chromosome from which most genes are deleted, the gaps will become larger than expected. We checked whether biased fractionation occurs in yeast, but in yeast, genes are lost equally from both chromosomes and biased fractionation therefore can also not explain the large gap sizes.

The model we use for yeast is almost entirely neutral. The only selection we implemented is that duplicate genes have only a small probability to be both deleted. Therefore, we conclude that the observed pattern in gene loss in yeast is caused by the mutational dynamics and not by selection. Indeed, it has been shown that the selective constraints on duplicate genes is very modest shortly after the duplication (Lynch & Conery, 2000) and selection increases approximately 10-fold later on. Given that most of the duplicates are lost very shortly after WGD (Scannell *et al.*, 2006), it is to be expected that most genes are lost in a relatively neutral way. In later stages of genome reduction, we expect that selection plays a larger role. The fact that we find that in yeast selection does not cause the clustering of deleted genes does not mean, however, that selection is not important during genome shrinkage after a WGD, but more probably that the amount

of clustering in functionally related genes in *S. cerevisiae* is very moderate.

In bacteria, the situation is quite different. We found that deleted genes are much more clustered in *B. aphidicola* than in yeast. We propose that, in contrast to yeast, clustering of genes in operons, and hence selection, explains the large gap sizes in *B. aphidicola*. However, also in *B. aphidicola*, we expect that the mutational dynamics partially cause the large gaps because large base pair deletions are also observed in bacteria.

Interestingly, the gap size distribution of base pairs in pseudogenes we found for *S. cerevisiae* (see Fig. 4.7) was very similar to the distribution already found for *Rickettsia* (Andersson & Andersson, 1999a,b). There it was found that 35% of the gaps were of one base pair, whereas 6% of the gaps had a size larger than 500 bp (Andersson & Andersson, 1999a,b). We found in *S. cerevisiae* that 17% of the gaps were single base pair gaps, whereas 10% of the gaps had a size larger than 500 bp. Such a distribution, where most gaps are very short and some are very long, is indicative of a power-law distribution. We assumed a geometric distribution of base pair deletions. Which distribution is used in the model is, however, not crucial for the outcome. What is crucial is that most deletions are only a few base pairs long (and hence cause single gene deletions) and a few are large (causing deletion of multiple genes).

Some studies concerning very recent deletions, using experimental evolution or different natural isolates of a certain species, observed much larger gaps than studies that focused on the comparison of different species. For example in Ochman & Jones (2000), different strains of *E. coli* were compared and many large gaps ($>$ 10 kbp) were observed. A similar study for *Mycobacterium tuberculosis* (Kato-Maeda *et al.*, 2001) also found such large gaps. In Nilsson *et al.* (2005), such large gaps were observed after experimental evolution of *Salmonella enterica* of mutants that are defective in mismatch repair. Finally, in *S. cerevisiae*, chromosomal deletions are not uncommon (Dunham *et al.*, 2002). All this might indicate that these large deletions do occur but are only selectively advantageous in specific environments, and over a longer evolutionary time scale, these deletions are lost by purifying selection.

Previously, it was observed (Scannell *et al.*, 2006) that, when one ohnolog of a pair is lost, then most often the same gene is lost in different lineages, even if corrected for the fact that many of these cases are because that gene is already lost in the ancestor. This is called convergent gene loss. Functional divergence of genes after WGD is mentioned as a possible explanation for convergent gene loss. We hypothesized that, when deletions would span several genes, another explanation might also play a role. During evolution after WGD, 2 ohnologs will get different neighboring genes. If a certain gene becomes a neighbor of 2 essential genes in one chromosome, but not in the other, this gene cannot be deleted by a gene deletion of length 2 or larger, whereas its ohnolog can. Therefore, its ohnolog will have a higher probability to be deleted. However, because the frequency of single gene deletions is very large in yeast ($q = 0.92$), deletions of multiple neighboring genes cannot explain the observed amount of convergent gene loss.

Our results for *S. cerevisiae* are in contrast with the findings of Byrnes *et al.* (2006), who study the pattern of gene deletions in the alignment between *S. ce-*

revisiae and *A. gossypii* (which is found in Dietrich *et al.* (2004). They compare whether a single gene deletion model or models with larger deletions (a uniform distribution with maximum size 2 or a Poisson distribution with mean 1 or 2) can best describe the data. It was found that the single gene deletion describes the data best ($P = 0.11$).

We find a much lower P value for the single gene deletion model. This is because we found a different gap size distribution, which is based on the alignment of 7 yeast species which each other in the YGOB instead of one. In this way, genes that were deleted in one of the species will be present in others, and this will give a better alignment. This is particularly important because *A. gossypii* has the least annotated protein-coding genes of all the yeast species we take into account (Byrne & Wolfe, 2005).

In addition, instead of using distributions with arbitrarily chosen parameters, we fitted q and therewith the mean of the distribution to the data. Therefore, we were able to find that distributions with a mean size of approximately 1.1 gene fit the data best.

There are more species that have undergone a WGD and from which the pattern of gene deletions is available. For *A. thaliana*, it has been shown that its genome is shaped by 3 WGDs (Initiative, 2000; Bowers *et al.*, 2003). The histograms of gap sizes are given in Thomas *et al.* (2006). *Paramecium tetraurelia* has also undergone several WGDs, and the pattern of gene deletions is available from the supplementary materials in Aury *et al.* (2006). For these 2 species, there is, however, no alignment available with a pre-WGD ancestor. Only an alignment between both chromosomes is available, but from this we cannot infer the gap sizes.

Very recently, an alignment between *Tetraodon nigroviridis*, *Danio rerio*, and several post-WGD species was used to study WGD in these teleost fishes (Semon & Wolfe, 2007). Unfortunately, this alignment was not given in the article. Moreover, this alignment is not made for the whole genome, because, unlike in yeast, extensive rearrangements have frequently caused loss of synteny. Therefore, we did not use this method for the teleost fishes.

Whether genome reduction is driven by selection or by neutral evolution is a much debated question. Noncoding DNA (like introns, transposons, etc.) is much more abundant in eukaryotes than in prokaryotes. It has been proposed that, by changing the amount of nuclear DNA, the cell size is changed, which has selective power (Cavalier-Smith, 1978, 2005). A totally opposite hypothesis states that noncoding DNA is slightly deleterious, because of the potential mutational burden, but only species with high effective population sizes and high per nucleotide mutation rates are capable of keeping their genome devoid of this noncoding DNA (Lynch, 2006; Lynch *et al.*, 2006).

In this study, however, we look at loss of genes, instead of loss of noncoding DNA. In *B. aphidicola*, we find evidence that selection is an important factor in determining which genes are lost. Although genome size reduction in *B. aphidicola* might very well be caused by a lack of selection pressure, due to the small effective population size (Mira *et al.*, 2001; Moran & Mira, 2001), this does not mean that selection is not important in determining which genes are lost. In *S. cerevisiae*, we find strong evidence that selection does not influence the pattern of

gene deletions after its WGD. However, we do not believe that this means that selection is not important during genome shrinkage in yeast.

In summary, we show that gene loss not only occurs in a gradual, gene by gene manner, but that larger base pair deletions can cause simultaneous loss of several neighboring genes. Furthermore, we show that this mechanism is responsible for clustering of deleted genes in *S. cerevisiae*. In *B. aphidicola*, however, we argue that the excess amount of large gaps is due to the clustering of functional genes in operons.

Acknowledgments

We thank A. Crombach for his help with data retrieval and B. Snel for giving an inspiring seminar, which led us to do this research. This work has been supported by the Faculty of Biology at Utrecht University.

5

Evolutionary Modeling of the Metabolic Network of Yeast After its Whole Genome Duplication

M.J.A. van Hoek and P. Hogeweg
Theoretical Biology/Bioinformatics Group, Utrecht University
Padualaan 8, 3584 CH Utrecht, The Netherlands.
In Preparation

Abstract

In several evolutionary lineages there is strong evidence for a whole genome duplication (WGD). WGDs have been hypothesized to be responsible for major transitions in evolution. However, the effect of a WGD on cellular behavior and metabolism are still poorly understood. Here we study, from a metabolic modeling perspective, the effect of a WGD on the metabolic capacities of a cell and in this way on its fitness. Conveniently, there exists an organism for which a WGD has been convincingly proved and for which the metabolism is very well known. In the yeast lineage a WGD has occurred approximately 100 million years ago. We here study the effect of such a WGD and the subsequent process of massive gene loss on *Saccharomyces cerevisiae*, using a previously published genome-scale metabolic model. We develop a model in which it is possible to describe the effect of a WGD on the metabolism of a cell. This model is based on Flux Balance Analysis (FBA) and it assumes that the maximal flux through a reaction increases if the corresponding gene is duplicated (dosage dependence). It has been proposed that, because the WGD in the yeast lineage occurred around the same time as the appearance of angiosperms, this WGD played a role in the adaptation of yeast to glucose-rich fruits. We find that our model captures some essential elements of the WGD in yeast. Genes that are more often retained in duplicate in the model are also more often retained in duplicate after the WGD in yeast. We also find that, both in the model and in the data, glycolysis and transporter genes are more likely to be retained in duplicate, which leads to an increase in maximal glycolytic flux after WGD. Furthermore we find that a WGD can, in environments to which the cell is not perfectly adapted, infer an immediate fitness advantage. These findings confirm the hypothesis that the WGD has been important in the adaptation of yeast to the new glucose-rich environment that arose after the appearance of angiosperms.

5.1 Introduction

The occurrence of whole genome duplications (WGD) was already proposed by Ohno (1970). Nowadays, there is conclusive evidence for WGDs in several lineages, such as yeast (Wolfe & Shields, 1997; Kellis *et al.*, 2004), plants (Initiative, 2000; Bowers *et al.*, 2003) and teleost fishes (Amores *et al.*, 1998; Taylor *et al.*, 2001; Semon & Wolfe, 2007). It has been speculated that these WGDs caused major transitions in evolution (Huerta-Cepas *et al.*, 2007), but as yet it is unknown what the precise effects of a WGD are on cellular behavior.

One would expect that cellular behavior, and specifically the metabolism, will change after a WGD. For example, it is known that cell size increases with DNA content (Cavalier-Smith, 1978). An increased cell size would alter the surface area to volume ratio of a cell, which could have serious consequences on the metabolic capacities of a cell. After a WGD it is common that most duplicate genes are lost and only a minority of the genes remain in duplicate (for example approximately 500 gene pairs in *Saccharomyces cerevisiae*). In this way, if we assume a gene-dosage effect, the relative abundance of gene products may alter.

In this paper we study the effects of a WGD using a modeling approach. To study how a WGD changes the metabolic capacities of a cell, we need a model of the whole metabolic network of an organism that underwent WGD. For *S. cerevisiae*, which also underwent a WGD, such a model is already available. Therefore, we use *S. cerevisiae* as a model system to study the possible effects of WGD and subsequent gene loss on the metabolism of a cell.

Very recently it has been proposed that the WGD that occurred in yeast led to an increase in glycolytic flux (Conant & Wolfe, 2007). Using a kinetic model of glycolysis and assuming a dosage effect for duplicated genes, it was shown that these duplicated genes could indeed enhance glycolytic flux.

Furthermore, using a comparative genomics approach, it has been established that in general yeast species that underwent WGD are Crabtree positive, which means that they produce ethanol in the presence of oxygen, while most species that did not undergo WGD are Crabtree negative (Merico *et al.*, 2007). Therefore it appears that the WGD caused yeast to be able to rapidly consume glucose.

Here we use a fully compartmentalized, genome-scale metabolic model (Duarte *et al.*, 2004a), which is publicly available. Because obviously no such model is available for the pre-WGD ancestor of yeast, we, strictly speaking, study the effect that a WGD would have if it would occur now in *S. cerevisiae*. In this way we hope to get a better understanding of how a WGD can change a metabolic network.

The fluxes through such a metabolic network can be calculated using flux balance analysis (FBA). FBA is a constraint-based modeling approach which, instead of modeling the dynamics of a metabolic network, focuses on determining the steady state fluxes. This approach is necessary, because in a genome-scale model it is very difficult to assess the kinetic parameters of all reactions. Using stoichiometric (and possibly other) constraints, FBA tries to find a flux distribution that optimizes for a certain objective function, for example the growth rate of the cell. Optimizing for the growth rate of a cell has successfully been used to

reproduce the growth rates for gene deletion studies (Edwards & Palsson, 2000; Famili *et al.*, 2003) and by-product secretion of cells (Famili *et al.*, 2003; Duarte *et al.*, 2004b).

We develop a model, based on FBA, in which we can study the effect of WGD and subsequent gene loss on the metabolism of a cell. This model assumes that the cell size is correlated with genome size. Indeed, haploid yeast cells are smaller than diploid cells, which are again smaller than tetraploid cells (Hennaut *et al.*, 1970). Furthermore, even between organisms such a correlation is also observed (Cavalier-Smith, 1978). Therefore, we assume that after a WGD the cell size increases and gradually decreases when genes are lost. This changes the surface area to volume ratio, which is an important factor determining the metabolic capacities of a cell (see for example Kooijman *et al.* (1991)). Furthermore we assume dosage dependence in our model, such that reactions for which the genes are duplicated can reach higher fluxes.

We study the effect of a WGD on the whole metabolic network, using evolutionary modeling. By performing evolutionary simulations, we study the WGD in yeast and the subsequent loss of genes and we ask under which circumstances the WGD is initially adaptive and under which circumstances it can eventually lead to a higher growth rate.

We find that our model can describe the essential features of the WGD in *S. cerevisiae*. The WGD is followed by massive gene loss, during which the relative abundance of genes change. Our model satisfactorily predicts which genes are retained in duplicate after WGD: genes that are never (sometimes) retained in duplicate in our simulations have a low (high) probability to have been retained in duplicate after the WGD in yeast. We also find that genes coding for transport and glycolysis reactions have a higher probability to be retained in duplicate, both in our model as in the data, which leads to an increase in glycolytic flux as has been proposed previously (Conant & Wolfe, 2007). Finally, we find that a WGD can lead to an instant fitness increase in environments for which the cell was not perfectly adapted initially, in line with the idea that the WGD in yeast helped to adapt to the newly arisen environment of glucose-rich fruits.

5.2 Materials and Methods

To study the evolution of yeast after a WGD we used a previously developed genome-scale model of yeast metabolism (Duarte *et al.*, 2004a). Here we will first shortly describe this model. It consists of 1061 metabolites and 1266 reactions. The reversibility of each reaction is also given in the model. Therefore, the model consists of a 1266×1061 stoichiometric matrix.

All reactions are linked with genes of *S. cerevisiae*. Some reactions have no corresponding genes, either because there is no gene found that catalyzes this reaction, or because it does not need a protein to be catalyzed. Other reactions have more than one corresponding gene. For simplicity, it has been assumed that genes work together in a Boolean way. For example, reaction X is performed by GENE A OR GENE B, reaction Y is performed by (GENE C AND GENE D) OR

(GENE C AND GENE E), etcetera.

We now use this model as the basis for an evolutionary model that can describe the effect of a WGD and subsequent gene loss on the metabolic network. For this we use a flux balance approach. FBA maximizes the growth rate of the cell, given the stoichiometric constraints and possible constraints on the flux values of each reaction. Our approach to model the WGD is to change the constraints on the network and in this way the growth rate, rather than changing the network itself. During evolution, some genes will be kept in duplicate, some as singleton and some will be entirely deleted. We assume that the maximal flux value of a certain reaction is determined by the number of genes present to catalyze that particular reaction. Therefore we assume dosage dependence, such that the maximal flux through each reaction depends on the copy number of the associated genes. When genes are deleted during evolution, these maximal fluxes change and in this way the constraints on the cells metabolism change and cells can adapt to the environment.

We used the Matlab COBRA Toolbox (Becker *et al.*, 2007) to perform FBA and all evolutionary simulations were performed using Matlab.

5.2.1 A Model to Determine Changes in Flux Constraints.

To model the evolution of *S. cerevisiae* after a WGD we need a formalism to describe the changes in flux constraints during evolution. Here we develop a model that describes the changes in flux constraints of each reaction, dependent on the copy number of all genes, the cell volume and surface area. We assume dosage dependence, gene differentiation is excluded in this model.

We distinguish two type of reactions, exchange reactions and intracellular reactions. First we explain how we model the changes in constraints of these intracellular reactions. We assume that, after a WGD, the cell size changes. This appears to be a natural assumption, given the good correlation between DNA content and cell size that is ubiquitous in nature (Cavalier-Smith, 1978; Gregory, 2001; Cavalier-Smith, 2005). Furthermore, it has been observed that polyploid yeast cells are larger than haploid yeast cells. Hennaut *et al.* (1970) observed that a diploid cell is 1.87 as large as a haploid cell. Furthermore, the surface area increased with a factor 1.56. Using these data, we now assume for our model that $V \sim N^{0.9}$ and $A \sim V^{0.7}$, where V is the volume of a cell, N the number of genes and A the surface area. We also assume that cells cannot become smaller than 1, i.e. not smaller than before WGD. Accordingly, after a WGD the volume increases by a factor 1.87. For intracellular reactions that are not associated with a gene, we assume that the maximum flux (measured in mmol/(gram Dry Weight hour)) remains constant.

For intracellular reactions that *are* associated with a gene we take the following approach. When a reaction is associated with several genes, the model gives us, as mentioned above, whether this occurs in an "AND" or "OR" like fashion. If reaction X can be performed by multiple genes, in an "OR"-like fashion, we assume that every gene contributes equally to the maximum flux. If reaction X can be performed by N genes and one of them is deleted, the maximum flux decreases

by a factor $(N-1)/N$. If a certain reaction Y is performed by several genes in an "AND"-like fashion, all these genes are necessary to perform this reaction and if one is lost the maximum flux becomes zero.

For more complicated situations, it is helpful to define the concept "reaction multiplicity", $m(i)$, for each reaction i, as the total number of proteins or protein complexes that can perform a certain reaction. For reaction X this would be N before gene deletion and $N-1$ after gene deletion. For reaction Y it would be 1 before gene deletion and 0 after. A somewhat more complicated example: assume that reaction Z is associated with three genes in the following way: (gene A AND gene B) OR (gene A AND gene C). m_i for this reaction would be 2 before WGD and 4 after WGD. If after WGD one of the duplicates of gene B would be deleted, $m(i)$ would decrease from 4 to 3, while if one of the duplicates of gene A would be deleted, $m(i)$ would decrease from 4 to 2. Using this concept it is now easy to calculate the maximal flux through a reaction as:

$$F_{max}(i) = F_{max,0}(i)\frac{m(i)V_0}{m_0(i)V} \equiv F_{max,0}(i)\frac{m(i)}{m_0(i)\beta} \tag{5.1}$$

Here the index 0 means before WGD and we defined $\beta \equiv V/V_0$.

