5

Optimal Adaptations

Robustness of Optimal Metabolic Performance by Optimal Gene Expression in Varying Environments

In collaboration with:
Evert Bosdriesz, Frank J. Bruggeman

Manuscript in preparation
Abstract

To understand the principles of the remarkable adaptivity of metabolic networks in dynamic environments, their underlying gene regulatory networks need to be characterised. Typically, little is known about the kinetics and topology of those gene networks in contrast to the metabolic networks that they regulate. Here, we present a theoretical approach based on metabolic control analysis to elucidate gene regulatory networks that can achieve a specific metabolic control task on the basis of information about the metabolic network alone. We focus on optimal gene networks that are capable of restoring the metabolic network to an optimal state upon an environmental perturbation. We show how the demand for optimal regulation of metabolism sets constraints on the input-output relationship of gene networks and leaves degrees of freedom for the gene network to meet other demands. We derive the theory to identify the constraints acting on the gene network to guarantee perfect adaptation of metabolic optimal behaviour by regulation of metabolic gene expression. We also show how gene network structures can be found that give rise to this optimal steering of metabolism using a linear programming approach, given gene network design criteria and a mathematical model of metabolism. Our approach is applicable to synthetic biology studies for rational engineering of metabolism.

Introduction

In previous chapters we have introduced optimal metabolic performance under constraints (Chapter 3) and made use of the regulatory interactions between various cellular networks using an input-output relationship (Chapter 4). The input-output relationship was then used to fit the kinetic parameters of an existing model of the galactose regulatory network. In this chapter we aim to incorporate all these concepts into a coherent theoretical framework. With this framework we aim to: (i) identify constraints and limits of regulatory networks on the basis of information of the metabolic level only, (ii) identify adaptive changes in enzyme levels and metabolites upon a parameter perturbation required for optimal regulation of metabolism and (iii) use these optimal adaptations to construct the simplest regulatory network that can generate the desired adaptations.

Cellular function emerges from the concerted action of the signalling, gene expression and metabolic networks. When cells are faced with environmental perturbations, regulatory mechanisms are induced to give rise to adaption. Cellular adaptations restore vital network functions and, ideally, improve fitness. To come to a coherent understanding of cellular adaptations, integration of measurements from different cellular layers is therefore required (Rossell et al, 2006; Daran-Lapujade et al, 2007; Ishii et al, 2007; Bruggeman et al, 2008). With current technological possibilities, the availability
Optimal Adaptations

of experimental data is rarely limiting, however, the real challenge is to understand how regulation of molecular networks give rise to these adaptations.

An encouraging approach to unravel how molecular interactions of regulatory networks give rise to cellular adaptation is by looking for patterns or motifs that occur more often than one would expect if the design was random. These recurring regulatory patterns are called network motifs (Milo et al., 2002; Alon, 2007, 2006). Network motifs have been studied extensively experimentally (Basu et al., 2004; Mangan et al., 2006; Kaplan et al., 2008) and theoretically (Tyson et al., 2003; Prill et al., 2005). Despite the success of these examples, it is to be expected that for networks composed of multiple interactions from different cellular layers—such as metabolism—extensive knowledge of the motifs alone is not sufficient. One hypothesis that is used to overcome this, is to assume that—due to long term evolutionary selective pressures—biological adaptations can be understood as the outcome of optimization for cellular fitness (Chapter 4). Application of optimality hypotheses (Parker and Smith, 1990) has a long history in biological studies (cf. Smith, 1978; Gould and Lewontin, 1979) and has been successfully applied to explain adaptive responses (Segrè et al., 2002; Chubukov et al., 2012), and optimal control structures (Heinrich and Schuster, 1998; Klipp and Heinrich, 1999; Kalisky et al., 2007; Goyal et al., 2010). Furthermore, it forms the core of Flux Balance Analysis (FBA, Orth et al., 2010); the most commonly used method to simulate genome-scale metabolism. Further evidence for optimal performance in terms of the microbial growth rate comes from studies on unneeded protein synthesis (Novick and Weiner, 1957; Dean et al., 1986; Dong et al., 1995; Snoep et al., 1995; Lunzer et al., 2002; Dekel and Alon, 2005; Stoebel et al., 2008). When non-functional, or unneeded proteins are expressed a reduction in growth rate is observed, that is likely to arise because of a limitation in the protein expression machinery (e.g. limited amount of energy or (free) ribosome abundance). Experimental evidence of a cellular constraint that limits the total protein content was provided by showing that unneeded protein expression results in a decreased expression of the other growth-related proteins (Vind et al., 1993; Dong et al., 1995). Here we use this constraint to limit the functioning of metabolic networks and ask the question to what extent gene networks are capable of optimally regulating metabolism.

