“Nothing in Biology Makes Sense Except in the Light of Evolution”
Abstract

Biological sciences deal with a gigantic complexity. They study not only interactions between, but also within organisms. These organisms, or the networks/components that they are composed of, span over a wide range of length- (from nanometers to kilometres) and timescales (from picoseconds to years). Due to this complexity a systems biology approach was until recently hardly possible. However, over the last decade advances in genomics, proteomics, metabolomics and various other omics have allowed a paradigm shift. A reductionist approach focused predominantly on simplification was supplemented by an approach that allows to investigate and ultimately understand the system as a whole. The central aim in this thesis is to study the regulation of complex biological networks. In the previous chapters, different aspects of metabolic and regulatory networks have been studied, all inspired by the idea that we can obtain insights into network functioning by assuming that these networks are in one way or another functioning optimally. In this chapter, I will emphasis the main findings and try to bridge the results and put them into a coherent perspective. I will end with some concluding remarks and my ideas on future directions on this topic.

Constraints and Objective Functions

Due to its abundant and important role inside cells, metabolism has always received a lot of attention. As such, it has first benefited from reductionistic approaches, leading to detailed information of central parts of metabolism. These detailed descriptions of the components turned out to be valuable for current systems biology approaches e.g. kinetic models of glycolysis (Bakker et al., 1997; Teusink et al., 2000; Hynne et al., 2001). Later, due to genome-scale measurements and modelling, also less well studied parts of metabolism were taken into account. Remarkably, metabolism is one of the best described type of biological networks, in terms of their topology, whereas their regulation is less well understood. A metabolic flux for example, can be regulated simultaneously at the levels of transcription, translation and metabolism, or even at the metabolic level only. Lack of understanding of regulation principles seriously hampers us in our understanding and predictive power. This lack of understanding was illustrated by the outcome—in terms of enzyme expression levels—of a serial batch evolutionary experiment with yeast propagated on galactose (Chapter 3). We observed that one out of the four measured enzyme changed most, whereas the other enzymes had hardly changed. Together, this resulted in an increase of 18% in yeast's maximal growth rate. A combination of the results of a cost-benefit analysis and an enzyme fitness landscape, based on a mathematical model of this pathway, did reproduce the change in PGM expression but failed to do so for the other enzymes. Analysis indicated that PGM had the largest influence on fitness and the others were of less importance but still in-
fluenced fitness. These results suggest that the adaptive changes at the galactose metabolic pathway level, cannot fully explain the changes at the level of the phenotype. Alternative explanations that can be raised are (i) the mathematical model used turned out to wrong or incomplete or (ii) the duration of the evolutionary experiment was such that we picked up only the mutation with the largest effect on fitness. Overall these results indicate that the constraints and tradeoffs governing the regulation of a cellular system are not always clear. Note, that the galactose network is a relative simple network, consisting of only five metabolic enzymes, which has been studied extensively over the past decades. Even more so because usually only the outcomes of evolutionary experiments are subjected to experimental determination. Attempts to track back the occurrence of mutations and the reproducibility of them across independent conditions is usually not the main aim of laboratory evolution experiments, the enduring work of the group of Lenski being an exception (cf. Elena and Lenski, 2003; Kawecki et al, 2012). The constraints and objective functions considered in this thesis are a direct function of the enzyme expression levels. This led to valuable insights into the interplay between enzyme kinetics and their biochemical properties. It also allowed us to derive general theoretical frameworks for cost and benefit of enzyme expression (Chapter 3) and the identification of regulatory networks that can steer a metabolic network to a desired state (Chapter 4). Despite that, the experimental results in (Chapter 3) indicate that additional constraints and/or objective functions are needed to fully grasp the “endpoint” of this evolutionary experiment.

Even if only a single objective function applies to a cell, it does not guarantee that the mapping between molecular network (adaptations) and fitness is unique. Therefore, optimisations can still be useful but one should also come up with alternative hypothesis (Papp et al, 2009). To illustrate this, let us think about the outcome of the evolutionary galactose experiment again (i.e. PGM changed > twofold; other enzymes had hardly changed). The hypothesis we had is that the measured adaptations are leading to an increase in fitness, i.e. in an increase in growth rate. With this result alone we cannot strictly proof that the adaptations were actually selected for the proposed hypothesis. By only measuring adaptations (comparison between wild type and evolved strains) it cannot be ruled out whether the adaptations are directly shaped by natural selection or are by-products of other evolutionary processes, i.e. so-called hitch-hiking mutations. To overcome this, the observed mutations in the adapted strain can be introduced into the wild type and than placed under different environmental selective pressures.