For exchange reactions, the situation is somewhat more complicated. It is often assumed that the surface area to volume ratio is a crucial factor that determines the uptake of nutrients. Indeed, this factor can be important, but another factor is the amount of permease proteins that are available. It has been shown previously that in yeast both the surface area to volume ratio and the amount of permease proteins can be the rate limiting factor for the amount of nutrient uptake (Hennaut *et al.*, 1970). It was shown that some uptake reactions are not (or hardly) different between haploid and diploid yeast cells, while other reactions (almost) change by the surface area to volume ratio between haploid and diploid cells. The authors conclude that some uptake reactions are limited by the surface area of the cell and others by the amount of permease protein.

We would like to have a general formula that describes both these situations, the saturation for surface area and number of proteins. Suppose a cell has a surface area A and a volume V. Furthermore we assume that there are N permease proteins on the cell surface. Every permease protein can transport molecules that hit the surface of the cell around that particular protein, lets say within an area A_p of that protein. The total area that is than covered by permease proteins can be described by the following formula

$$A = \frac{AN}{N + A/A_p} \tag{5.2}$$

Indeed, the limit for very small N is correct, the area then increases linearly with the area covered by one permease protein and if N goes to infinity the whole area of the cell is covered. Now we can express the exchange flux for metabolite i as a function of the metabolite concentration, before the WGD as

$$F_{max,0}(i) = \frac{1}{V}\frac{rY}{Y+K}\frac{AN(i)}{N(i)+A/A_p} = \frac{1}{V}\frac{rY}{Y+K}\frac{Ax(i)}{1+x(i)} \tag{5.3}$$

Here r is the influx rate per unit surface area and Y the metabolite concentration and we introduced $x(i) \equiv \frac{N(i)A_p(i)}{A}$ as the initial area of all permease proteins divided by the initial cell surface area. Note again that flux is measured in mmol/(gram Dry Weight hour), hence the factor $1/V$. When $x(i) \gg 1$, the flux is proportional to the surface area to volume ratio, while when $x(i) \ll 1$ the flux is proportional to the protein concentration. We are now interested in how the flux changes if A, V and N change by a factor α, β and γ respectively (for example after a WGD). The ratio between the flux after and before such a change is given by the following formula

$$\begin{aligned}\frac{F_{max,end}(i)}{F_{max,0}(i)} &= \frac{\frac{1}{\beta V}\frac{rY}{Y+K}\frac{\alpha A\gamma(i)N(i)}{\gamma(i)N(i)+\alpha A/A_p(i)}}{\frac{1}{V}\frac{rY}{Y+K}\frac{A\gamma(i)N(i)}{N(i)+A/A_p(i)}} = \frac{\alpha\gamma(i)}{\beta}\frac{N(i)+A/A_p(i)}{\gamma(i)N(i)+\alpha A/A_p(i)} \equiv \\ &\frac{\alpha\gamma(i)}{\beta}\frac{1+x(i)}{\gamma(i)x(i)+\alpha} \equiv \frac{\alpha m(i)}{\beta m_0(i)}\frac{1+x(i)}{\frac{m(i)}{m_0(i)}x(i)+\alpha}\end{aligned} \tag{5.4}$$

where, in the last part, we again introduced the reaction multiplicity, because $\gamma(i) = \frac{m(i)}{m_0(i)}$.

This formula can be compared to Eq. 5.1, which describes how intracellular fluxes change after WGD. Indeed, the limit for $x(i)$ goes to zero gives Eq. 5.1. In that case the uptake rate does not depend on the surface area of the cell. Finally, uptake and excretion fluxes that are not associated with genes (such as oxygen uptake) are modeled using the limit for x to infinity of Eq. 5.4, such that these fluxes are only determined by the surface area to volume ratio.

Unfortunately, it is not possible to know the $x(i)$ values for all exchange reactions that occur in yeast. We choose $x(i) = 1$ for all reactions, such that both the surface area to volume ratio and the amount of permease proteins are important in determining the exchange fluxes.

5.2.2 Initialization

In the previous section we described how the flux constraints change during the course of evolution. However, we also need a way to establish the initial constraints on the fluxes. As there is no experimental data for all reactions, we approach this problem from a different point of view. We assume that the initial maximal flux for each reaction is determined by evolution. Genes that correspond to reactions that need high fluxes, will have evolved high transcription rates and vice versa.

To determine these maximal fluxes we performed FBA 1000 times in different environments (the environments in which we performed the evolutionary simulations, which we will explain in the next section) and calculated the fluxes through each reaction. To avoid futile, thermodynamically unrealistic fluxes, we minimized the sum of the absolute value of all fluxes after we maximized for the growth rate. A similar approach was followed by Kuepfer *et al.* (2005) and Bilu

et al. (2006). As maximal flux value for a certain reaction we now took the maximum of the 1000 obtained flux values. If this value was smaller than a certain cut-off, we put the maximum flux value to this cut-off. For reversible reactions, maximum fluxes for both directions are taken into account separately.

5.2.3 Different Environments

Papp *et al.* (2004) showed that an important reason that so many enzymes are not essential for viability in *S. cerevisiae*, is that some enzymes are essential in environments different from laboratory conditions. This was done using a genome-scale model of yeast (Forster *et al.*, 2003), the precursor of the model we use (Duarte *et al.*, 2004a). They tested enzyme essentiality in 9 different environments that might have been important for the evolution of yeast. We now use 9 environments, similar to the ones used by Papp *et al.* (2004). These environments were also used in the initialization procedure. We used one rich medium (environment 1, taken from Bilu *et al.* (2006)) and 8 minimal media (2: minimal glucose, low O_2, 3: minimal glucose, anaerobic, 4: minimal ethanol, low O_2, 5: minimal acetate, low O_2, 6: minimal glucose, carbon limited, 7: minimal glucose, nitrogen limited, anaerobic, 8: minimal glucose, phosphate limited, anaerobic, 9: minimal glucose, sulfate limited, anaerobic). A precise description of these media is given in the Supplementary Material.

5.2.4 Evolutionary Algorithm

At $t = 0$ we start with a (constant) population size of 100 cells that have a duplicated genome. Randomly, each time step one of the nine environments is chosen and the growth rate of each cell is calculated in that environment, using FBA. We continue the simulations for 100000 time steps, at which time the population has gone to an evolutionary steady state.

According to the growth rates of the cells we decide which cells are allowed to reproduce and which not. After every time step we allow 50 reproductions of cells. For every reproduction, the probability that we pick cell j is calculated using a Boltzmann distribution: $P_j = \frac{\exp(\mu_j/T)}{\sum_i \exp(\mu_i/T)}$. We used a "temperature" of $T = 0.001$, such that growth rate differences of 0.001 mmol/(gram Dry Weight hour) still are selected on.

Cells that have a growth rate of zero are removed from the population. When a cell reproduces, it has a probability of 10% to mutate. The mutational procedure will be explained in the next section. After reproduction the population size is restored to 100 by random death and we continue to the next time step. Our approach is comparable to the OptGene algorithm developed in Patil *et al.* (2005), who designed a genetic algorithm to find gene deletions that lead to a certain flux distribution.

5.2.5 Mutations

If a cell reproduces, it has a probability of 10% to mutate. After WGD, we only allow for gene deletion. Every gene can therefore be present 0, 1 or 2 times. When a mutation occurs, the constraints for every reaction are recalculated as described in section 5.2.1. Previously we have shown that after WGD in *S. cerevisiae*, genes are not only lost one by one, but there is also a (relatively small) chance that neighboring genes are deleted simultaneously (see **chapter 4**). It was found that a geometric distribution describes the data well. Therefore, we model the probability that n genes are deleted simultaneously as $P(n) = q(1-q)^{(n-1)}$, using $q = 0.9$.

5.3 Results

5.3.1 Is Gene Retention After WGD in Yeast Correlated With Flux?

A major assumption in our evolutionary model is that the number of gene copies determines the maximal flux through each reaction. To check this assumption, we studied whether we could find a correlation between genes that have been retained in duplicate after WGD and the flux through the network. For this, we do not use the evolutionary model described in the previous section, but simply the metabolic model of Duarte *et al.* (2004a) on which our evolutionary model is based.

We tested whether in this model, without any constraints, reactions that are associated with genes that are retained in duplicate after the WGD in yeast, have higher fluxes than on average. To do this, we performed FBA using the original model and subsequently minimized the total flux through the network, like we did when determining the initial constraints on the network (see section 5.2.2). We did this for three environments: the anaerobic environment with high glucose concentration, the glucose-limited aerobic environment with high glucose concentration and the glucose-limited aerobic environment with low glucose concentration. These environments correspond with the environment 2 and 6 as explained in the Supplementary Material, using a maximal glucose uptake rate of 0.1 or 3 mmol/(gram Dry Weight hour).

We calculated the average flux through all reactions that are associated with genes that were duplicated during the WGD and compared this number with the average flux of reactions associated with an equal number of randomly chosen genes. By calculating the average flux for random genes 10000 times, we constructed a P value that indicates the significance of whether the average flux for reactions that are associated with duplicated genes is higher than expected by chance. These P values are given in Table 5.1, for all three environments. Indeed we observe that reactions that are associated with genes that are retained in duplicate after the WGD have significantly higher fluxes than on average, in both high glucose environments.

Table 5.1: *P*-values that average flux is higher than expected randomly

	aerobic high glucose	aerobic low glucose	anaerobic high glucose
WGD	0.032	0.088	0.0018
overexpressed (Ferea *et al.*, 1999)	<0.0001	0.0001	0.29
underexpressed (Ferea *et al.*, 1999)	0.88	0.97	<0.0001
duplicated (Dunham *et al.*, 2002)	0.57	0.51	0.68

This indicates that there is indeed selection for keeping reactions with high fluxes duplicated. Or alternatively, that the WGD increased the flux through certain reactions by duplicating them. In any case, this observation confirms our hypothesis that gene dosage is an important issue in deciding whether genes are kept in duplicate after WGD. Note also that the signal is stronger in the anaerobic environment, in correspondence with the hypothesis that the WGD in yeast was important in establishing the fermentative lifestyle of yeast. Furthermore, in the environment with low glucose concentrations, the result is not significant, indicating that the WGD helped to adapt to a glucose-rich environment.

Inspired by this result, we checked whether selection for high dosage could also be observed in other evolutionary circumstances. In Ferea *et al.* (1999), yeast cells were evolved in a low glucose, aerobic environment for a few hundred generations. After evolution, it was observed that the expression of many genes had changed compared to the ancestor, when growing in this environment. From their data we selected genes that were overexpressed and underexpressed after evolution and again calculated the average flux through reactions that are associated with these genes. Indeed we found that genes that were overexpressed in this experiment had higher flux in both aerobic environments, but not in the anaerobic environment. Conversely, genes that were underexpressed had higher flux in the anaerobic environment. This shows that in the Ferea-data, reactions that initially already have a high flux have a larger chance to become overexpressed and will therefore increase their flux even more. Again this indicates that dosage effects are important during evolution and that the model can capture this effect well.

Finally we looked at data published by Dunham *et al.* (2002). In similar experiments as those by Ferea *et al.* (1999), it was observed that gross chromosomal rearrangements had occurred. We also studied whether genes that were duplicated in these chromosomal duplications had a higher flux than expected by chance. We found that this was not the case. Apparently, although such a duplication could be selective because several important genes are duplicated, the duplication as a whole does not make much sense. The reason for this could be that there is too little clustering of functional genes in yeast (see **chapter 4**) and therefore many genes that are not important in the given environment will also

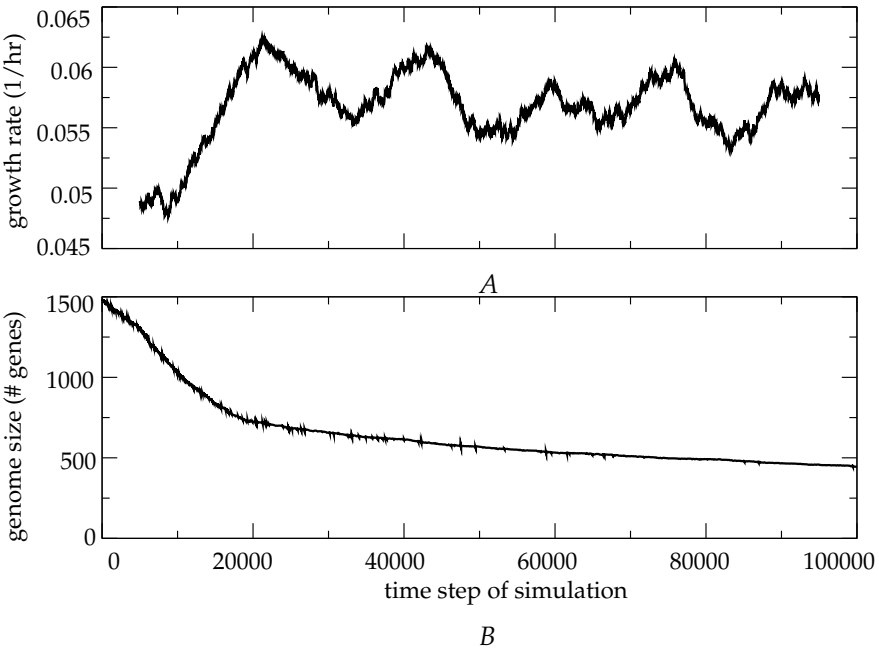


Figure 5.1: (*A*) Running average of the growth rate. (*B*) Genome size.

be duplicated. Only after selection has deleted all unnecessary genes, like after WGD, a signal can be observed. We conclude that dosage selection is important in determining both expression level of a gene and whether a gene will remain duplicated after WGD.

5.3.2 Gene Loss After WGD: Evolutionary Simulations

In this section we discuss the results of the evolutionary model that is described in the Materials and Methods. As explained there, we evolve a population of 100 cells in a changing environment, after WGD. Because this is a small population size, we expect much evolutionary drift. Therefore, we performed 10 simulations using a different random seed. Then we performed competition experiments, where 10 individual cells of each of the 10 simulations (yielding 100 cells) were competed against each other. The mutation rate was set to zero in these simulations. Here we report the findings of the evolutionary simulation yielding the population that most often won the competition experiments. In **part I** we used a similar method to cope with evolutionary drift.

In our model, a fitness increase can be gained by decreasing the genome size and therewith the cell size. When the cell size decreases by the deletion of a gene, the maximal fluxes of all other reactions increase, especially of exchange

reactions, which increase by the surface area to volume ratio (see Eq. 5.1 and 5.4). This leads to a selection pressure for a small genome. Therefore, we find that the WGD is followed by massive gene loss, as indeed has been observed for all WGDs (see for example Wolfe & Shields (1997); Aury *et al.* (2006)).

In Fig. 5.1 both the genome size and average growth rate of the population are shown. Initially, the model consists of 743 genes (and 7 mitochondrial genes). We observe that the genome sizes indeed decreases even under 743 and stabilizes around approximately 450 genes. This is because many genes in the model are dispensable, as has already been studied elsewhere (Papp *et al.*, 2004). The average growth rate initially increases and later stays more or less constant. We observe quite some variation over time in the growth rate, which is due to the changing environment.

From Fig. 5.1 it is not clear in which way cells adapt to the different environments. Therefore we took, at 20 different time points, one individual cell out of the population and calculated the growth rate of this cell in each of the nine environments that are explained in the Supplementary Material, using the maximal metabolite concentrations. We took more time points in the beginning of the evolutionary simulation, because there the growth rate and fitness change the most. We also calculated the growth rate in the model without WGD. Because we observed that population heterogeneity is very small we only focus on these individual cells.

The results are shown in Fig. 5.2. We observe that the growth rate in all environments initially (this means at the WGD) decreases. Because the volume of the cell increases approximately two-fold, the surface area to volume ratio decreases, which makes uptake of nutrients more difficult. Therefore, it appears that in these circumstances WGDs are generally unfavorable for a cell (we will come back to this later). However, during evolution the growth rate increases in most environments and in some environments (1,2,3,7,8 and 9, i.e., the nutrient rich environment, the aerobic, oxygen limited environment and all anaerobic environments) the growth rate becomes larger than the initial growth rate, while the volume of the cells at the end of the simulations equals the volume at $t = 0$. The increase in growth rate is accomplished by duplication of certain genes, for example the hexose transporters. In this way, in the long run a WGD can be adaptive.

5.3.3 Duplicated Genes

At the end of the simulations, the genome size of the cells is much reduced. However, some genes are retained in duplicate, on average approximately 50. Some genes are retained in each of the ten simulations we performed, others only in a few. We wondered whether our model can predict which genes are retained in duplicate after the WGD in yeast. In Fig. 5.3 the fraction of genes that are retained in duplicate after the WGD in yeast is plotted against the number of times that these genes are retained in duplicate in our 10 simulations, ranging from 0 to 10 times. A similar approach of measuring the predictive power of a metabolic model was used in Pal *et al.* (2006).

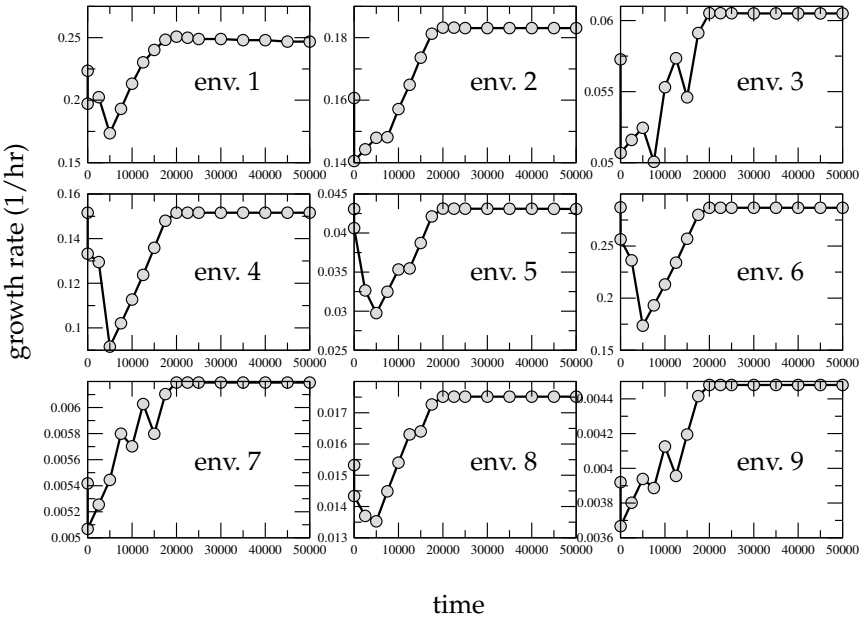


Figure 5.2: Growth rate as a function of time in all nine environments. Note that we zoomed in on the first half of the evolutionary simulation, during which the genome shrinks most.

We observe that genes that are never retained in duplicate in our model, have the smallest probability to be retained in duplicate in *S. cerevisae*, also significantly lower than on average (the dashed line). This probability increases for genes that are 1-3 out of ten times retained in duplicate. Interestingly, for genes that are even more often retained in duplicate in our model, this probability decreases again.

This surprising behavior is an artefact of using *S. cerevisiae* instead of a pre-WGD ancestor, because the metabolic network of *S. cerevisiae* is already the result of a WGD. How does this lead to the observed drop in Fig. 5.3?

More than 80% of the ohnolog pairs (a pair of duplicated genes that arose from a WGD), for which both genes are present in the model, function in the same reaction. Furthermore, these reactions always remain functional if one ohnolog is removed, i.e. the ohnologs function in an "OR"-like way. Note that in our model, duplicated genes are also implemented in an "OR"-like way. This means that reactions that are retained in duplicate after the WGD in *S. cerevisae*, are now performed by at least two genes. Therefore, if such a reaction would again remain duplicated in our simulations, each of the two ohnologs can remain duplicated, while the other can be deleted. Therefore, in only 50% of the cases (or less, if more genes code for this reaction) a certain gene that codes for this reaction will be retained in duplicate. This means that genes that are *always* retained in duplicate, either belong to reactions for which all genes belonging to that reaction

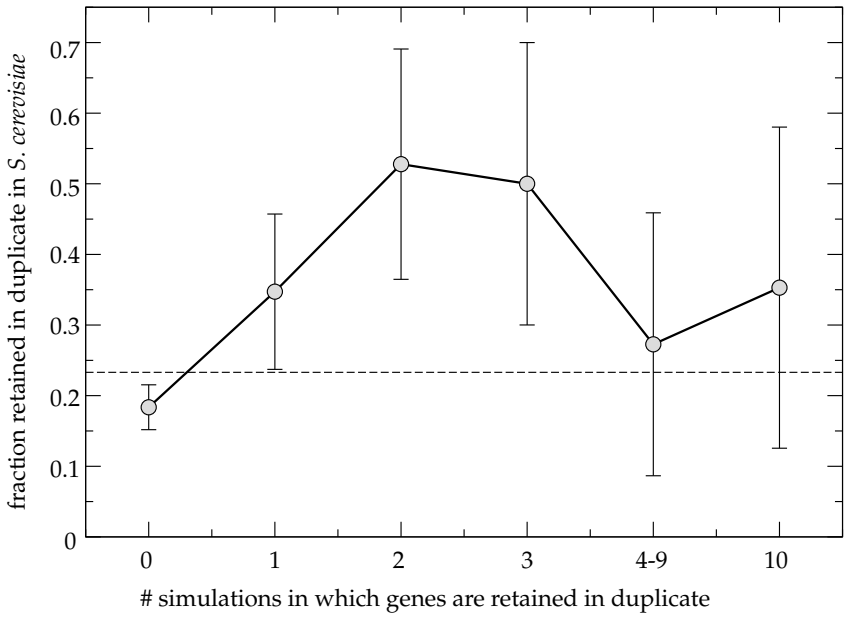


Figure 5.3: Fraction of genes that are retained in duplicate after the WGD in *S. cerevisiae*, for 6 categories of genes (genes that are 0, 1, 2, 3, 4-9 or 10 times retained in the simulations). The average retention fraction is given by the dashed line.

are retained in duplicate (which happens very seldom), or code for a reaction on their own. If genes code for a reaction on their own, they most most often do *not* belong to an ohnolog-pair. This explains the fact that for larger probabilities to be retained in our model, the probability to belong to an ohnolog-pair decreases.

To find out what kind of reactions are retained in duplicate in the simulations, we looked at reactions that have a higher reaction multiplicity at the end of the simulations than at the beginning, in at least two out of ten simulations. There are 44 of such reactions. 20 of these reactions belong to the glycolysis/gluconeogenesis, 19 are extracellular transport reactions, 2 belong to amino acid metabolism, 1 to the pentose phosphate pathway, 1 to pyruvate metabolism and 1 to fatty acid biosynthesis.

Our model therefore predicts that genes that function in transport reactions and glycolysis have a higher chance to be kept in duplicate after WGD. For example, all of the 17 glucose transporters genes that are described in the model are retained in duplicate in the simulations. In *S. cerevisiae*, 6 out of 17 glucose transporters described in the model form an ohnolog pair. Indeed it has been found that yeast species that underwent WGD have much more hexose transporter genes than yeast species that did not (Conant & Wolfe, 2007), which is again in correspondence with our results. The duplication of hexose transporters and glycolysis genes leads to an increase in the maximal glycolytic flux.

Is there also evidence that glycolysis and transport reactions are more often retained in duplicate in *S. cerevisae*? Each gene belongs to one or more reactions and for every reaction the pathway or function is given in the model. Therefore, each gene belongs to at least one pathway/function. We now counted how often genes are assigned to glycolysis and transport reactions, both for genes that are known to be retained in duplicate after WGD (see Byrne & Wolfe (2005)) and genes that are not. Genes that are assigned to different reactions were counted more than once. The result is shown in Table 5.2. We calculated the P values in this table using a χ^2-test.

Table 5.2: Correlation between transport reactions and glycolysis reactions and WGD.

	transport			glycolysis		
	yes	no	percentage	yes	no	percentage
WGD	67	256	21%	18	305	5.6%
non-WGD	138	895	13%	25	1008	2.4%
percentage	33%	22%		42%	23%	
P value	0.017			0.035		

We observe for example that 21% of the ohnologs are functional in a transport reaction, while of the other genes this is only 13%. For glycolysis genes, these percentages are 5.6% and 2.4% respectively. The fact that genes belonging to glycolysis reactions are often retained in duplicate has already been noted by (Conant & Wolfe, 2007). It must be noted however that after the WGD in yeast many more genes are remained in duplicate and only a minority of these genes have a function in glycolysis or transport (as can also be seen from Table 5.2). All other genes are kept in duplicate for reasons our model does not explain. Furthermore, many ohnologs are not even incorporated in the model (only 750 out of over 6000 genes are incorporated in the model), partially because they are not metabolic genes.

5.3.4 The Effect of WGD on Adaptation to New Environments.

Up to now we have seen that the effects of a WGD in our simulations are comparable to the effect of the WGD in *S. cerevisiae*. The WGD is followed by massive gene loss, transporter genes and glycolysis genes are more often retained, which leads to an increase in glycolytic flux. Furthermore we have seen that on the long term, the fitness can increase compared to before the WGD. However, it appears that a WGD can not increase the growth rate instantaneously (see Fig. 5.2).

In these simulations we have assumed that cells were already perfectly adapted to the environment, because we used the maximal flux through each reaction experienced in 1000 environments as the initial constraints on the network (see section 5.2.2). Therefore, because of the decrease in surface area to volume ratio, the fitness is initially decreased after WGD.

We wondered whether a WGD could lead to a fitness increase if cells were not previously adapted to the environment. To test this, we changed the initialization procedure described in section 5.2.2 somewhat. We performed 9 different simulations, in which we assumed that cells were only adapted to eight of the nine environments. This was accomplished by using as initial constraints the maximal fluxes attained in these 8 environments. During the simulations we let cells evolve mostly in the previously unknown environment. With a probability of 90% the "new" environment was experienced, with a probability of 10% one of all nine environments was chosen randomly.

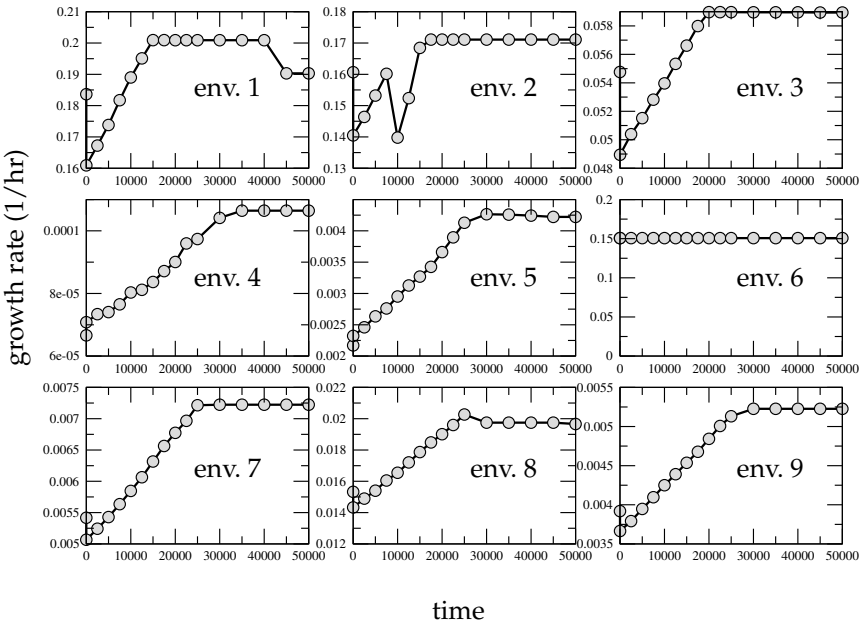


Figure 5.4: Growth rate as a function of time in all nine environments, cells were fully adapted to all environments except the environment for which we report the growth rate. In the simulation for environment 4, a "temperature" of 1e-05 was used, because the growth rate was so low in this environment. Again we zoomed in on the first half of the evolutionary simulation.

The results of these nine simulations are shown in Fig. 5.4. For every simulation, the growth rate in the "new" environment is shown. For most environments, we still observe a drop in the fitness. For environments 4 and 5 however, we observe that the fitness increases instantly at the WGD and keeps increasing continuously, until it reaches a certain level. These two environments are aerobic growth on ethanol and acetate respectively.

Why is it that in these two environments fitness immediately increases? All other seven environments are glucose based environments. Four of these are anaerobic and three are aerobic environments. Therefore, these environments

mostly differ in the amounts of other metabolites (e.g. oxygen, ammonia or phosphate). This means that adapting to one environment also gives a pretty good adaptation to the others. Environments 4 and 5 however, are based on different carbon sources (ethanol and acetate). Hence, adaptation to all other eight environments does not give a good adaptation to growth on ethanol or acetate.

A WGD will always decrease the maximal exchange fluxes, because the surface area to volume ratio becomes lower (see Eq. 5.4, although for very low values of x_i, which we did not take into account, a WGD could increase exchange fluxes, because in that case the surface area is not limiting). The maximal intracellular fluxes however *do* increase, because the increase in dosage (γ_i) is larger than the increase in volume (β). Therefore, if the intracellular fluxes are perfectly adjusted, cells do not benefit from the increase of these maximal fluxes and cells only experience the negative effect of the lower surface area to volume ratio. If these intracellular fluxes are not perfectly adjusted, the cells can gain a benefit by the increase in the maximal flux of these reactions and therefore cells can increase their fitness.

5.3.5 Pathway Usage During Evolution

During evolution, we expect that pathways are activated and inactivated, because genes are deleted or metabolic capacities change. To study how the fluxes through reactions can change during evolution, we performed flux variability analysis (Mahadevan & Schilling, 2003). In FBA, very often there exist several alternative optimal flux distributions with equal growth rate. Flux variability analysis identifies, for every reaction, the minimal and maximal flux value that that reaction can have, given that the growth rate is optimal. We now performed flux variability analysis over time, like we did for the growth rate in Fig. 5.2, using the evolutionary simulation we also used in that figure.

In Fig. 5.5 we have shown the minimal absolute value for several reactions in all environments. Only reactions that have a certain minimum average flux are shown. We clustered the reactions using their minimal flux value. Therefore, reactions that are active or inactive together are clustered. In this way we can visualize different pathways, as reactions that act as one functional unit. Indeed we observe that reactions in the same pathway are often clustered together.

From this picture it is clear that different reactions (or pathways) are activated in different environments. The reactions in the lower part of the figure for example are only active in environment 1, the rich environment. These reactions are uptake reactions for metabolites that are only present in environment 1. The large cluster that is only inactive in the rich environment all function in amino acid production. This makes sense, because environment 1 is the only environment with amino acids in it.

In Fig. 5.6 we have selected the behavior of a few reactions in the anaerobic environment (env. 3). Only reactions that change significantly over time in this environment are shown. Note that the extracellular environment is constant over time. As we saw in Fig. 5.5, we observe that reactions are most often not "turned off" or "on" on their own. For example, reaction 278, 1148, 1149 and 1186 are

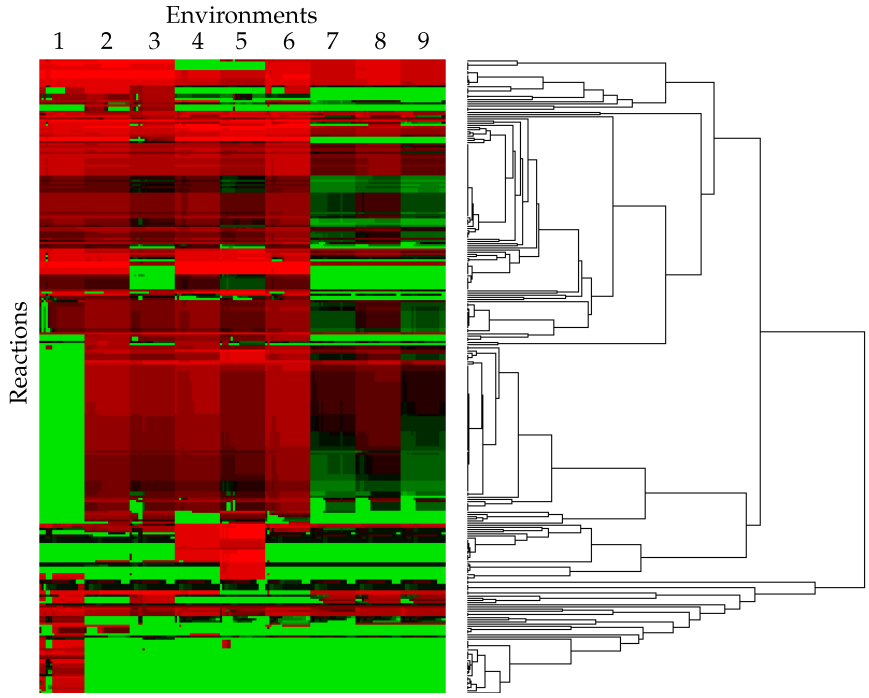


Figure 5.5: Minimal flux through reactions in all 9 environments over time. Green indicates low flux, black intermediate flux and red high flux. The clustering is performed using EPCLUST (http://ep.ebi.ac.uk/), using a linear correlation based distance and average linkage. Columns indicate the different environments over time, rows the different reactions. We included four more time points than in Fig. 5.2 ($t = 62500$, $t = 75000$, $t = 87500$ and $t = 100000$). See also the color plate on page 150.

turned on together. These reactions all function in amino acid metabolism and they indeed form a connected subgraph in the metabolic network. At the same time that this pathway is activated, cells start excreting acetate and valine (reaction 78 and 1255). Under these circumstances, the pre-WGD cell does not excrete acetate (nor valine), but due to the increase in glycolytic flux, cells start acetate excretion. This shows that during evolution, the intracellular environment changes, due to a change in uptake. This change in intracellular environment drives the metabolic network to a different steady state, often in a non-obvious way.

Another interesting pattern can be seen for reactions 169, 180, 761, 112 and 763. These reactions function in amino acid metabolism. At time point 5 and 6, these reactions are shut down. At the same time a number of other reactions become active, which appear to take over its function. In later stages of evolution none of these pathways is active. Again, it seems that for a short period of time, the steady state of the network changes drastically, by activating some pathways

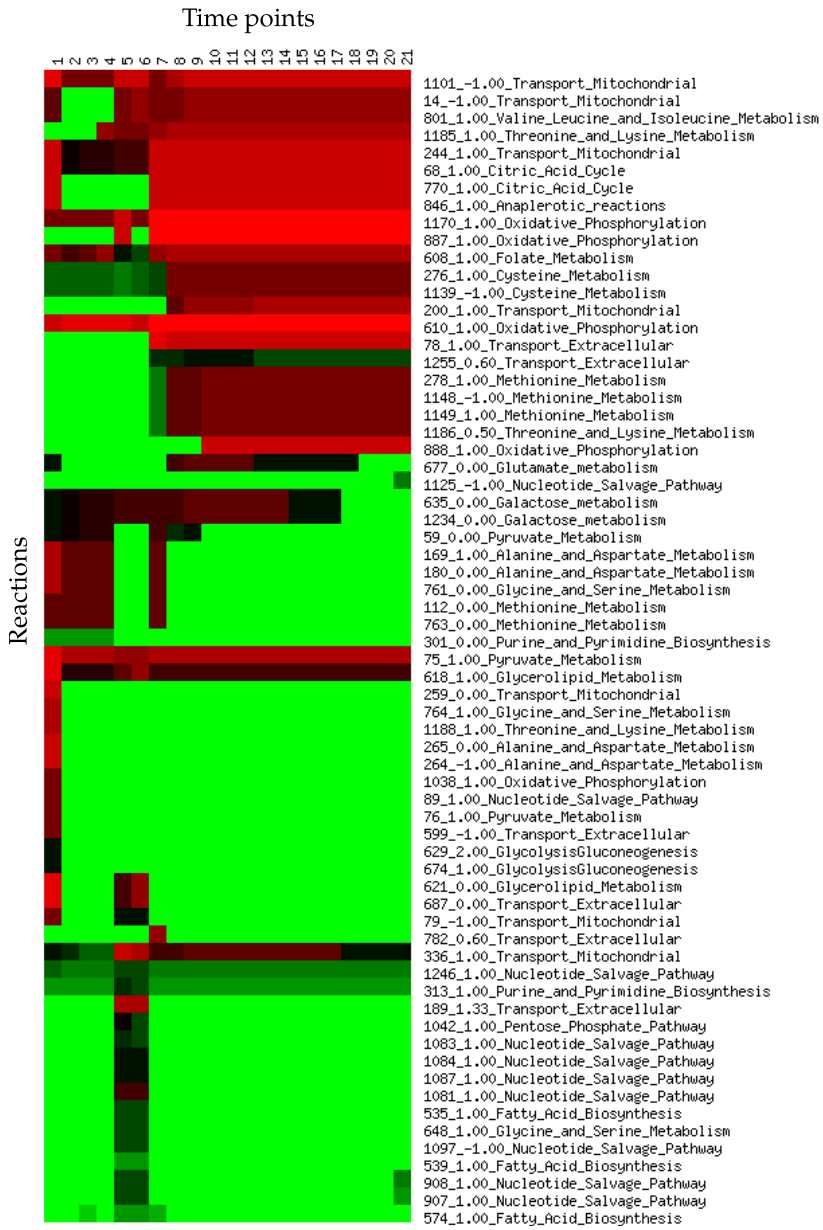


Figure 5.6: Minimal flux through reactions in environment 3. At the right the reaction number, γ_i and the pathway/function for each reaction is shown. γ_i (see Eq. 5.4) indicates the reaction multiplicity at the end of the simulation divided by the reaction multiplicity at the beginning of the simulation, a value of -1 means that no gene is associated with the reaction. Clustering is done as in Fig. 5.5. The time scale is also as in Fig. 5.5, $t = 1$ indicates pre-WGD. See also the color plate on page 151.

and deactivating other. Therefore, an other intriguing explanation why so many yeast genes are dispensable could be that they are only used in certain evolutionary circumstances.

It is also interesting to note that the flux through some reactions show an all-or-nothing behavior, with very abrupt changes (for example 78, 770, 846), while other reactions change rather continuously over time (for example, 244,68,635, 1234, 244, and 336), indicating that a complex metabolic network can give rise to different kinds of evolutionary dynamics.