An example of an alternative hypothesis could be that Gal1p induces expression of the GAL genes. Note that Gal1p is one of the enzymes for which we did not measure a significant change in its expression level. Gal1p is a bifunctional protein in that it has galactokinase activity and is able to induce the expression of the GAL genes (Bhat and Hopper, 1992; Platt and Reece, 1998). Gal1p and Gal3p are highly homologous
proteins (73% identity and 92% homology at the amino acid level, Platt et al., 2000). Although this induction is less efficient than that of Gal3p (Platt and Reece, 1998), it could well be that cells have increased Gal1p expression to improve the induction of the \textit{GAL} genes. Yet, the increased Gal1p concentration remains undetectable in our enzymatic assays because it's bound to other regulatory proteins that prevent it from its galactokinase activity. Unlike Gal1p, Gal3p does not possess galactokinase activity. However, it was experimentally shown that the addition of only two amino acids could impart the galactokinase activity of this protein (Platt et al., 2000), which is not so surprising given that closely related yeast species, e.g., \textit{Kluyveromyces lactis}, contain a single galactokinase-like molecule that functions both as a galactokinase and as a transcriptional inducer (Meyer et al., 1991). Overall, both Gal1p and Gal3p require both galactose and ATP for their mode of action. Although they perform seemingly different functions, we cannot rule out that Gal1p and Gal3p act through a similar mechanism. Therefore care should be taken with interpretation of the experimental results.

When studying living organisms, it should be acknowledged that adaptations will always occur. As such there is no endpoint to evolution. Therefore a full understanding of an organism's behaviour can only be obtained when the environment it has been evolved in, is taken into account. To illustrate this, I will use this section to discuss some additional experimental results. In addition to the galactose evolutionary experiment, a similar experiment with glucose was performed. In a nutshell: two genetically identical yeast strains (CEN.PK113-7D) were used to inoculate two evolutionary experiments: one in medium containing glucose as carbon source and one with similar medium but supplemented with (same concentration of) galactose. After about 400 generations the improvement in maximal growth rate (compared to the wild type) was 5 and 18%, for glucose (unpublished results) and galactose, respectively. How can two such chemically similar carbon sources (same chemical formula, they only differ in the arrangement of two side groups) lead to such different growth patterns? One might argue that this is because glucose is by far the most preferred sugar of yeast. Though this answer seems convenient, it does not answer \textit{why} glucose is so preferred. To come to an answer one therefore has to consider the environments to which yeast has been exposed in the past. It is speculated that yeast has been mainly exposed to glucose in it's environment and not so much to other sugars such as galactose. Following this reasoning it can be better understood why certain regulatory mechanism are present in yeast (e.g. glucose catabolite repression) and why the observed growth patterns differ so much. Finally, the experiment also nicely illustrates that even though yeast has been better adapted to glucose than galactose, when placed under the right selective pressures, there is still room for some improvement (e.g. a 5% increase in maximal growth rate on glucose).

Although it is convenient to consider one objective function at the time, I found it hard to believe that such simplistic view represents the natural environment of cells.
The fact that we study populations of living organisms implies that these organisms continuously interact and compete with each other. They encounter ongoing dynamics in nutrient concentration and levels of stress. Even when (in a perfect world) environmental conditions would not change, cells have to maintain themselves and generate offspring. During these processes errors are made and different strategies may arise that can be selected for (Gore et al, 2009; Bachmann et al, 2011). Recently, experimental studies indeed showed that metabolism in different bacteria operates at the optimal surface defined by three objective functions (Schuetz et al, 2007, 2012, see also Shoval et al, 2012). In Chapter 6 a new framework (called Feasibility Analysis, FA) is presented that allows to inspect multiple objective functions and the network parameters that are necessary to achieve (one, or all of) them. With this framework we used experimental data of a long-term chemostat experiment (Jansen et al, 2005; Mashego et al, 2005; Wu et al, 2006) to study the modes of regulation on the glycolytic pathway. Interestingly, we were able to untangle a tradeoff of two competing objective functions. FA attempts to combine kinetic and constraint-based modelling. By means of sampling, the kinetic model is used to create a parameter solution space, followed by the selection of a subset of feasible parameter samples by applying constraints. Currently, FA is only considering steady state samples. It would be of particular interest (though of course computationally more expensive) to extend the sampling method to dynamic states. As such, FA could –in contrast to the steady state based approaches in FBA– (i) make a significant contribution in our understanding of the dynamic regulation of metabolic networks and (ii) benefit from the growing interest to develop genome-scale kinetic models (Jamshidi and Palsson, 2008; Smallbone et al, 2010).