The abrupt stop of activity of reactions 635 and 1234 is due to a gene deletion, as we see that at the end of the simulation both reaction 635 and 1234 have a reaction multiplicity of 0. However, in most other cases, a drop in activity is not caused by a gene deletion of that particular reaction. For example, there is a large cluster of reactions (259-674) that is only active before WGD, but after the WGD loses its activity, which is never regained. However, some of these reactions remain active after WGD in other environments. This shows that the WGD in itself, without the subsequent gene loss, can already change the dynamics of the metabolic network.

5.4 Discussion

We have studied the evolution of a metabolic network after whole genome duplication. Using very simple assumptions (gene dosage and a correlation between cell size and genome size), we have been able to formulate a model that can describe network evolution after WGD. This simple model captures some essential features of this process.

A WGD is in our model always followed by massive gene loss, because the WGD increases the cell size, which leads to a larger surface area to volume ratio. This causes a slower uptake of nutrients and therefore, the cell size is slowly restored to the pre-WGD size. In this process, many duplicate genes are lost, which changes, in contrast to the WGD itself, the relative protein abundances of each gene. In this way, the constraints on the metabolic network change.

Furthermore, our model predicts gene retention quite well. Genes that are never retained in our model have a significantly lower probability to be an ohnolog in *S. cerevisiae*. We also find that the WGD in yeast leads to an increase in glycolytic flux, because both glycolysis and transporter genes are retained more than on average.

This last result is in precise correspondence with the results of Conant & Wolfe (2007). They showed that glycolysis genes are retained in duplicate more often than expected by chance. Furthermore they found, using a kinetic model for glycolysis and assuming gene dosage, that the retention of duplicated genes leads to an increase in glycolytic flux. In contrast, we have studied the effect of a WGD on the whole metabolic network of yeast, using evolutionary simulations. From this model we conclude that it is indeed to be expected that, at least in a glucose-rich environment, a WGD will lead to an increase in the glycolytic flux.

In our model, reactions coming from pyruvate are not retained in duplicate.

This is in contrast to what is found in *S. cerevisiae*. Pyruvate can be regarded as a "hub" in the metabolic network, because many different pathways have pyruvate as a starting point. Therefore, in the model, an increased glycolytic flux can be shuttled to different pathways after pyruvate, which could explain why in the model reactions coming from pyruvate have not been duplicated. In contrast, the glycolysis is a very linear pathway, in which it is much more difficult to reroute fluxes. Whether this explanation also holds in *S. cerevisiae* is doubtful, as several reactions coming from pyruvate have been duplicated.

In our model we assume that duplicated genes increase the maximal possible flux through a reaction. Indeed, there is ample evidence for this assumption. Papp *et al.* (2004) showed that the average *in silico* flux for duplicated enzymes in yeast is higher than of single-copy enzymes. Furthermore, Kuepfer *et al.* (2005) showed that gene dosage explains the retention of some duplicate genes in yeast, because they were associated with reactions with high (experimentally measured) carbon fluxes. Finally, genes that are retained in duplicate after WGD have on average higher expression rates (Seoighe & Wolfe, 1999). Here we now show that also for genes that have been retained in duplicate after WGD, the average *in silico* flux for duplicated enzymes is higher than expected by chance. Furthermore we have shown that this is more so in an anaerobic environment than in an aerobic environment and also more so in a glucose-rich environment than in a glucose-poor environment. This confirms the hypothesis that the WGD has been important in the adaptation to fermentation in a glucose-rich environment.

Functional divergence through subfunctionalization or neofunctionalization is of course also a possible evolutionary outcome of gene duplication, which we did not include. However, we should expect that divergence of duplicated genes is not important in the initial stages of evolution after WGD. Furthermore, functional divergence of duplicate genes through subfunctionalization can also partly be accounted for in our model. For example, the glucose transporter gene HXT 1 of *S. cerevisiae* is expressed at high glucose concentrations, while HXT 2 and 4 are expressed at low glucose concentrations (Ozcan & Johnston, 1999). By gene duplication, it has been possible for *S. cerevisiae* to specialize genes for different glucose concentrations and hence different fluxes. If there were only one glucose transporter gene, this gene would probably be expressed under all glucose concentrations. Because this gene would not be specialized for high glucose concentrations, the maximum attainable flux would be lower. Generally, we expect that if duplicated genes undergo subfunctionalization, both duplicates will perform their function better and we expect a higher possible flux for these reactions.

We did not take genetic regulation into account in our model, in contrast to for example Covert *et al.* (2001), Herrgard *et al.* (2006) and Shlomi *et al.* (2007), because genetic regulation will change during evolution, in a way that is very difficult to predict. Therefore we chose to assume that genetic regulation is optimized during evolution such as to maximize the growth rate. Furthermore, in our approach, we do not distinguish between subfunctionalization (and changes in gene expression) or increased gene dosage, but simply assume that both phenomena lead to higher fluxes.

Previously, it has been found that ribosomal proteins are retained in duplicate

after WGD more often than expected by chance (Wolfe, 2004). In our model we find a strong correlation between growth rate (in a certain environment) and the total flux through the network. Therefore, if the growth rate would increase after WGD, what we expect, the total flux through the network should also increase. To be able to reach this higher total flux, it might be necessary to have more ribosomes, which could explain, again assuming dosage dependence of ribosomal genes, the retention of ribosomal genes.

We also tried to evolve the network using only single gene duplications, instead of a WGD. In this simulation, we used equal gene duplication and gene deletion rates. Surprisingly, we found very similar results. Again we found that glycolysis and transporter genes were duplicated. It has been assumed that, in contrast to single gene duplications, WGD can lead to the duplication of whole pathways, such as the glycolysis. We found however that using single gene duplications, cells were still able to duplicate the whole glycolysis pathway. This can be explained because one reaction after the other becomes rate-limiting in the glycolysis and in this way the whole pathway is duplicated. We expect however that single gene duplication do not work for reactions which are in our model, described by a Boolean "AND" (GENE A "AND" GENE B) (such as protein complexes). In the glycolysis, only one reaction is described using an AND function of only two genes (fructose biphosphate aldolase). Duplicating this reaction is apparently still possible by single gene duplication.

Pal *et al.* (2006) have adopted a similar approach to model genome reduction in *Buchnera aphidicola*. Performing sequential gene deletions in the metabolic network of *Escherichia coli*, they arrived to a minimal genome, that could still function in a nutrient-rich environment. They indeed showed for the first time that FBA can also be used to study metabolic network evolution and they were able to predict the evolutionary outcome of genome shrinkage in *B. aphidicola* with more than 80% accuracy. Their algorithm is much simpler as ours, because we study genome evolution after WGD. Therefore, genes can also be present twice in our model and we needed a way to describe the difference in fluxes for reactions that are present once or twice. Furthermore, we used evolutionary simulations instead of the greedy search that was used in Pal *et al.* (2006).

Concludingly, we have developed a model, based on flux balance analysis, with which it is possible to study the effects of a WGD and subsequent gene loss on a metabolic network. This model can predict part of the evolutionary outcome of the WGD in *S. cerevisiae*. There is satisfactory agreement between genes that are retained in duplicate in our model and genes that have been retained in duplicate in *S. cerevisiae*. Furthermore, we have shown that WGDs can be selective in environments, for which a cell was not yet perfectly adapted. This is in nice correspondence with the hypothesis that the WGD helped yeast to adapt to the newly arisen environment of glucose-rich fruits (Merico *et al.*, 2007; Conant & Wolfe, 2007). It must be noted however, that also other events have been crucial for yeast to evolve the ability to grow fast under anaerobic conditions, such as horizontal gene transfer (Gojkovic *et al.*, 2004). Finally, we found that during evolution, metabolic pathways are activated and inactivated, indicating that gene dispensability changes during evolution, which could for some part explain why

so many genes are dispensable in *S. cerevisiae* (Papp *et al.*, 2004).

5.5 Supplementary Material

In this section we precisely describe the environments we used in the evolutionary simulations. In Table 5.3 the lower and upper bounds of the uptake rates are given. For every nutrient the actual uptake rate is chosen from a uniform distribution between the lower and upper bounds that are given in the table. Note that these uptake rates are the rates before WGD. After WGD, these values change according to Eq. 5.4. Metabolites that are not mentioned for a particular environment have a maximal uptake rate of 0 mmol/(gram Dry Weight hour) in that environment. The first environment is comparable to the rich medium used by Bilu *et al.* (2006). The other eight are similar to the minimal media used by Papp *et al.* (2004).

Table 5.3: description of the environments (values in mmol/(gram DW hour)

Environment	Metabolite	Lower Bound	Upper Bound
1	adenine	0.0075	0.7575
1	alanine	0.0075	0.7575
1	arginine	0.0075	0.7575
1	asparagine	0.0075	0.7575
1	aspartate	0.0075	0.7575
1	cysteine	0.0075	0.7575
1	deoxycytidine	0.0075	0.7575
1	glutamine	0.0075	0.7575
1	glycine	0.0075	0.7575
1	guanine	0.0075	0.7575
1	histidine	0.0075	0.7575
1	isoleucine	0.0075	0.7575
1	leusine	0.0075	0.7575
1	lycine	0.0075	0.7575
1	methionine	0.0075	0.7575
1	phenylalanine	0.0075	0.7575
1	proline	0.0075	0.7575
1	serine	0.0075	0.7575
1	thymidine	0.0075	0.7575
1	threonine	0.0075	0.7575
1	tryptophan	0.0075	0.7575
1	tyrosine	0.0075	0.7575
1	valine	0.0075	0.7575
1	glucose	0.03	3.03
1	oxygen	0.03	3.03
1	ammonia	unbounded	unbounded
1	carbondioxide	unbounded	unbounded
1	phosphate	unbounded	unbounded
1	potassium	unbounded	unbounded
1	sodium	unbounded	unbounded
1	sulfate	unbounded	unbounded
1	water	unbounded	unbounded
2	glucose	0.03	3.03
2	oxygen	0.03	3.03
2	ammonia	unbounded	unbounded
2	carbondioxide	unbounded	unbounded
2	phosphate	unbounded	unbounded
2	potassium	unbounded	unbounded
2	sodium	unbounded	unbounded
2	sulfate	unbounded	unbounded
2	water	unbounded	unbounded

Table 5.3: continued.

Environment	Metabolite	Lower Bound	Upper Bound
3	glucose	0.03	3.03
3	ammonia	unbounded	unbounded
3	ergosterol	unbounded	unbounded
3	carbondioxide	unbounded	unbounded
3	hexadecenoate	unbounded	unbounded
3	octadecanoate	unbounded	unbounded
3	octadecenoate	unbounded	unbounded
3	octadecynoate	unbounded	unbounded
3	phosphate	unbounded	unbounded
3	potassium	unbounded	unbounded
3	sodium	unbounded	unbounded
3	sulfate	unbounded	unbounded
3	water	unbounded	unbounded
3	zymosterol	unbounded	unbounded
4	ethanol	0.09	9.09
4	oxygen	0.09	3.09
4	ammonia	unbounded	unbounded
4	carbondioxide	unbounded	unbounded
4	phosphate	unbounded	unbounded
4	potassium	unbounded	unbounded
4	sodium	unbounded	unbounded
4	sulfate	unbounded	unbounded
4	water	unbounded	unbounded
5	acetate	0.09	9.09
5	oxygen	0.09	3.09
5	ammonia	unbounded	unbounded
5	carbondioxide	unbounded	unbounded
5	phosphate	unbounded	unbounded
5	potassium	unbounded	unbounded
5	sodium	unbounded	unbounded
5	sulfate	unbounded	unbounded
5	water	unbounded	unbounded
6	glucose	0.03	3.03
6	oxygen	unbounded	unbounded
6	ammonia	unbounded	unbounded
6	carbondioxide	unbounded	unbounded
6	phosphate	unbounded	unbounded
6	potassium	unbounded	unbounded
6	sodium	unbounded	unbounded
6	sulfate	unbounded	unbounded
6	water	unbounded	unbounded

Table 5.3: continued.

Environment	Metabolite	Lower Bound	Upper Bound
7	glucose	1	3
7	ammonia	0.0003	0.0303
7	ergosterol	unbounded	unbounded
7	carbondioxide	unbounded	unbounded
7	hexadecenoate	unbounded	unbounded
7	octadecanoate	unbounded	unbounded
7	octadecenoate	unbounded	unbounded
7	octadecynoate	unbounded	unbounded
7	phosphate	unbounded	unbounded
7	potassium	unbounded	unbounded
7	sodium	unbounded	unbounded
7	sulfate	unbounded	unbounded
7	water	unbounded	unbounded
7	zymosterol	unbounded	unbounded
8	glucose	1	3
8	ammonia	unbounded	unbounded
8	ergosterol	unbounded	unbounded
8	carbondioxide	unbounded	unbounded
8	hexadecenoate	unbounded	unbounded
8	octadecanoate	unbounded	unbounded
8	octadecenoate	unbounded	unbounded
8	octadecynoate	unbounded	unbounded
8	phosphate	0.00003	0.00303
8	potassium	unbounded	unbounded
8	sodium	unbounded	unbounded
8	sulfate	unbounded	unbounded
8	water	unbounded	unbounded
8	zymosterol	unbounded	unbounded
9	glucose	1	3
9	ammonia	unbounded	unbounded
9	ergosterol	unbounded	unbounded
9	carbondioxide	unbounded	unbounded
9	hexadecenoate	unbounded	unbounded
9	octadecanoate	unbounded	unbounded
9	octadecenoate	unbounded	unbounded
9	octadecynoate	unbounded	unbounded
9	phosphate	unbounded	unbounded
9	potassium	unbounded	unbounded
9	sodium	unbounded	unbounded
9	sulfate	0.000003	0.000303
9	water	unbounded	unbounded
9	zymosterol	unbounded	unbounded

6

Summarizing Discussion

6.1 Review

In this thesis we have studied the evolutionary dynamics of metabolic adaptation, using a modeling approach. We did this in two different systems, the *lac* operon of *Escherichia coli* and the metabolic network of *Saccharomyces cerevisiae* after a whole genome duplication (WGD).

Part I of this thesis is concerned with the evolution of the *lac* operon. Although the *lac* operon is maybe the best studied example of genetic regulation, after more than 40 years of research it is still not fully understood.

In **chapter 2** we showed that, using *in silico* evolution of the *lac* operon in a fluctuating environment, we can surprisingly well reproduce the experimentally observed promoter function (Setty *et al.*, 2003). As in these experiments, we find that the promoter function has a relatively high repressed transcription rate and shallow shifts between the different levels of transcription. Both these properties of the promoter function result in a continuous response to the natural inducer lactose, although the *lac* operon has the potential for bistability because of an inherent positive feedback loop.

We showed that the reason that bistability has been observed in experiments of Novick & Weiner (1957) and Ozbudak *et al.* (2004) is that artificial inducers were used in these experiments. As the positive feedback loop for artificial inducers is much stronger than for lactose, because they are not degraded by β-galactosidase, the results for these artificial inducers cannot be generalized to the natural inducer lactose.

Since the experiments of Novick & Weiner (1957) that showed bistability of the *lac* operon with respect to TMG, people have tried to explain these experimental findings from an evolutionary point of view. Our results emphasize that interpreting experimental results is often not a trivial task. Although the use of artificial inducers makes much sense from an experimental point of view, it is very tempting, but wrong, to generalize the results to the biologically relevant inducer lactose.

In **chapter 3** we took a closer look at the effect of stochasticity on the evolution and dynamics of the *lac* operon. Because it has been proposed that stochasticity in gene expression would enable switching between the two equilibria (Thattai & Van Oudenaarden, 2004), this is indeed an important factor to take into account.

We were able to incorporate stochasticity by adding only one extra parameter, which has been experimentally measured, namely the burst size of protein production. In this way we were able to qualitatively reproduce noise measurements that were done in *E. coli* (Elowitz *et al.*, 2002).

Again we found that the dynamics are very different for artificial inducers than for lactose. The *lac* operon behaves much noisier when induced by TMG or IPTG than by lactose, again due to the stronger positive feedback loop. Furthermore we found that, given that *E. coli* grows in a spatially structured environment and has a genetically structured population, we should expect the noise due to spatial and genetic heterogeneity to be at least as large as the noise due to stochasticity in gene expression.

Finally we showed that in the stochastic model, cells evolved higher repressed transcription rates than in the deterministic model. These higher transcription rates lead to lower noise in gene expression. From this we concluded that cells avoided high amounts of noise in gene expression. Therefore, instead of using stochasticity to switch between equilibria, cells evolve to regimes where noise in gene expression is low.

The reason that both in the deterministic model (**chapter 2**) and in the stochastic model (**chapter 3**) graded switches are favored over bistable switches, is that a graded switch allows for a faster response to the environment. Because protein dynamics is slow, of the order of the generation time, fast switching is very important for the cells. In the deterministic model, bistable cells need to wait before the external lactose concentration increases over the point at which bistability is lost, which takes much time. In the stochastic model, cells can switch stochastically from one equilibrium to the other, which in theory could render bistability advantageous. We showed that this is not the case. On the contrary, cells evolve even further from the bistable region in the stochastic simulations. When a cells switches stochastically from one equilibrium to the other, this still does not occur instantaneously, as has been assumed in models explaining the advantageousness of stochasticity in bistable systems (Thattai & Van Oudenaarden, 2004; Wolf *et al.*, 2005). Furthermore, we have shown that bistability is caused by a low repressed transcription rate, which *increases* the delay in lactose uptake. Therefore, in the *lac* operon, bistable switching is always slower than continuous switching.

Different models have been proposed to explain the advantageousness of bistability in for example the *lac* operon (Thattai & Van Oudenaarden, 2004; Wolf *et al.*, 2005; Kussell & Leibler, 2005). However, these models are all very general and not specific for the *lac* operon. Thattai & Van Oudenaarden (2004) find that stochastic switching can be advantageous in a periodic changing environment, Wolf *et al.* (2005) if the sensory mechanism is imperfect and Kussell & Leibler (2005) explain the advantageousness of bistability by introducing a cost for sensing the environment. All these assumptions do not appear very realistic for the *lac* operon. For example, we know that *E. coli* does sense the external lactose levels and apparently is willing to "pay the price". It appears that, because the cell is able to sense lactose, it will use this information instead of switching stochastically.

Interestingly, Kashiwagi *et al.* (2006) have shown that for environments that

E. coli cannot sense, there still exists a mechanism for cells to switch to the "right" state, which means the state with the highest growth rate. This behavior is explained using a stochastic model, in which both the transcription, translation and degradation/dilution rates increase with the growth rate. If these conditions are met, cells will experience more stochasticity in the state with lower growth rate and hence will more easily switch to the state with less stochasticity than vice versa. Combined with our study and the results of Kussell & Leibler (2005), this indicates that a bistable response can be advantageous for environments for which cells do not have a sensory mechanism. When cells do have a sensory mechanism, a graded response is favored.

In **Part II** we studied the evolution of *S. cerevisiae* after its WGD. In **chapter 4** we approached this question from a mutational perspective. More generally, we looked at the mutational dynamics of genome shrinkage, which occurs after WGD, but also for example in the evolution of parasitic or endosymbiotic bacteria. We observed that the pattern of gene loss, both in yeast and the bacterium *Buchnera aphidicola*, could not be explained by random loss of single genes. We found that, in yeast, the data could entirely be explained by assuming that gene loss occurs via deletion of stretches of base pairs, with an average length of approximately 500 base pairs. Therefore, most genes will be deleted on their own, because the average gene length of *S. cerevisiae* is longer than 500 base pairs. However some genes will be deleted in pairs, triples etcetera.

We were able to show that in yeast the clustering of deleted genes was due to the mutational dynamics and not by clustering of functional genes in the genome of yeast. For example, we showed that neighboring genes that have a small intergenic distance in pre-WGD species, have a larger probability of being both deleted in post-WGD species. Secondly, large genes have a larger probability of being deleted on their own. Thirdly, two neighboring genes in a pre-WGD species that are both deleted in a post-WGD species, have a larger probability of being deleted on the same chromosome in the post-WGD species. Finally we also measured the size distribution of base pair deletions in pseudogenes of *S. cerevisiae*. We found that, using precisely this distribution of deletion sizes, we could satisfactorily explain the clustering of deleted genes in *S. cerevisiae*.

When we tried to use this base pair deletion model to simulate gene loss in *B. aphidicola*, we found that larger deletions (approximately 1000 base pairs) were required, although the average gene length in *B. aphidicola* is smaller than in yeast. We propose that this difference is due to the organization of bacterial genomes in operons. We found that the amount of clustering of genes in operons in *E. coli*, which is closely related to *B. aphidicola*, is sufficient to explain the larger amount of clustering of deleted genes in *B. aphidicola*.

In **chapter 5** we studied the evolution of the metabolic network of *S. cerevisiae* after its WGD. Using Flux Balance Analysis we were able to formulate a model in which we could study the consequences of a WGD, both from a metabolic and evolutionary perspective. We found that during evolution different pathways are used in the same environment, which could account for part of the apparent dispensability of many yeast genes. Furthermore we found a significant correlation between genes that are retained in duplicate in *S. cerevisiae* and in our model. We

predicted that both transporter genes and glycolysis genes are retained in duplicate more than on average. Indeed, we found that in yeast both these categories of genes are retained in duplicate more often than expected by chance. In the model this causes the glycolytic fluxes to increase after WGD, as has been proposed to have happened in yeast after its WGD (Conant & Wolfe, 2007).

Furthermore we have shown that a WGD can lead to an immediate fitness increase, if cells are not yet perfectly adapted to the environment. The reason for this is that intracellular fluxes that were rate-limiting before WGD can increase their maximal flux by WGD, which can increase fitness. Exchange fluxes however always become more constrained after WGD. Therefore, if cells are perfectly adapted to their environment, a WGD will lead to an immediate fitness decrease. Both the increase of glycolytic flux and the fact that a WGD can be adaptive in new environments confirm the hypothesis that the WGD in yeast has helped yeast to adapt to a new environment of glucose-rich fruits, that occurred after the appearance of angiosperms (Merico *et al.*, 2007; Conant & Wolfe, 2007).

6.2 Model Complexity and the Importance of Different Time Scales

There is a long history in mathematical modeling of complex biological systems. Most often, this is done from an "engineering" point of view: can we understand the dynamics of a biological system by modeling it in detail. There also exists a large body of research of evolutionary modeling of biological systems. These models are most often general, abstract models. However, very few attempts have been made to combine these two approaches to study the evolution of such complex systems. Exceptions are Pal *et al.* (2006), who studied the reductive evolution of *B. aphidicola*, using a genome-scale metabolic model and Gatenby & Vincent (2003), who studied somatic evolution of cancer cells, using a quantitative model of carcinogenesis.

In this thesis, we studied how metabolic adaptation interacts with the evolutionary dynamics. To this end, we developed models in which both the metabolic and the evolutionary time scales are incorporated. We combined evolutionary modeling and detailed modeling of the metabolic system under consideration. This led to relatively complex and computationally demanding models. In this section we discuss why and how we adopted this approach.

In **part I** of this thesis, we combined an individual oriented, spatial model with a detailed differential equation model to simulate *lac* operon evolution. In this model, very different time scales are taken into account. The cellular behavior is modeled in detail using a differential equation model. These dynamics range from seconds-minutes (metabolites), to hours (proteins). Cells grow, divide and die on a time scale of one generation, which is also in the order of hours. They live in a fluctuating environment, which fluctuates on different time scales, as we will discuss later. Finally, cells evolve their promoter function over many generations.

There already exists a long history of modeling the *lac* operon using differential equations, as we described in the Introduction. However, to our knowledge,

we are the first to study the *lac* operon using such an individual based, evolutionary approach. This approach has the advantage that we can study the detailed dynamics of the operon in a fairly realistic environment, in contrast to most studies, that use a bimodal environment (Thattai & Van Oudenaarden, 2004; Kalisky *et al.*, 2007). Obviously, the environment is a crucial factor in determining the evolutionary outcome, for in a constant environment metabolic regulation would not even be needed.

Space can be crucial in the outcome of evolutionary processes, as has first been shown by Boerlijst & Hogeweg (1991), who showed that hypercycles are stable against parasites in a spatial environment, because space allows for extra levels of selection. In the evolution of metabolic regulation, space can also be of crucial importance (Pfeiffer *et al.*, 2001), because it can prevent the "tragedy of the commons". Furthermore, because we modeled space explicitly, we took into account the feedback between the metabolism of the cells and the environment, which in **part I** proved to be important in the evolution of the *lac* operon. Finally, space gives a fluctuating environment over different time scales in a very natural way. The long time scale is defined by the imposed environment, which changes after several hours, while the short time scale is defined by the fluctuations in the spatial environment, due to consumption and diffusion.

Furthermore, we used an evolutionary approach, in which we let the promoter function evolve in a spatial, fluctuating environment. Surprisingly little attention has been paid to the *lac* operon from an evolutionary point of view. However, we believe that, when we want to understand how the *lac* operon works, evolution is of crucial importance. Dekel & Alon (2005) already showed that *E. coli* cells can adapt their promoter functions in only a few hundred generations. As we saw in **part I** of this thesis, because experimental results seemed to indicate that the *lac* operon is a bistable switch, evolutionary reasons have been sought why the *lac* operon behaves this way. By performing evolutionary modeling, we let the operon "decide for itself". In this way we gained insight in how the operon behaves, namely bistably for artificial inducers and continuously for lactose, and why it has evolved this way.

We observed that it is not possible to separate the metabolic, ecological and evolutionary time scales. It turned out that the metabolic and ecological time scales are crucial for the results of the evolutionary simulations. When the fluctuations of the environment are too fast, we observe no regulation whatsoever. If they are too slow, also no regulation is needed (cells will lose their regulation if it is not needed for a long time). Therefore, fluctuations in the environment need to be somewhat slower than the response time of the organism (which is in the order of an hour, due to the slow protein dynamics). In that case, we observed that cells evolve to minimize the switching times between the environments. In our model, this is accomplished by having a relatively high repressed transcription rate, which is in itself costly, but already "prepares" the cells for the nutrient-rich environment. These high repressed transcription rates have also been observed in experiments (Setty *et al.*, 2003; Ozbudak *et al.*, 2004). Clearly, combining the time scales of metabolic regulation and evolution is crucial for this result.

A second advantage of using evolutionary modeling as we did in **part I** is

that it is less vulnerable to parameter uncertainties. Yildirim & Mackey (2003) tried to predict whether bistability occurs if the *lac* operon is induced by lactose, by developing a differential equation model describing induction by lactose and performing an exhaustive literature search for parameter values. Using these parameters, the *lac* operon turned out to be bistable for lactose. However, as we have shown, with the support of experimental observations (Ozbudak *et al.*, 2004), this is not the case.

Indeed, many parameter values of the *lac* operon can be found in literature. However, even for the *lac* operon, there is much parameter uncertainty. Some of the parameters are measured *in vitro*, others *in vivo*, in different environments, etcetera. Even more, different models, that all take their parameter values from literature, can end up with highly different parameter values (compare for example Yildirim & Mackey (2003) and Santillan & Mackey (2004)). Using our approach, the precise parameter values are less crucial, because cells can adapt the parameters describing their promoter function to the other imposed parameter values.

A drawback of such detailed evolutionary modeling, where metabolic, ecological and evolutionary time scales are not separated, is that it is computationally demanding. The detailed dynamics of a population of hundreds of cells must be integrated over a time scale of thousands of days, with a timestep of only 0.2 seconds. Furthermore, due to the relatively small population size, fluctuations are quite large. We solved this problem by performing multiple simulations and choosing the simulation that led to the best performing last common ancestor.

In **part I**, we used differential equations to model the intracellular *lac* operon dynamics. Differential equations are easy to use and there is already a long history of modeling the *lac* operon using differential equations. Another advantage is that they can be analytically solved, as we did in **chapter 2**.

However, differential equations assume that cells are well-mixed entities, which is not the case. Kennell & Riezman (1977) already wondered how it is possible that an *E. coli* cell, which is almost completely filled with proteins, ribosomes, DNA, RNA, etcetera can still function properly. One way cells have solved this problem is that metabolic pathways are compartmentalized and are therefore less diffusion-limited (Ovadi & Saks, 2004), again indicating that cells are not well-mixed systems. Furthermore, differential equations neglect the fact that cells are small entities that are inherently stochastic. The latter disadvantage was solved in **chapter 3**, by using a stochastic modeling approach.

In **chapter 5** of this thesis, we used Flux Balance Analysis (FBA) to simulate evolution of *S. cerevisiae* after its WGD. Considering that our model of the *lac* operon already has so many parameters, it becomes immediately clear that, when modeling the complete metabolic network of *S. cerevisiae*, we have to take a different approach. FBA is a constraint-based way of calculating the flux through a metabolic network. It can only find the steady state fluxes and does not take the dynamics of the metabolic network into account. This means that, in contrast to **part I**, we assume that the optimal activity of all reactions only depend on the nutrient concentrations and not on the precise dynamics of the nutrients. Therefore, in this chapter, one time scale less is taken into account compared with **part I**.

Furthermore, genetic regulation is not explicitly taken into account. Instead, it is assumed that the genetic regulation has evolved such that the growth rate of the organism is optimized. There have been attempts to incorporate genetic regulation in FBA (Covert *et al.*, 2001; Herrgard *et al.*, 2006; Shlomi *et al.*, 2007). However, because we study the evolution of the network, we should expect that genetic regulation is also modified by evolution. It is not easy to predict in which way genetic regulation will be modified during evolution and therefore we assumed that genetic regulation changes during evolution of the network such that the growth rate is optimized.

This metabolic network is again implemented in an individual oriented model. Here we did not embed the cells in a spatial grid. Because, in contrast to **part I**, we assume that optimal activity only depends on the nutrient concentrations, instead of on the dynamics of the nutrients, adding space to the model is not so useful.

6.2.1 The Relationship Between Genetic Regulation and Evolution

One of the major goals we had when starting this research was to gain a better understanding of how genetic regulation and evolution interact. When studying this question, different time scales are of crucial importance. Using genetic regulation, cells adapt to the environment in the order of an hour. However, as we saw in **part I** (and experimentally in Dekel & Alon (2005)), organisms can also evolve their genetic regulation in only a few hundred generations. Given a certain changing environment, cells evolve a certain promoter function. When the environment changes in a fixed way, evolution will reach more or less a steady state. However, when the environments starts fluctuating differently, for example by changing the frequency of environmental changes, the promoter function will change.

This in itself is already an interesting observation. Apparently, genetic regulation is not sufficient to adapt an organism perfectly to each possible environment. A priori we might expect that the promoter function has a certain optimal value for each glucose and lactose concentration and that therefore in every changing environment the same promoter function would evolve. However, it turns out that this is not the case and that the optimal transcription rate also depends on how the environment precisely fluctuates.

Indeed, this observation has very recently been confirmed by Kalisky *et al.* (2007). As we did in **part I**, Kalisky *et al.* (2007) try to explain the experimentally observed shape of the *lac* operon promoter function (Setty *et al.*, 2003). Instead of explicit evolutionary modeling, they used a cost-benefit theory. Using a very different approach from ours, they confirmed our results that the precise shape of the experimentally observed promoter function cannot be understood from optimizing the fitness for every lactose and glucose concentration. However, if pulses of lactose were assumed, there was better correspondence between the experimentally observed promoter function and the promoter function expected by the cost-benefit theory. As we already noted in **part I**, the reason for this is that

the *lac* operon cannot respond to very fast changes in the environment. According to Kalisky *et al.* (2007), this effect decreases the optimal maximal transcription rate. Interestingly, in our study, a fluctuating environment leads to an *increase* in the optimal *minimal* transcription rate, because this avoids bistability and decreases the response time to the environment. It should be noted that Kalisky *et al.* (2007) do not consider the possibility of bistability, but simply assume a minimal transcription rate of zero and therefore do not consider that crucial part of the promoter function.

Furthermore, Kalisky *et al.* (2007) established that, if noise in gene expression was incorporated, this could change the optimal promoter function. Again, this confirms our results found in **chapter 3**. In agreement with our results, Kalisky *et al.* (2007) observed that steep promoter functions lead to high amounts of noise. Therefore, using a very different approach, Kalisky *et al.* (2007) actually arrive at similar conclusions, namely that the precise shape of the promoter function can only be understood when considering that it has evolved in a fluctuating environment, while minimizing the effect of stochasticity in gene expression.

Except for evolving genetic regulation, there are also other mutational operators with which organisms can adapt their metabolism to a changing environment. In **part II** we studied how gene duplication and deletion, or even whole genome duplication can be used to adapt oneself to the environment. An interesting question is under which circumstances evolution will change the form of genetic regulation and under which circumstances gene duplication/gene deletion (and hence increased/decreased dosage) will be used.

There is one obvious difference between genetic regulation and gene duplication. Every gene has a certain maximal transcription rate. If higher transcription rates are needed, gene duplication will be favorable. Indeed, in **part I** we assumed a maximal transcription rate, while in **chapter 5** we assumed that if higher maximal fluxes through a certain reaction are needed, gene duplication would be favorable.

Because gene duplication/deletion is a very different kind of mutation than mutations that change genetic regulation (which most often involve the change of affinity of a regulatory protein to its binding site), the mutation rates may be very different. A priori one would expect that gene deletions/duplications are more rare than changes in the affinity of proteins. This is also clear from the very different time scales in these two processes, in the order of days or million of years! Therefore one could speculate that both processes allow for adaptation to the environment on different time scales. However, this does not appear to be the case, because Dunham *et al.* (2002) showed that extensive chromosomal duplications and deletions already occur in a few hundred generations.

Finally, there is even a longer time scale on which adaptations to the environment might occur. There is a long controversy whether directed mutations occur in evolution, e.g. Lenski & Mittler (1993); Brisson (2003). Interestingly, one of the first observations of directed mutation was in the *lac* operon (Cairns *et al.*, 1988). Furthermore, there is evidence that genomes can structure themselves such that beneficial duplications occur more often than others (Dunham *et al.*, 2002; Crombach & Hogeweg, 2007). If organisms also evolve their evolvability,

an even longer time scale is involved.

6.3 Future Directions

While our study of the *lac* operon is, we believe, quite definite, our work on the evolution of *S. cerevisiae* after its WGD is much more open-ended. As we mentioned, space could be an important factor in the evolution of *S. cerevisiae* and it would therefore be interesting to embed our model into a spatial environment.

However, in this case, this is not as easy as in the case of the *lac* operon. First of all, in the simulations in **chapter 5** a population size of 100 cells was used, which is too small to usefully put on a grid. Furthermore, when we would include space, we would also like to incorporate nutrient dynamics and hence the feedback between the environment and the cells. For this we would need to implement dynamic Flux Balance Analysis (Varma & Palsson, 1994; Mahadevan *et al.*, 2002), which would add an extra timescale to the model. Whether it is computationally feasible to add dynamic FBA into the already demanding model remains to be seen.

Adding space could also be useful for studying speciation after WGD. It is known that the WGD in yeast was followed by many speciation events (Scannell *et al.*, 2006). This was explained by reciprocal gene loss, which could lead to reproductive isolation. However, a WGD also leads to much evolutionary freedom, which could lead to specialization and possibly speciation. It would be interesting to study whether this effect in itself would be sufficient to lead to speciation, without reproductive isolation. Furthermore, adding sexual reproduction would then be a natural extension, to study the role of reciprocal gene loss.

Finally, as we already mentioned, it would be interesting to study how the evolution of gene regulation and metabolic network evolution (by gene duplication and deletion) would interact. In our thesis both these topics have only been studied separately. However, combining these two processes is probably only possible in a small network, instead of a genome-scale metabolic network. In this way it might be possible to understand under which circumstances evolution changes gene regulation and under which circumstances the metabolic or regulatory network.

6.4 Conclusion

In this thesis we have shown that combining detailed, quantitative modeling with an evolutionary, individual based approach is a promising way to study the evolution of biological systems. This approach has not only given us insight how, from an evolutionary point of view, such systems came about, but also how such systems work. Ignoring the evolutionary aspects of biological systems has, in the case of the *lac* operon, led to a wrong understanding of the dynamics of the system. Furthermore we have shown that even for such a large scale system as the metabolic network of *S. cerevisiae*, this approach leads to accurate predictions of the evolutionary outcome.

Bibliography

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M. & Postlethwait, J. H. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282: 1711–1714 (1998).

Andersen, K. B. & Von Meyenburg, K. Are growth rates of Escherichia coli in batch cultures limited by respiration? *J. Bacteriol.* 144: 114–123 (1980).

Andersson, J. O. & Andersson, S. G. Genome degradation is an ongoing process in Rickettsia. *Mol. Biol. Evol.* 16: 1178–1191 (1999a).

Andersson, J. O. & Andersson, S. G. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* 9: 664–671 (1999b).

Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A. M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J. & Wincker, P. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444: 171–178 (2006).

Babloyantz, A. & Sanglier, M. Chemical instabilities of "all-or-none" type in β - galactosidase induction and active transport. *FEBS Lett.* 23: 364–366 (1972).

Barkley, M. D., Riggs, A. D., Jobe, A. & Burgeois, S. Interaction of effecting ligands with lac repressor and repressor-operator complex. *Biochemistry.* 14: 1700–1712 (1975).

Barnett, J. A. A history of research on yeasts. 1: Work by chemists and biologists 1789-1850. *Yeast.* 14: 1439–1451 (1998).

Barnett, J. A. A history of research on yeasts 7: enzymic adaptation and regulation. *Yeast.* 21: 703–746 (2004).

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. O. & Herrgard, M. J. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.* 2: 727–738 (2007).

Beckwith, J. Escherichia coli *and* Salmonella, chapter The Operon: an Historical Account. ASM Press, Washington D.C. (1996).

Bilu, Y., Shlomi, T., Barkai, N. & Ruppin, E. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS. Comput. Biol.* 2: e106 (2006).

Blake, W. J., Balazsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., Cantor, C. R., Walt, D. R. & Collins, J. J. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol. Cell* 24: 853–865 (2006).