What is not discussed so far, is how cells are able to accurately signal and integrate multiple (competing) signals. For certain signals, cells have evolved dedicated receptors whereas other signals are sensed via indirect, or, yet unknown signalling systems. This observation is a consequence of a common organisational structure found in biological networks, a so-called, bow-tie structure (Csete and Doyle, 2004). A bow-tie architecture of metabolism learns us that a wide range of inputs (e.g nutrients), produces a large variety of signals (e.g. products and complex macromolecules) using relatively few intermediates (e.g. NAD(P)H, ATP). Based on this, it is of interest to understand how cells accurately sense and integrate signals that derive from multiple objective functions. We tested this idea on a simple model as shown in Figure 7.1A. This model consists of a linear metabolic pathway with an underlying gene network. The gene network receives signalling metabolites ($x_1$ and/or $x_2$) from the metabolic network as input, and outputs the three metabolic enzymes. Note, that this approach is similar as the input-output coupling as presented in Chapter 4.
Figure 7.1. Dual optimisation of a metabolic pathway. (A) Overview of a metabolic network that consists of three reversible enzymatic reactions and an underlying gene network. $S$ is the substrate of this pathway that is converted into product $P$. $R$ represents an external signal that is sensed by, though not catalytically affected, $e_2$. Metabolite $X_1$ inhibits the first enzymatic step. Synthesis of mRNA is influenced by signalling metabolite $X_1$ and/or $X_2$. mRNA in turn stimulates translation of the three metabolic enzymes. The mRNA’s and proteins are degraded using first order mass action kinetics. (B) Schematic overview of the coupling between metabolic and gene network. The gene network gets metabolic concentrations as input and generates enzyme expression levels as output. (C) 3D-plot showing the result of a dual optimisation of the metabolic network for optimal steady state flux and maximal responsiveness to $R$. For sake of comparison, both objectives were normalised (using equal importance) and then summed. (D) Optimal input-output relationship between $e_2$ and $X_1$. For discrete values of $S$ and $R$, the dual optimization was performed, using the enzyme concentrations as variables. (E) Optimal input-output relationship between $e_3$, $X_1$, and $X_2$. Note that due to the feedback at the metabolic level, the range of optimal concentration of $X_2$ is much smaller (better homeostatically regulated) than $X_1$.

When the metabolic network is optimised for a single objective function (optimal steady state flux over a range of substrate concentrations), smooth, monotonic increasing or decreasing functions between optimal metabolites and enzyme levels were found. Consequently, the gene network could be parameterised using a single signalling metabolite (either $X_1$ or $X_2$) as input. Next, we are interested in another external signal $(R)$. We require the system to be responsive towards $R$ (this was quantified using a response coefficient, Hofmeyr and Cornish-Bowden, 1996). With the responsiveness alone as objective function we were again able to find a (though different) gene network that could produce the desired enzyme expression level using a single
metabolite as input. Interestingly, when these two objective functions are applied simultaneously, a more complicated input-output relationship emerges. Although the combined response is still monotonically increasing with respect to the external signals (S and R, see Figure 7.1C), the underlying enzyme-(single-)metabolite relationship is not. In other words for a given value of signalling metabolite (in this example $x_1$), there exist multiple values of $e_2$ enzyme concentrations (see the 2D-plot, Figure 7.1D). For adequate regulation, a combination of both signalling metabolites is now required, which can be seen when plotting the relationship between enzyme concentration and the two signalling metabolites in a 3D-plot (Figure 7.1E). Based on these results, it seems suggestive that the number of signalling metabolites (e.g. transcription factors) forms an indication for the number of objective function that apply to the metabolic network. Whether these results can also be extended to more complicated networks remains to be shown.

Another way to analyse systems exposed to multiple objective functions is by converting objective functions into cellular constraints. In the above example, we could have had optimised the metabolic flux under the constraint that the response-coefficient of $e_2$ towards $R$ should be higher than a threshold. It should be noted that these two approaches are fundamentally different. But when the bounds of the constraints are carefully chosen (e.g. when they are based on experimental measurements), the outcome of a multi-objective optimization should in principle fall within the allowed solution space as defined by the constraints. Furthermore, it should be realised that constraints outside the system of interest could also delimit systems behaviour. To understand cellular behaviour, such as the growth rate, integration of the relevant modules into a cellular model is therefore a very promising approach (Molenaar et al, 2009).