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S. & Van de Peer, Y. The gain and loss of genes during 600 million years of vertebrate evolution.

Genome Biol. 7: R43 (2006).

Blumenthal, T. Operons in eukaryotes. *Brief. Funct. Genomic. Proteomic.* 3: 199–211 (2004).

Boerlijst, M. C. & Hogeweg, P. Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. *Phys. D* 48: 17–28 (1991).

Bordenave, G. Louis Pasteur (1822-1895). *Microbes. Infect.* 5: 553–560 (2003).

Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438 (2003).

Brisson, D. The directed mutation controversy in an evolutionary context. *Crit. Rev. Microbiol.* 29: 25–35 (2003).

Burhans, D. T., Ramachandran, L., Wang, J., Liang, P., Patterton, H. G., Breitenbach, M. & Burhans, W. C. Non-random clustering of stress-related genes during evolution of the S. cerevisiae genome. *BMC. Evol. Biol.* 6: 58 (2006).

Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15: 1456–1461 (2005).

Byrne, K. P. & Wolfe, K. H. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic. Acids. Res.* 34: D452–D455 (2006).

Byrnes, J. K., Morris, G. P. & Li, W. H. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol. Biol. Evol.* 23: 1136–1143 (2006).

Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* 440: 358–362 (2006).

Cairns, J., Overbaugh, J. & Miller, S. The origin of mutants. *Nature* 335: 142–145 (1988).

Carlson, R. & Srienc, F. Fundamental Escherichia coli biochemical pathways for biomass and energy production: creation of overall flux states. *Biotechnol. Bioeng.* 86: 149–162 (2004).

Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* 34: 247–278 (1978).

Cavalier-Smith, T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot. (Lond).* 95: 147–175 (2005).

Chance, B., Garfinkel, D., Higgins, J. & Hess, B. Metabolic control mechanisms. 5. A solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells. *J. Biol. Chem.* 235: 2426–2439 (1960).

Cheng, B., Fournier, R. L., Relue, P. A. & Schisler, J. An experimental and theoretical study of the inhibition of Escherichia coli lac operon gene expression by antigene oligonucleotides. *Biotechnol. Bioeng.* 74: 220–229 (2001).

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. & Davis, R. W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2: 65–73 (1998).

Chung, J. D. & Stephanopoulos, G. On physiological multiplicity and population heterogeneity of biological systems. *Chem. Eng. Sci.* 51: 1509–1521 (1996).

Clarke, R.T.J. *Microbial Ecology of the Gut*, chapter The Gut and its Microorganisms. Academic Press Inc. London (1977).

Cole, J. Nitrate reduction to ammonia by enteric bacteria: redundancy, or a strategy for survival during oxygen starvation? *FEMS. Microbiol. Lett.* 136: 1–11 (1996).

Conant, G. C. & Wolfe, K. H. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol. Syst. Biol.* 3: 129 (2007).

Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J. theor. Biol.* 213: 73–88 (2001).

Crombach, A. & Hogeweg, P. Chromosome rearrangements and the evolution of genome structuring and adaptability. *Mol. Biol. Evol.* 24: 1130–1139 (2007).

Dagan, T., Blekhman, R. & Graur, D. The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Mol. Biol. Evol.* 23: 310–316 (2006).

Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS. Biol.* 3: e314 (2005).

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

Delmotte, F., Rispe, C., Schaber, J., Silva, F. J. & Moya, A. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. *BMC. Evol. Biol.* 6: 56 (2006).

DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680–686 (1997).

Dienert, F. Sur la fermentation du galactose et sur l'accoutumance des levures à ce sucre. *Ann Inst Pasteur* 14: 139–189 (1900).

Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., Wing, R. A., Flavier, A., Gaffney, T. D. & Philippsen, P. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 304: 304–307 (2004).

Duarte, N. C., Herrgard, M. J. & Palsson, B. O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 14: 1298–1309 (2004a).

Duarte, N. C., Palsson, B. O. & Fu, P. Integrated analysis of metabolic phenotypes in Saccharomyces cerevisiae. *BMC. Genomics.* 5: 63 (2004b).

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. & Botstein, D. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* 99: 16144–16149 (2002).

Edwards, J. S. & Palsson, B. O. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* 274: 17410–17416 (1999).

Edwards, J. S. & Palsson, B. O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* 97: 5528–5533 (2000).

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expres-

sion in a single cell. *Science* 297: 1183–1186 (2002).

Englesberg, E., Irr, J., Power, J. & Lee, N. Positive control of enzyme synthesis by gene C in the L-arabinose system. *J. Bacteriol.* 90: 946–957 (1965).

Famili, I., Forster, J., Nielsen, J. & Palsson, B. O. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. U.S.A.* 100: 13134–13139 (2003).

Fay, J. C. & Benavides, J. A. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS. Genet.* 1: 66–71 (2005).

Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 96: 9721–9726 (1999).

Forster, J., Famili, I., Fu, P., Palsson, B. O. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* 13: 244–253 (2003).

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. Noise minimization in eukaryotic gene expression. *PLoS. Biol.* 2: e137 (2004).

Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16: 805–814 (2006).

Furlong, R. F. & Holland, P. W. Were vertebrates octoploid? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 357: 531–544 (2002).

Gatenby, R. A. & Vincent, T. L. An evolutionary model of carcinogenesis. *Cancer Res.* 63: 6212–6220 (2003).

Gennis, R.B. & Stewart, V. *Escherichia coli and Salmonella*, chapter Respiration. ASM Press, Washington D.C. (1996).

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. Life with 6000 genes. *Science* 274: 546 (1996).

Gojkovic, Z., Knecht, W., Zameitat, E., Warneboldt, J., Coutelis, J. B., Pynyaha, Y., Neuveglise, C., Moller, K., Loffler, M. & Piskur, J. Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol. Genet. Genomics.* 271: 387–393 (2004).

Goldstein, M. L., Morris, S. A. & Yen, G. G. Problems with fitting to the power-law distribution. *Eur. Phys. J. B.* 41: 255–258 (2004).

Gomez-Valero, L., Silva, F. J., Christophe Simon, J. & Latorre, A. Genome reduction of the aphid endosymbiont Buchnera aphidicola in a recent evolutionary time scale. *Gene.* 389: 87–95 (2007).

Goodwin, B. C. Oscillatory behavior in enzymatic control processes. *Adv. Enzyme. Regul.* 3: 425–438 (1965).

Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76: 65–101 (2001).

Griffith, J. S. Mathematics of cellular control processes. I. Negative feedback to one gene. *J. theor. Biol.* 20: 202–208 (1968a).

Griffith, J. S. Mathematics of cellular control processes. II. Positive feedback to

one gene. *J. theor. Biol.* 20: 209–216 (1968b).

Hennaut, C., Hilger, F. & Grenson, M. Space limitation for permease insertion in the cytoplasmic membrane of Saccharomyces cerevisiae. *Biochem. Biophys. Res. Commun.* 39: 666–671 (1970).

Hermida, L., Brachat, S., Voegeli, S., Philippsen, P. & Primig, M. The Ashbya Genome Database (AGD)–a tool for the yeast community and genome biologists. *Nucleic. Acids. Res.* 33: D348–D352 (2005).

Herrgard, M. J., Lee, B. S., Portnoy, V. & Palsson, B. O. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res.* 16: 627–635 (2006).

Hogema, B. M., Arents, J. C., Bader, R. & Postma, P. W. Autoregulation of lactose uptake through the LacY permease by enzyme IIAGlc of the PTS in Escherichia coli K-12. *Mol. Microbiol.* 31: 1825–1833 (1999).

Huber, R. E., Kurz, G. & Wallenfels, K. A quantitation of the factors which affect the hydrolase and transgalactosylase activities of beta-galactosidase (E. coli) on lactose. *Biochemistry.* 15: 1994–2001 (1976).

Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldon, T. The human phylome. *Genome Biol.* 8: R109 (2007).

Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5: 299–310 (2004).

Huynen, M. A., Stadler, P. F. & Fontana, W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 93: 397–401 (1996).

Initiative, Arabidopsis Genome. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815 (2000).

Ippen, K., Miller, J. H., Scaife, J. & Beckwith, J. New controlling element in the Lac operon of E. coli. *Nature* 217: 825–827 (1968).

Jacob, F., Perrin, D., Sanchez, C. & Monod, J. [Operon: a group of genes with the expression coordinated by an operator.]. *C. R. Hebd. Seances. Acad. Sci.* 250: 1727–1729 (1960).

Jobe, A. & Bourgeois, S. lac Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.* 69: 397–408 (1972).

Joshi, A. & Palsson, O. Metabolic dynamics in the human red cell part i-a comprehensive kinetic model. *J. theor. Biol.* 141: 515–528 (1989).

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

Kashiwagi, A., Urabe, I., Kaneko, K. & Yomo, T. Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. *PLoS. ONE.* 1: e49 (2006).

Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N. & Small, P. M. Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res.* 11: 547–554 (2001).

Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428: 617–624 (2004).

Kennell, D. & Riezman, H. Transcription and translation initiation frequencies of the Escherichia coli lac operon. *J. Mol. Biol.* 114: 1–21 (1977).

Kierzek, A. M., Zaim, J. & Zielenkiewicz, P. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J. Biol. Chem.* 276: 8165–8172 (2001).

Koch, A. L. The protein burden of lac operon products. *J. Mol. Evol.* 19: 455–462 (1983).

Kooijman, S. A., Muller, E. B. & Stouthamer, A. H. Microbial growth dynamics on the basis of individual budgets. *Antonie. Van. Leeuwenhoek.* 60: 159–174 (1991).

Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res.* 15: 1421–1430 (2005).

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

Lafontaine, I., Fischer, G., Talla, E. & Dujon, B. Gene relics in the genome of the yeast Saccharomyces cerevisiae. *Gene.* 335: 1–17 (2004).

Lee, S. B. & Bailey, J. E. Genetically structured models for *lac* promoter-operator function in the *Escherichia coli* chromosome and in multicopy plasmids. *Biotechnol. Bioeng.* 26: 1372–1389 (1984).

Lenski, R. E. & Mittler, J. E. The directed mutation controversy and neo-Darwinism. *Science* 259: 188–194 (1993).

Lin, Y. S., Byrnes, J. K., Hwang, J. K. & Li, W. H. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. *Proc. Natl. Acad. Sci. U.S.A.* 103: 14412–14416 (2006).

Liti, G. & Louis, E. J. Yeast evolution and comparative genomics. *Annu. Rev. Microbiol.* 59: 135–153 (2005).

Lynch, M. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* 60: 327–349 (2006).

Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155 (2000).

Lynch, M., Koskella, B. & Schaack, S. Mutation pressure and the evolution of organelle genomic architecture. *Science* 311: 1727–1730 (2006).

Mahadevan, R., Edwards, J. S. & Doyle, F. J. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys. J.* 83: 1331–1340 (2002).

Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5: 264–276 (2003).

Malan, T. P., Kolb, A., Buc, H. & McClure, W. R. Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. *J. Mol. Biol.* 180: 881–909 (1984).

Malan, T. P. & McClure, W. R. Dual promoter control of the Escherichia coli lactose operon. *Cell* 39: 173–180 (1984).

Maniloff, J. Evolution of wall-less prokaryotes. *Annu. Rev. Microbiol.* 37: 477–499 (1983).

Maquat, L. E. & Reznikoff, W. S. In vitro analysis of the Escherichia coli RNA polymerase interaction with wild-type and mutant lactose promoters. *J. Mol. Biol.* 125: 467–490 (1978).

Martinez-Bilbao, M., Holdsworth, R. E., Edwards, L. A. & Huber, R. E. A highly

reactive beta-galactosidase (Escherichia coli) resulting from a substitution of an aspartic acid for Gly-794. *J. Biol. Chem.* 266: 4979–4986 (1991).

McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 94: 814–819 (1997).

Merico, A., Sulo, P., Piskur, J. & Compagno, C. Fermentative lifestyle in yeasts belonging to the Saccharomyces complex. *FEBS J.* 274: 976–989 (2007).

Meugnier, E., Rome, S. & Vidal, H. Regulation of gene expression by glucose. *Curr. Opin. Clin. Nutr. Metab. Care.* 10: 518–522 (2007).

MIPS (http://mips.gsf.de/genre/proj/yeast/index.jsp/).

Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17: 589–596 (2001).

Monod, J. *Recherches sur la croissance des cellules bactériennes.* Hermann & Cie, Paris (1942).

Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol.* 2: RESEARCH0054 (2001).

Morgenstern, B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic. Acids. Res.* 32: W33–W36 (2004).

Mortimer, R. K. Evolution and variation of the yeast (Saccharomyces) genome. *Genome Res.* 10: 403–409 (2000).

Müller-Hill, B. *The lac operon*. Walter de Gruyter, Berlin (1996).

Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res.* 12: 1370–1376 (2002).

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. & Andersson, D. I. Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U.S.A.* 102: 12112–12116 (2005).

Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. U.S.A.* 43: 553–566 (1957).

Ochman, H. & Jones, I. B. Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J.* 19: 6637–6643 (2000).

Ohno, S. *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg (1970).

Ovadi, J. & Saks, V. On the origin of intracellular compartmentation and organized metabolic systems. *Mol. Cell Biochem.* 256: 5–12 (2004).

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nat. Genet.* 31: 69–73 (2002).

Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden, A. Multistability in the lactose utilization network of Escherichia coli. *Nature* 427: 737–740 (2004).

Ozcan, S. & Johnston, M. Function and regulation of yeast hexose transporters. *Microbiol. Mol. Biol. Rev.* 63: 554–569 (1999).

Pagie, L. & Hogeweg, P. Information integration and red queen dynamics in coevolutionary optimization. In *Proceedings of the 2000 Congress on Evolutionary Computation CEC00*, pp. 1260–1267. IEEE Press, La Jolla Marriott Hotel La Jolla, California, USA (2000).

Pal, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G. & Hurst, L. D. Chance

and necessity in the evolution of minimal metabolic networks. *Nature* 440: 667–670 (2006).

Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429: 661–664 (2004).

Pardee, A. B., Jaboc. F. & Monod, J. The genetic control and cytoplasmic expression of "inducibility" in the synthesis of β-glactosidase by *E. coli*. *J. Mol. Biol.* 1: 165–178 (1959).

Patil, K. R., Rocha, I., Forster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC. Bioinformatics* 6: 308 (2005).

Pfeiffer, T., Schuster, S. & Bonhoeffer, S. Cooperation and competition in the evolution of ATP-producing pathways. *Science* 292: 504–507 (2001).

Piskur, J., Rozpedowska, E., Polakova, S., Merico, A. & Compagno, C. How did Saccharomyces evolve to become a good brewer? *Trends Genet.* 22: 183–186 (2006).

Postma, P. W., Lengeler, J. W. & Jacobson, G. R. Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiol. Rev.* 57: 543–594 (1993).

Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* 304: 1811–1814 (2004).

Reznikoff, W. S. The lactose operon-controlling elements: a complex paradigm. *Mol. Microbiol.* 6: 2419–2422 (1992).

Reznikoff, W. S., Winter, R. B. & Hurley, C. K. The location of the repressor binding sites in the lac operon. *Proc. Natl. Acad. Sci. U.S.A.* 71: 2314–2318 (1974).

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A. & Collado-Vides, J. Regulondb (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic. Acids. Res.* 34: D394–D397 (2006).

Santillan, M. & Mackey, M. C. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. *Biophys. J.* 86: 1282–1292 (2004).

Santillan, M., Mackey, M. C. & Zeron, E. S. Origin of bistability in the lac Operon. *Biophys. J.* 92: 3830–3842 (2007).

Savageau, M. A. Genetic regulatory mechanisms and the ecological niche of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* 71: 2453–2455 (1974).

Savageau, M. A. Design of molecular control mechanisms and the demand for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 74: 5647–5651 (1977).

Savageau, M. A. Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos.* 11: 142–159 (2001).

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440: 341–345 (2006).

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S. & Palsson, B. O. Genome-scale metabolic model of Helicobacter pylori 26695. *J. Bacteriol.* 184: 4582–4593 (2002).

Schmitz, A. Cyclic AMP receptor proteins interacts with lactose operator DNA. *Nucleic. Acids. Res.* 9: 277–292 (1981).

Schrödinger, E. *What is Life?* Cambridge University Press, Cambridge (1944).

Semon, M. & Wolfe, K. H. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* 23: 108–112 (2007).

Seoighe, C. & Wolfe, K. H. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2: 548–554 (1999).

Setty, Y., Mayo, A. E., Surette, M. G. & Alon, U. Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci. U.S.A.* 100: 7702–7707 (2003).

SGD (http://www.yeastgenome.org/).

Shapiro, J. A. Genome organization, natural genetic engineering and adaptive mutation. *Trends Genet.* 13: 98–104 (1997).

Shlomi, T., Eisenberg, Y., Sharan, R. & Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* 3: 101 (2007).

Soule, C. Graphic requirements for multistationarity. *ComPlexUs* 1: 123–133 (2003).

Soule, C. Mathematical approaches to differentiation and gene regulation. *C. R. Biol.* 329: 13–20 (2006).

Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99: 12795–12800 (2002).

Taylor, J. S., Van de Peer, Y., Braasch, I. & Meyer, A. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356: 1661–1679 (2001).

Thattai, M. & Van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.* 98: 8614–8619 (2001).

Thattai, M. & Van Oudenaarden, A. Stochastic gene expression in fluctuating environments. *Genetics.* 167: 523–530 (2004).

Thomas, B. C., Pedersen, B. & Freeling, M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16: 934–946 (2006).

Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids. Res.* 22: 4673–4680 (1994).

Van Dedem, G. & Moo-Young, M. A model for diauxic growth. *Biotechnol. Bioeng.* 17: 1301–1312 (1975).

Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* 60: 3724–3731 (1994).

Vilar, J. M., Guet, C. C. & Leibler, S. Modeling network dynamics: the lac operon, a case study. *J. Cell Biol.* 161: 471–476 (2003).

Vilar, J. M. & Leibler, S. DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.* 331: 981–989 (2003).

Winge, O. On haplophase and diplophase in some *Saccharomycetes*. *C.R. Trav. Lab. Carlsberg Ser. Physiol.* 21: 77–111 (1935).

Woese, C. R. Bacterial evolution. *Microbiol. Rev.* 51: 221–271 (1987).

Wolf, D. M., Vazirani, V. V. & Arkin, A. P. Diversity in times of adversity: probabilistic strategies in microbial survival games. *J. theor. Biol.* 234: 227–253 (2005).

Wolfe, K. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol.* 14: R392–R394 (2004).

Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708–713 (1997).

Wong, P., Gladney, S. & Keasling, J. D. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol. Prog.* 13: 132–143 (1997).

Yildirim, N. & Mackey, M. C. Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophys. J.* 84: 2841–2851 (2003).

Yildirim, N., Santillan, M., Horike, D. & Mackey, M. C. Dynamics and bistability in a reduced model of the lac operon. *Chaos.* 14: 279–292 (2004).

YMPD (http://bmerc.bu.edu/projects/mito/).

Zubay, G., Schwartz, D. & Beckwith, J. Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proc. Natl. Acad. Sci. U.S.A.* 66: 104–110 (1970).

Color Plates From Chapter 3

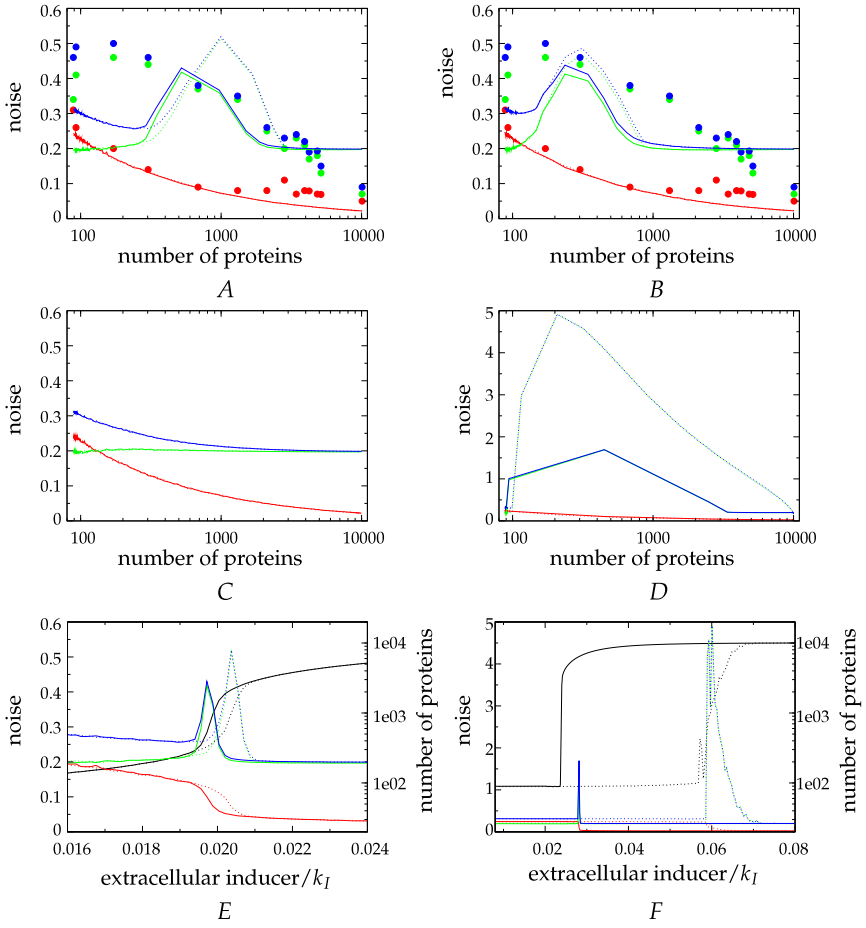


Figure 3.1 on page 54

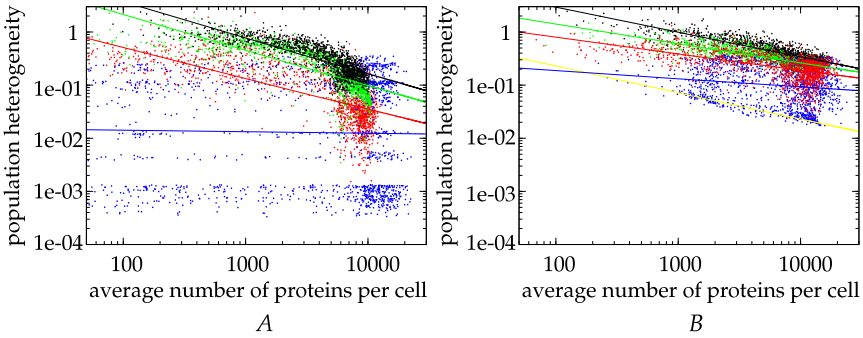


Figure 3.2 on page 57

Color Plates From Chapter 5

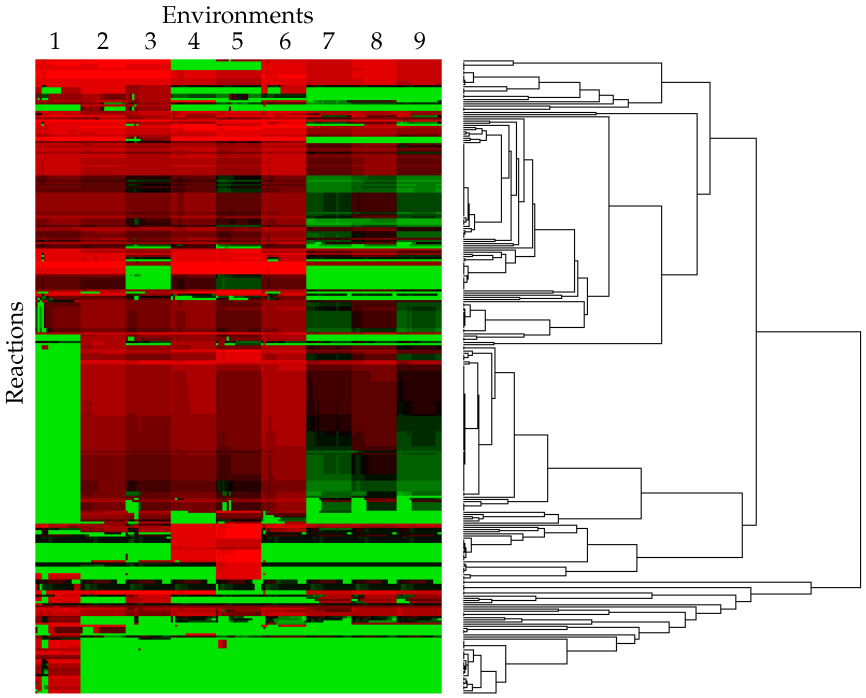


Figure 5.5 on page 115

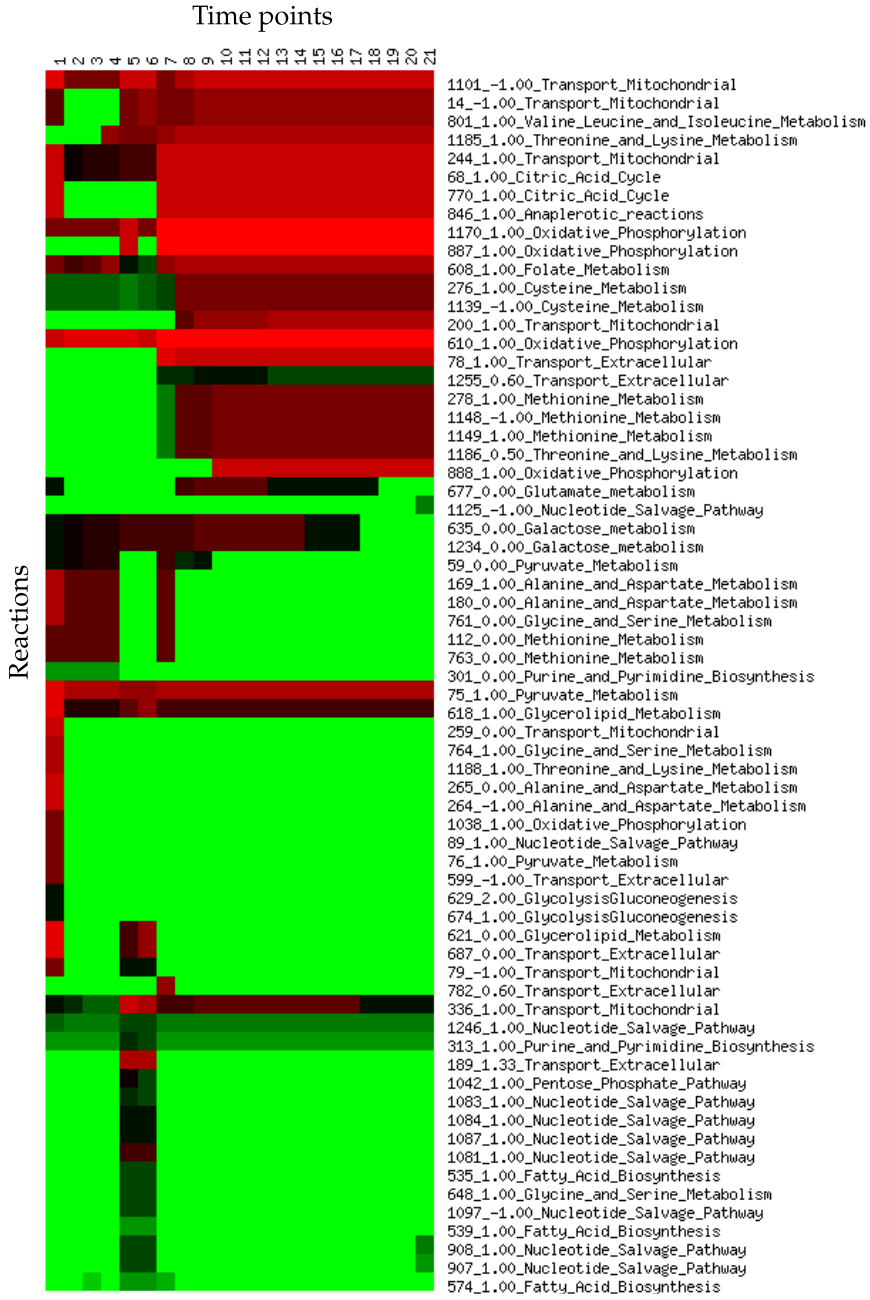


Figure 5.6 on page 116

Samenvatting

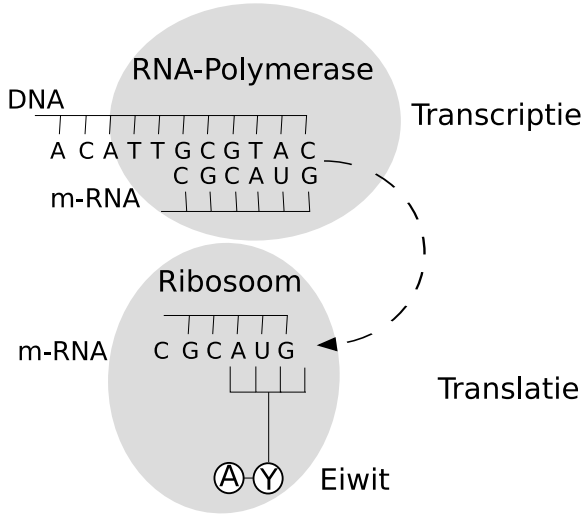
Metabolisme, oftewel stofwisseling, is één van de belangrijkste eigenschappen van het leven. Organismen eten om de energie te genereren die nodig is om te leven en om de bouwstenen te kunnen maken om te groeien. Het is opvallend hoe het metabolisme van totaal verschillende organismen toch zoveel op elkaar lijkt. Zowel mensen, planten, gisten en bacteriën gebruiken dezelfde methode om suiker te verbranden: de citroenzuurcyclus. "Het wiel" is dus maar één keer uitgevonden in de evolutie en alleen van soort tot soort aangepast aan de specifieke omstandigheden. Voor biologen maakt dit het leven een stuk eenvoudiger. Door het metabolisme van een paar soorten heel goed te bestuderen, kunnen we toch veel leren over metabolisme in het algemeen.

In dit proefschrift hebben we het metabolisme van twee ééncelligen bestudeerd. De eerste is *Escherichia coli* (*E. coli*), een darmbacterie die enigszins berucht is omdat hij buikgriep (gastro-enteritis) kan veroorzaken. Wat minder mensen weten is dat *E. coli* bij iedereen voorkomt en dat er maar relatief weinig soorten *E. coli* zijn die voor problemen zorgen. *E. coli* is een model-organisme voor biologen en is daarom erg goed bestudeerd, zo ook het metabolisme van *E. coli*.

Het tweede organisme dat we bestudeerd hebben is *Saccharomyces cerevisiae* (*S. cerevisiae*), oftewel bakkersgist. *S. cerevisiae* is ook een ééncellige. In tegenstelling tot *E. coli* is bakkersgist geen bacterie maar behoort het tot de schimmels. *S. cerevisiae* wordt al duizenden jaren door mensen gebruikt om bier, wijn en brood te produceren. Ook *S. cerevisiae* is een zeer goed bestudeerd model-organisme, waarvan het metabolisme goed bekend is.

Cellen en organismen passen hun metabolisme continu aan aan de veranderende omgeving. Dit gebeurt op twee verschillende manieren. De eerste manier is door middel van metabolische regulatie. Een bekend voorbeeld hiervan is de aanpassing van het menselijk lichaam aan de veranderende bloedsuikerspiegel. Als de bloedsuikerspiegel stijgt gaat de alvleesklier insuline uitscheiden. De insuline stimuleert afbraak van glucose en zo blijft de bloedsuikerspiegel min of meer constant. Dit is een voorbeeld van metabolische regulatie. Het is duidelijk dat metabolische regulatie belangrijk is voor het overleven van het organisme: als de regulatie van de bloedsuikerspiegel faalt, is suikerziekte het gevolg.

De tweede manier waarop organismen hun metabolisme aanpassen aan een veranderende omgeving is via evolutie. Evolutie wordt vaak gezien als een proces dat zich "lang geleden" afspeelde. Minder bekend is dat evolutie ook een proces is dat zich op veel kortere tijdschalen afspeelt. In laboratoria bijvoorbeeld laat men bacteriën en gistcellen in enkele honderden generaties evolueren. In een paar weken zijn de bacteriën of gistcellen dan al veel beter aangepast aan de laboratoriumomgeving.



Genetica: Voor een beter begrip van de rest van deze samenvatting, geef ik hier een korte inleiding in de genetica. In bovenstaande figuur staat een schematische weergave van de beschreven processen. De erfelijke eigenschappen van organismen zitten opgeslagen in het genoom. Het genoom van alle organismen bestaat uit DNA. DNA is een molecuul dat gevormd wordt door een keten bestaande uit vier verschillende "baseparen" (A, G, C en T). Baseparen vormen de letters van de genetische taal. Genen zijn een serie baseparen die coderen voor een bepaalde eigenschap van een organisme. Of preciezer, elk gen codeert voor het aanmaken van een specifiek eiwit. Eiwitten zijn moleculen die allerlei functies in de cel kunnen vervullen. Sommigen versnellen chemische reacties, anderen zijn belangrijk voor de structuur van de cel en weer anderen zorgen voor het transport van bepaalde stoffen. Er is ook een bepaald eiwit (RNA-polymerase) dat het DNA afleest en zo RNA produceert (transcriptie). RNA is een molecuul dat veel op DNA lijkt en dat onder andere wordt gebruikt om de informatie uit het DNA te verspreiden door de cel. Dit stuk RNA wordt vervolgens in de ribosomen (weer een complex van eiwitten) vertaald tot een eiwit (translatie).

In dit proefschrift hebben wij de aanpassing van cellen aan een veranderende omgeving, zowel door middel van metabolische regulatie als evolutie, bestudeerd. Dit hebben we gedaan vanuit twee tegenovergestelde oogpunten. In **deel I** van dit proefschrift hebben we ons geconcentreerd op het *lac* operon van *E. coli*. Het *lac* operon bestaat uit drie genen en is dus een klein onderdeel van het metabolisme van *E. coli*. In **deel II** bekeken we de evolutie van het hele metabolisch netwerk van bakkersgist, bestaand uit 750 genen.

De mens heeft ongeveer 28000 genen, bakkersgist ongeveer 6000 en *E. coli* ruim 4000. Al die genen produceren niet altijd en overal eiwitten. In de jaren vijftig en zestig werd ontdekt hoe de activiteit van genen gereguleerd werd. Jacob en Monod bestudeerden hoe *E. coli* de genen die verantwoordelijk zijn voor lactose (melksuiker) metabolisme reguleert en wonnen daarvoor de Nobelprijs voor de fysiologie en geneeskunde in 1965.

Het genetisch systeem dat dit reguleerde noemden zij het *lac* operon. Zij vonden dat de twee genen die zorgen voor lactose metabolisme alleen geactiveerd

worden als er lactose in de omgeving is. Bovendien vonden zij dat als er glucose aanwezig is in de omgeving van de cel, dat die twee genen *niet* actief zijn. Blijkbaar consumeert *E. coli*, als het de keus heeft, liever glucose dan lactose. Om iets preciezer te zijn, de activiteit van het *lac* operon wordt bepaald door twee variabelen, glucose en lactose, en is maximaal als de lactose concentratie hoog is en de glucose concentratie laag (zie figuur 2.4 *B* voor de experimenteel gemeten functie).

De vraag was nu hoe *E. coli* dit voor elkaar krijgt. Ook dit werd uitgevonden door Jacob en Monod en dit model van genetische regulatie bleek een goed algemeen beeld te geven hoe genetische regulatie werkt. Daarom is het *lac* operon het tekstboekvoorbeeld van genetische regulatie geworden.

In figuur 6.1 geven we een schematische weergave van de genetische regulatie van het *lac* operon. De twee genen die verantwoordelijk zijn voor lactose metabolisme hebben allebei hun eigen taak. Eén gen codeert voor een eiwit dat lactose de cel in transporteert (permease). Het andere gen codeert voor een eiwit (β-galactosidase) dat lactose omzet in glucose en galactose, die dan verder verbrand kunnen worden. Hoe werkt dit nu? Om een gen af te lezen moet het eiwit RNA-polymerase binden aan het DNA. Als er geen lactose aanwezig is in de cel, bindt er een ander eiwit (repressor) aan het DNA, zodat er geen plaats meer is voor RNA-polymerase. Zodoende wordt dit gedeelte van het DNA niet meer afgelezen als er geen lactose in de cel zit. Omdat de twee genen van het *lac* operon naast elkaar op het DNA zitten, worden ze zo altijd tegelijkertijd geactiveerd of gedeactiveerd. Zo kan *E. coli* de genen die verantwoordelijk zijn voor het lactose metabolisme alleen activeren als ze nodig zijn. Later werd ook gevonden dat genen kunnen worden geactiveerd door eiwitten in plaats van gedeactiveerd, zoals in het *lac* operon. Maar in zijn algemeenheid, bleef het model van Jacob en Monod staan, zeker voor bacteriën.

In het geval van het *lac* operon zit er echter nog een addertje onder het gras. Zoals eerder uitgelegd, bindt het repressor eiwit aan het DNA als er geen lactose in de cel zit (lactose bindt namelijk aan dit repressor eiwit en deactiveert het). Maar om lactose de cel in te krijgen moet het *lac* operon actief zijn (het permease eiwit zorgt namelijk voor het transport van lactose de cel in). Hoe kan de cel dan ooit beginnen met lactose metabolisme? Blijkbaar heeft de cel lactose in de cel nodig om het *lac* operon te activeren, maar andersom is een actief *lac* operon nodig om lactose de cel in te krijgen, een voorbeeld van positieve terugkoppeling. Op deze manier blijft het operon actief als het al actief is en blijft het inactief als het al inactief is. Afhankelijk van de voorgeschiedenis van de cel zal de cel zich dus anders gedragen. Er zijn dus twee stabiele evenwichten mogelijk, een effect dat bistabiliteit wordt genoemd. Dit probleem wordt opgelost doordat het *lac* operon altijd een beetje actief is, dus altijd een paar eiwitten produceert die het proces dan verder op gang kunnen brengen.