**Control vs. Regulation**

Mathematical analysis of biochemical systems has been successful in quantifying the control exerted by component properties upon system variables such as flux and metabolite concentrations (e.g. control coefficients in MCA). These frameworks, however, do not address the question how living systems actually regulate their system properties when challenged with an environmental change. A valuable framework to quantify regulation of metabolic networks is regulation analysis. As already mentioned in the Introduction this framework allows to perform the “bookkeeping” of adaptations (e.g. how much of the observed adaptation can be attributed to a certain process), what it fails to do is to understand, and ultimately predict, these adaptations. Possibly, this lack of predictive power arises because of the choice to study large adaptations (e.g. different types of nutrients, different operational modes of metabolism). From an experimental point of view, such big changes are advantageous because they can be accurately measured, are industrially relevant and dissection of various underlying pro-
cesses is possible. From a theoretical perspective, large changes are less favourable because linearisation of the state of the system cannot be done, thereby complicating the predictions about the state changes upon perturbations. This might also explain why an extensive endevour of the available data-sets used in regulation analysis by me, did not reveal any correlations between environmental conditions, enzyme properties and/or the observed adaptive patterns (unpublished results).

In Chapter 5 we have studied optimal adaptive responses of molecular networks. We particularly focussed on the constraints that optimal metabolic performance imposes on the regulatory gene network. We assumed that metabolic enzyme expression is optimal for an objective function, and that the gene network will restore the enzyme expression levels to a new optimal state when perturbed. When the perturbations are (infinitesimal) small, our framework also allows to relate hierarchical regulation and flux control coefficients. Regulation analysis could greatly benefit if the adaptive responses can be made more insightful by relating it to the extensive knowledge already accumulated within MCA (Kacser and Burns, 1973; Fell, 1997). Below, I will describe some initial results that were obtained towards this goal.

We will consider pathways that obey: (i) they operate at steady states of maximal flux under the constraint of total enzyme conservation and (ii) they display optimal adaptation upon a parameter change (such that the new steady state again has maximal flux). For such systems, we can obtain a relationship between flux control and hierarchical regulation coefficients. The derivation of this relationship starts from the enzyme conservation equation, \[ \sum_{i=1}^{r} e_i \delta \ln e_i = 0 \] (7.1)
Division of Eq. 7.1 by the total enzyme concentration, \( e_T \), and the fractional change in the flux, \( \delta \ln J \), through the pathway gives,
\[ \sum_{i=1}^{r} \frac{e_i}{e_T} \rho_{h,i} = 0 \] (7.2)
Here the hierarchical regulation coefficient \( \rho_{h,i} \) was substituted for \( \frac{\delta \ln e_i}{\delta \ln J} \) (see Introduction). For linear pathways operating at maximal flux states under the constraint of enzyme conservation, the flux control coefficient of the \( i^{th} \) enzyme, \( C_i' \), equals \( e_i/e_T \) (Heinrich and Klipp, 1996; Klipp and Heinrich, 1999). Eq. 7.2 then simplifies to,
\[ \sum_{i=1}^{r} C_i' \rho_{h,i} = 0 \] (7.3)
In Chapter 3 the proportionality between optimal enzyme levels and the flux control coefficient was shown (see also: Klipp and Heinrich, 1994; Heinrich and Klipp, 1996; Heinrich and Schuster, 1998; Klipp and Heinrich, 1999). This means that flux control coefficients are always positive and range between 0 and 1. This shows that those pathways will always have both positive and negative hierarchical regulation coefficients. Suppose that one enzyme has a flux control coefficient of 0.9. Then all others have a summed flux control coefficient of 0.1. Which of the enzymes now has a positive or negative hierarchical regulation is not indicated by Eq. 7.3, this will depend on other properties such as, for instance, the nature of the perturbation. A small reduction in the level of the flux controlling enzyme only leads to a small reduction in its flux control coefficient and its $\delta \ln e$. However, the other enzymes will tend to: (i) have a large change in their hierarchical regulation coefficients (as the change in their enzyme level is considered relative to their initial enzyme concentration) and (ii) to remain without appreciable flux control coefficients (their enzyme levels has increased by a factor that is still much smaller than the total enzyme level). In such a case, the hierarchical regulation coefficient of the enzyme with high flux control is smaller than the hierarchical regulation coefficients of the enzymes that have little control. This is due to the scaling of the enzyme concentrations in the hierarchical regulation coefficients. So control and hierarchical regulation coefficients are linked.