Toch is er in experimenten met het *lac* operon gezien dat als het *lac* operon geactiveerd wordt met kunstmatige activators, onder bepaalde concentraties van die stoffen in de omgeving, het *lac* operon inderdaad bistabiel is. Eerder heeft men dit fenomeen geprobeerd te verklaren vanuit evolutionair oogpunt. Als je aanneemt dat cellen af en toe kunnen switchen van een *lac* operon dat actief is

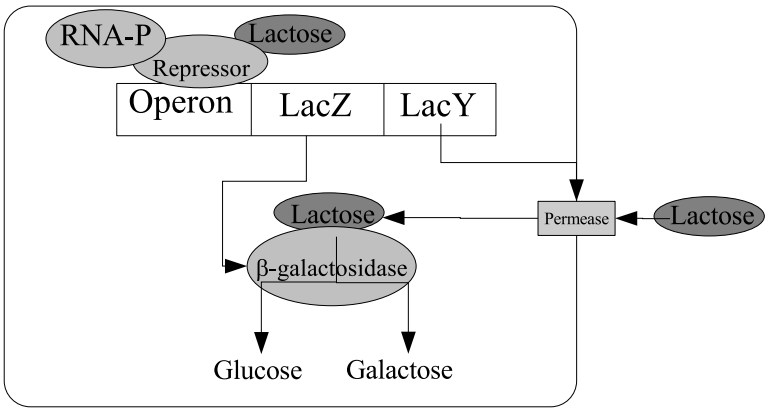


Figure 6.1: Een overzicht van de werking van het *lac* operon. Lactose wordt de cel in getransporteerd door permease. Eenmaal in de cel wordt het afgebroken door β-galactosidase. De productie van deze twee eiwitten wordt gereguleerd door het *lac* operon. Als er geen lactose in de cel aanwezig bindt het repressor-eiwit aan het operon. Dit verhindert het binden van RNA-polymerase met het operon en op deze manier transcriptie van de genen die coderen voor permease en β-galactosidase.

naar een *lac* operon dat inactief is (wat inderdaad het geval is, omdat cellen heel klein zijn en zich dus enigszins stochastisch (willekeurig) gedragen), kan dit zorgen voor variatie in de populatie cellen. Sommige cellen zullen een actief *lac* operon hebben, anderen een inactief. Dit kan uit evolutionair oogpunt een voordeel zijn, omdat zo altijd een gedeelte van de populatie het goede doet (als een soort van voorzorgsmaatregel).

Tot zover de beschrijving over hoe het *lac* operon werkt. In **deel I** van dit proefschrift bestudeerden we de evolutie van het *lac* operon door middel van computersimulaties. We ontwikkelden een ruimtelijk, individueel georiënteerd model van *E. coli* cellen. De dynamica van de individuele cellen (zoals de glucose en lactose concentraties binnen en buiten de cel) werden gemodelleerd met differentiaalvergelijkingen. De cellen nemen glucose en lactose op uit de omgeving, groeien aan de hand hiervan, vermenigvuldigen zich en sterven. Elke cel heeft zo een eigen *lac* operon dat de dynamica van de cel bepaalt. Als een cel zich vermenigvuldigt, kan de cel muteren en zo het *lac* operon een beetje veranderen en cellen met het best functionerende *lac* operon zullen overleven.

Op deze manier konden we de experimentele functie van het *lac* operon verrassend goed reproduceren (zie figuur 2.4 *A* en 2.4 *B*). Inderdaad vonden we dat als er geen lactose in de cel aanwezig is, het *lac* operon nog steeds een relatief hoge activiteit heeft. Dit voorkomt het probleem dat we eerder bespraken, namelijk dat cellen niet kunnen switchen van een inactief naar een actief operon. Inderdaad laten we zien dat deze relatief hoge activiteit als de lactose concentratie laag

is, voor een snelle switch zorgt als de omgeving verandert. Dit is cruciaal voor de overlevingskansen van het organisme.

Verder laten we zien dat het *lac* operon dat evolueert in onze simulaties, als het geactiveerd wordt met lactose, zich anders gedraagt als in de experimenten en niet bistabiel is. Hoe kan dit verschil verklaard worden? Daarvoor moeten we ons realiseren dat in deze experimenten niet lactose, maar een kunstmatige activator gebruikt is om het *lac* operon te activeren. Deze stof wordt niet afgebroken door β-galactosidase en heeft daardoor een sterkere positieve terugkoppeling dan lactose en leidt daarom eerder tot bistabiliteit. Inderdaad vinden we dat als we het in onze simulaties geëvolueerde *lac* operon met een dergelijke kunstmatige activator activeren, we *wel* bistabiliteit zien, in tegenstelling tot wanneer het met lactose wordt geactiveerd. Dit is dus wel in overeenstemming met de experimenten. Verder is er recentelijk gemeten hoe het *lac* operon zich precies gedraagt als het geactiveerd wordt met lactose. De verwachtte bistabiliteit kon niet worden waargenomen, mogelijk door fouten in de metingen. Onze resultaten laten nu zien dat we niet moeten verwachten dat het om een meetfout gaat, maar dat het *lac* operon zich niet bistabiel gedraagt als het geactiveerd wordt met lactose, terwijl het zich wel bistabiel gedraagt als het geactiveerd wordt met een kunstmatige activator.

In **hoofdstuk 3** bekeken we de evolutie van het *lac* operon in een stochastisch model van genregulatie. In **hoofdstuk 2** hadden we namelijk aangenomen dat cellen zich volledig voorspelbaar (deterministisch) gedragen, wat niet realistisch is. We zien dat, als we aannemen dat cellen stochastische entiteiten zijn, dat dit effect heeft op de evolutie van de cellen. De cellen evolueren dan namelijk naar een nog hogere minimale activiteit. Deze hoge minimale activiteit (de productie van eiwitten als er geen lactose in de omgeving is) zorgt er niet alleen voor dat, net zoals in **hoofdstuk 2**, de cellen niet bistabiel zijn. Het zorgt er ook voor dat het effect van de stochasticiteit wordt geminimaliseerd. Stochasticiteit is namelijk het sterkst als de aantallen eiwitten laag zijn, dus als de activiteit van het operon laag is. Zo zorgt een hoge minimale activiteit voor een vermindering in stochasticiteit. Op deze manier laten we zien dat stochasticiteit in het geval van het *lac* operon ongunstig is en dus niet, zoals werd aangenomen, wordt gebruikt om een hogere variatie in de populatie te krijgen. Verder hebben we laten zien dat genetische variatie en ruimtelijke variatie een zeker zo grote rol spelen in populatie variatie als stochasticiteit in genregulatie.

In **deel II** bekeken we de evolutie van metabolische regulatie vanuit een heel ander perspectief. Metabolische reacties staan niet op zichzelf. Een stof die door de ene reactie wordt geproduceerd, wordt door een volgende reactie weer gebruikt. Zo staan alle stoffen en reacties in verbinding met elkaar en vormen zo een "metabolisch netwerk" (zie figuur 1.1 *A* voor een klein deel van het metabolisch netwerk van gist). In **deel II** van dit proefschrift bestudeerden we de evolutie van een zo volledig mogelijk metabolische netwerk, dit in contrast met **deel I**, waarin we de evolutie van maar een heel klein onderdeel van het metabolisme van *E. coli* bekeken.

Als specifiek voorbeeld bestudeerden we de evolutie van het metabolische netwerk van bakkersgist na genoomduplicatie. Het genoom van bakkersgist

heeft zich ongeveer 100 miljoen jaar geleden verdubbeld. Dit is een extreem soort mutatie, waarbij alle chromosomen zich verdubbelen en een organisme een dubbel genoom overhoudt. Dit kan gebeuren bij een fout in de celdeling. Niet alleen in gist, maar ook in andere soorten is een genoomduplicatie aangetoond, zoals in veel planten, vissen, het pantoffeldiertje en ook de mens. Het is nog onduidelijk wat het effect van een genoomduplicatie is op het functioneren en de evolutie van een soort. Wel is bekend dat genoomduplicatie altijd gevolgd wordt door massaal genverlies en dat zo uiteindelijk het totale aantal genen maar weinig verandert.

In **hoofdstuk 4** hebben we laten zien dat het genverlies na de genoomduplicatie niet alleen veroorzaakt wordt door het één voor één verliezen van genen. Het patroon van verloren genen kan alleen worden verklaard als we aannemen dat genen soms ook tegelijkertijd verloren raken en dat er dus grotere mutaties nodig zijn. Dit soort mutaties vinden plaats op het niveau van baseparen in plaats van genen. Zo kunnen grote mutaties op het niveau van baseparen ervoor zorgen dat meerdere genen tegelijkertijd verdwijnen. Een verwijdering van een reeks baseparen heeft natuurlijk een grotere kans om een volgend gen te bereiken als het aantal baseparen tussen de genen in klein is (genen liggen namelijk niet precies naast elkaar maar worden gescheiden door stukjes DNA). Inderdaad hebben we kunnen laten zien dat naburige genen die weinig baseparen tussen zich in hebben vaker samen verloren zijn gegaan. Op dezelfde manier hebben we aangetoond dat grote genen vaker op zichzelf verwijderd worden dan kleine genen.

Deze methode hebben we ook kunnen gebruiken om massaal genverlies bij de bacterie *Buchnera aphidicola* (*B. aphidicola*) te bestuderen. Deze bacterie leeft in de cellen van de bladluis. Zowel de bladluis als de bacterie heeft hier baat bij. Omdat de bladluis veel van de metabolische taken zelf doet, kon de bacterie veel genen verliezen, omdat ze overbodig waren geworden. In *B. aphidicola* is het patroon van genverlies aanzienlijk anders dan in gist. In deze bacterie komt het veel vaker voor dat grote groepen naburige genen samen verloren zijn gegaan. We laten zien dat het feit dat het genoom van bacteriën veel meer geordend is dan van gist, de verklaring kan zijn voor deze observatie. Zoals we eerder al zagen zitten genen met vergelijkbare functies in bacteriën vaak naast elkaar in operons, in tegenstelling tot bij eukaryoten (organismen met celkern, zoals ook gist). Dit kan de verklaring zijn waarom genen in *B. aphidicola* vaker in groepen verloren zijn dan in gist.

Na een genoomduplicatie kan het metabolisch netwerk natuurlijk veranderen. In **hoofdstuk 5** hebben wij een model ontwikkeld waarmee we het effect van een genoomduplicatie op een metabolisch netwerk kunnen beschrijven. Zo kunnen we de evolutie van het metabolisch netwerk na een genoomduplicatie simuleren. Inderdaad zien we dat een genoomduplicatie wordt gevolgd door massaal genverlies. We kunnen ook redelijk goed voorspellen welke genen dubbel behouden blijven en welke niet. Ook kunnen we voorspellen dat een genoomduplicatie in bakkersgist leidt tot een verhoging van de maximale flux door de glycolyse, wat er op neer komt dat bakkersgist daardoor suiker vlugger kan metaboliseren. Tenslotte hebben we laten zien dat een genoomduplicatie in een nieuwe omgeving direct kan leiden tot een verhoging van de fitness.

Al deze resultaten bevestigen de theorie dat de genoomduplicatie van gist

heeft geholpen bij de aanpassing van gist aan zijn omgeving. De genoomduplicatie van gist heeft rond dezelfde tijd plaatsgevonden als het ontstaan van de bloeiende planten en dus het ontstaan van fruit. Rottend fruit is de natuurlijk leefomgeving van gist. Zoals we allemaal weten is fruit een erg suikerrijke leefomgeving en dus is voor gist maximalisatie van de suikeropname belangrijker dan het efficiënt gebruiken van die suikers. Omdat de genoomduplicatie heeft kunnen leiden tot een verhoging van de suikeropname, bevestigt ons onderzoek de hypothese dat de genoomduplicatie 100 miljoen jaar geleden gist heeft geholpen bij het aanpassen aan de nieuwe omgeving van rottend fruit.

Al met al hebben we laten zien dat het combineren van gedetailleerd, kwantitatief modelleren van biologische systemen met een evolutionair, individueel georiënteerde aanpak, een veelbelovende manier is om de evolutie van biologische systemen te bestuderen. Deze aanpak heeft ons niet alleen veel geleerd over hoe dit soort systemen geëvolueerd zijn, maar ook hoe ze functioneren. We hebben gezien dat het negeren van de evolutionaire aspecten in het geval van het *lac* operon heeft geleid tot een verkeerd begrip van de dynamica van het *lac* operon, namelijk de voorspelling dat het *lac* operon bistabiel zou zijn. Verder hebben we laten zien dat zelfs voor een grootschalig systeem als het hele metabolische netwerk van gist, deze aanpak leidt tot accurate voorspellingen over de uitkomst van evolutie.

Curriculum Vitæ

De auteur van dit proefschrift, Milan Johannes Adrianus van Hoek, werd op 28 juni 1978 geboren in Tilburg. Vanaf 1990 was hij leerling op het Sint Oduplhuslyceum te Tilburg. In mei 1996 behaalde hij daar zijn gymnasium diploma. In september 1996 begon hij met de studie natuur- en sterrenkunde aan de Universiteit Utrecht. In 1998 behaalde hij daar zijn propedeuse in wiskunde en natuurkunde. Hij koos als specialisatie voor de richting theoretische natuurkunde en studeerde in september 2002 af bij Prof. dr. H.T.C. Stoof met een scriptie getiteld "Scattering laserlight from an ultracold atomic Fermi gas". Na zijn studie ging hij op zoek naar een promotieplek in de biologie en kwam terecht bij de vakgroep theoretische biologie/bioinformatica van Prof. dr. P. Hogeweg aan de Universiteit Utrecht. Daar volgde hij de vakken bioinformatische patroonanalyse en bioinformatische processen. Vanaf juni 2003 was hij als assistent in opleiding (AIO) werkzaam bij deze vakgroep, begeleid door Prof. dr. P. Hogeweg. De resultaten van dit onderzoek staan beschreven in dit proefschrift.

The author of this thesis, Milan Johannes Adrianus van Hoek, was born on June 28th, 1978 in Tilburg. From 1990 onwards he attended the Sint Odulphuslyceum in Tilburg. In may 1996 he obtained his gymnasium diploma there. In september 1996 he started his study in physics at the Utrecht University. In 1998 he obtained his propedeuse in mathematics and physics there. He chose to specialize in theoretical physics and he graduated in september 2002 under the supervision of Prof. dr. H.T.C. Stoof with a thesis titled "Scattering laserlight from an ultracold atomic Fermi gas". After his graduation, he looked for a doctoral position in biology, which he found in the theoretical biology/bioinformatics group of Prof. dr. P. Hogeweg at Utrecht University. There he followed courses in bioinformatic pattern analysis and bioinformatics processes. In june 2003 he started his doctoral research in this group, supervised by Prof. dr. P. Hogeweg. The results of this research are described in this thesis.

List of Publications

Van Hoek, M. J. A. & Hogeweg, P. In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. *Biophys J.* 91: 2833-2843 (2006)

Van Hoek, M. J. A. & Hogeweg, P. The effect of stochasticity on the lac operon: an evolutionary perspective. *PLoS. Comput. Biol.* 3: e111 (2007a)

Van Hoek, M. J. A. & Hogeweg, P. The role of mutational dynamics in genome shrinkage. *Mol. Biol. Evol.* 24: 2485-2494 (2007b)

Van Hoek, M. J. A. & Hogeweg, P. Evolutionary modeling of the metabolic network of yeast after its whole genome duplication. *In Preparation*

Dankwoord

Dit proefschrift was nooit tot stand gekomen zonder de hulp van een aantal mensen, die ik bij deze daar graag voor wil bedanken. Ten eerste wil ik mijn promotor, Paulien Hogeweg bedanken. Toen ik 5 jaar geleden voor het eerst bij je kwam om te vragen of ik, als natuurkundige zonder veel biologie-kennis, een promotieonderzoek bij jou kon doen, heb je me die kans gegeven, waar ik erg dankbaar voor ben. Het heeft me altijd verbaasd hoe snel jij dingen doorhad, waar ik lange berekeningen of simulaties voor nodig had. Ook sta ik er nog steeds van te kijken dat, hoewel je op het laatst wel 7 AIO's had en een flink aantal studenten, je toch altijd tijd voor me had.

Daarnaast wil ik mijn kamergenoten, Marian, Daniel en Rikkert, bedanken voor de gezellige tijd, ik heb het heel erg naar mijn zin gehad op onze kamer. Marian, bedankt voor de steun als ik er doorheen zat. Daniel, bedankt voor de leuke gesprekken over wetenschap en politiek. Rikkert, bedankt voor de gezellige spelavonden en je hulp bij mijn Matlab-problemen.

Ook wil ik de rest van de vakgroep, die zo langzamerhand te groot is om volledig met name te noemen, graag bedanken. Ik denk niet dat er veel vakgroepen zijn die zo gezellig en inspirerend zijn als deze. Met name wil ik nog Stan en Berend bedanken, met hulp en discussie bij het tot stand komen van respectievelijk hoofdstuk 2 en 4 van dit proefschrift. Verder wil ik Anton bedanken voor zijn programmeerhulp bij hoofdstuk 4, waar ik zelf nooit uitgekomen zou zijn, en het doorlezen van dit proefschrift. Daarvoor wil ik ook Kirsten bedanken, zonder haar had er in de eerste alinea al een foutje gestaan. Nobuto, bedankt voor de interessante wetenschappelijke discussies die we geregeld hadden. Ook wil ik onze systeembeheerder Jan Kees bedanken. Als ik weer eens een probleem had, kon ik altijd bij je terecht en was het binnen de kortste keren opgelost. En natuurlijk de Dalmuti-groep, waardoor elke lunchpauze weer de moeite waard was.

I would also like to thank Peter Schuster, Kunihiko Kaneko, Guillaume Beslon and Alexander van Oudenaarden for being members of my reading committee. Bedankt ook Sander Tans, dat ik je laboratorium mocht bezoeken, om de experimenten met *E. coli* eens in levende lijve te zien.

Verder wil ik graag Mathijs bedanken voor het feit dat hij mijn paranimf wil zijn. Ook ben ik erg blij met je goede idee voor de kaft! Ook wil ik Rob Kreuger bedanken voor zijn hulp met de kaft, nuttige lay-out tips en het leren hoe ik een poster moet maken. Ook al ben ik eigenwijs, ik heb de hulp erg gewaardeerd!

Natuurlijk wil ik ook mijn ouders graag bedanken, voor het vertrouwen in me en de hulp in de totstandkoming van dit boekje! Als laatste wil ik mijn vriendin Nanda bedanken. Bedankt voor al die keren dat je me opbeurde als ik weer eens dacht dat ik het nooit zou gaan redden, zonder jou had ik dat misschien ook niet. Je wist me altijd weer het gevoel te geven dat ik het wel zou halen en je had inderdaad gelijk! Ik kan niet wachten om over een paar maanden samen met jou op de fiets te stappen, India tegemoet!