For systems that do not operate under the constraint of enzyme conservation, we can derive another relationship between flux control coefficients (see also Teusink et al., 2011) and hierarchical regulation coefficients. The response of the flux to a change in all the enzyme concentrations (induced by some parameter perturbation) equals,

$$\delta \ln J = \sum_{i=1}^{r} C_i^j \delta \ln e_i$$

(7.4)

If all the enzyme levels display the same fractional change Eq. 7.4 becomes the summation theorem for flux control coefficients; the fractional change in the flux will equal the fractional change in enzyme levels. If this is not the case, the precise fractional changes in the enzyme levels matter for the change in the flux. Division by the fractional change in the flux then gives for a linear pathway,

$$\sum_{i=1}^{r} C_i^j \rho_{h,i} = 1$$

(7.5)

For general linear pathways, flux control coefficients are also typically positive and bounded to 1; but this is now not a necessity as it was for a linear pathway operating at maximal flux under the constraint of total enzyme conservation. The main difference between Eq. 7.3 and 7.5 is that the sum equals 1 for regular pathways. This has an in-
teresting consequence. Suppose all enzymes have an equal flux control coefficient of $1/r$ ($r$ the total number of enzymes). Eq. 7.5 then becomes $\sum_{i=1}^{n} \rho_{h,i} = r$. The average hierarchical regulation coefficient is 1. For optimal linear pathways the relationship becomes $\sum_{i=1}^{n} \rho_{h,i} = 0$. The average coefficient is now 0. These results suggest that by combining the data on $\rho_{h}$ and the flux control coefficient (usually difficult to measure, so this value could rather be computed using a mathematical model), in principle, one should be able to test whether a metabolic network operates under an optimal flux regime.

Concluding Remarks and Future Directions

The results presented in this thesis are inspired by the statement that optimisation leads to the most efficient solution to a certain problem. The premise that optimisation applies to biology was initially inspired by Darwin’s theory of natural selection. He stated that more efficiently designed individuals will leave more offspring. This concept has already revolutionised some aspects of biology and is therefore a promising attempt to unravel the piles of experimental data being generated. The results in this thesis indeed show the potential of such approaches.

Optimisation has already been employed in fields other than biology. Economists have traditionally calculated the best options that result in the greatest profit, and engineers routinely calculate the best design solutions, such as the lightest airplane, given certain other requirements. Interestingly, these concepts are more and more being applied in biology. The general cost-benefit theory for enzyme expression developed in this thesis (Chapter 3), nicely illustrates how insightful the use of such “neighbouring concepts” for biological problems can be.

The optimality hypothesis should —just like any other hypothesis in science— be tested because only then it can be concluded whether the hypothesis was right or wrong. It is therefore important to realise that it is not the aim to test whether organisms are optimal. Actually, it is the assumptions of optimality that are tested. Throughout this thesis I have presented examples of this procedure. The failure to find support for a prediction can be used to determine whether an assumption is wrong. As already stated earlier, the identification of alternative possibilities is thus important; once they are identified, a new theory can be devised and tested.

Mathematical models are extremely useful to complement experimental endeavours to come to a coherent understanding of the behaviour of molecular networks. Optimisation theory is an active field in mathematics and as such, with more and more mathematical models becoming available, biology could greatly benefit. The results obtained within the last two decades in the field of FBA is arguably the clearest example of the success of optimisation hypotheses. In contrast to the fast and relative simple (linear) optimisation problem definitions used in FBA, the situation gets quickly more
complicated when optimisation of dynamic models is considered. Kinetic models by themselves can already contain a severe complexity. This seriously affects the time (and sometimes) quality of the solutions when performing optimisations. Therefore, there is a need for better optimisation algorithms that can deal with big, complex and dynamic models. Furthermore, it should always be checked whether the outcomes of optimisations correspond to a global optimum. In this thesis, I compared the obtained solutions, with randomly generated samples (that would still fulfil the requirements of the constraints). Alternatively, evolutionary algorithms that explicitly simulate the process of natural selection can be used (Patil et al., 2005; Kratz et al., 2008).

The potential of optimisations has, according to my view, a much wider application than considered so far. To fully exploit this potential certain issues have to be addressed: the most important being a precise formulation of what the selective pressure (fitness) is and how it emerges from molecular networks. Having accomplished that will greatly improve the use of optimisation approaches. It will still remain an enormous challenge to capture all molecular components within such fitness objective function(s). Therefore, multiple objectives can be proposed. But I think the biggest achievements can be made when the objective function(s) and the constraints are made dependent on for instance the environment or the state of the organism. In this way, it will reflect the everlasting changes that organism encounter and the continuous adaptations and tradeoffs that these organisms have to make in order to survive.