Thus, constraints set limits to the “phenotype space” in which cellular networks can perform. Implying that the identification of the proper constraints is crucial to understand cellular behaviour and adaptations. At present, adaptive responses of a metabolic network, e.g. why some enzymes go up in expression level whereas others go down, are poorly understood. Proper functioning, in a fitness sense, of a metabolic or signalling function depends on enzyme expression levels and their kinetic properties. This information suffices to predict responses of enzyme levels that restore optimal functioning upon changes in the environment. It is an open question whether gene net-
works can always realise those enzyme levels, given molecular information about the metabolic and signalling state. One cannot rule out that constraints and tradeoffs are at play, at the level of the gene network that prevent certain adaptive enzyme changes, as those would lead to a reduction of fitness through some mechanism independent of the subnetwork of interest.

In this work we provide a predictive theoretical framework for understanding adaptive network responses; rules of thumb and theory that relate changes in enzymes levels to their biochemical properties. Our analysis starts with finding the optimal enzyme distribution within a mathematical model of a metabolic system under the assumption that resources are limited. This optimal state is used as a starting condition. Next, we perturb the optimal metabolic network and seek for the changes in enzyme levels such that the perturbed state is again an (different) optimal state. Interestingly, we find that these changes can be calculated using metabolic information from the starting condition only. Subsequently, based on the changes in enzyme and metabolite levels, an input-output relationship is constructed. The gene networks are constrained by the input-output relationship. As such we are able to ensure a robust regulation of metabolic performance by the gene network. Finally, we use linear programming to find the minimal structure of such gene networks.

**Methods**

**Kinetic Model Description and Optimal Input-Output Relations**

The model depicted in Figure 5.3A is described by the following set of ODEs:

\[
\frac{dm_1}{dt} = v_1 - v_2, \quad \frac{dm_2}{dt} = v_3 - v_4
\]

where

\[
v_1 = \frac{k_{cat1} \theta_1}{K_{M1}} \left( 1 - \frac{m_1}{S_1 K_{eq1}} \right) \]

\[
v_2 = \frac{k_{cat2} \theta_2}{K_{M2}} \left( 1 - \frac{m_2}{S_2 K_{eq2}} \right) \]

\[
v_3 = \frac{k_{cat3} \theta_3}{1 + m_1/K_{M1}} \frac{m_2/K_{M2}}{1 + m_2/K_{M2}} \left( 1 - \frac{m_3}{K_{eq3}} \right)
\]
The total enzyme concentration was kept constant. Optimal input-output relations were calculated by perturbing $S_1$ and $S_2$ for different external conditions. The values used for both $S_1$ and $S_2$ were $10^n$, $n \in \{-1, -0.75, \ldots, 0.75, 1\}$. This gave a total of $9 \times 9 \times 2 = 162$ different input-output relations.

**Formulation of MILP Problem**

In order to formulate the problem described in the main text as a proper MILP problem with indicator constraints, we have to split each potential interaction into an activating and an inhibition one. One entry in the $r^c_i$ matrix then becomes

$$
\sum_{\gamma=1}^{N_{TF}} \sum_{T, U \in \{A, I\}} (-1)^{n_i} r_{i,TU}^c,
$$

where $n_i$ is the number of inhibiting in interaction (i.e. $(-1)^{n_i}$ is 1 if $\{t, u\} = \{A, A\}$ or $\{I, I\}$ and -1 if $\{t, u\} = \{I, A\}$ or $\{A, I\}$). For the optimisation problem as described in the main text, the full MILP problem is than formulated as follows:

\[
\min \sum_{i=1}^{N_c} \sum_{j=1}^{N_{TF}} \sum_{t \in I, A} I_{TF}^{T_f} + \sum_{k=1}^{N_e} \sum_{l \in I, A} I_{EF_k}^e
\]

subject to:

\[
\delta e_i^c = \sum_{t, u \in I, A} \sum_{\alpha=1}^{N_e} \sum_{j=1}^{N_{TF}} \left(\text{sign}\right) r_{i, j, U}^{c, \alpha} \delta x_j^c \quad \forall i, \xi
\]

\[
I_{A_i}^{TF} = 1 \Rightarrow I_{x_j}^{TF} = 0
\]

\[
I_{x_j}^{TF} = 1 \Rightarrow I_{A_i}^{TF} = 0
\]

\[
I_{A_i}^{e_i} = 1 \Rightarrow I_{T_i}^{e_i} = 0
\]

\[
I_{T_i}^{e_i} = 1 \Rightarrow I_{A_i}^{e_i} = 0
\]

\[
I_{T_i}^{e_i} = 0 \Rightarrow r_{iabTU} = 0 \quad \forall i, U
\]

\[
I_{R_i}^{T_{ik}} = 0 \Rightarrow r_{ik},TU = 0 \quad \forall k, T
\]

\[
0 \leq r_{ik},TU \leq r_{max}
\]

\[
I_{TF}^{e_i}, I_{TF}^{T_i} \in \{0, 1\}
\]

We used IBM ILOG CPLEX Interactive Optimiser 12.4 to solve these optimisation problems.
Chapter 5

Results

Mathematical Description of Metabolic Networks and a Cellular Protein Constraint

Our analysis starts with a kinetic model of a metabolic network. Kinetic models are used to study dynamics and control properties (Heinrich and Schuster, 1996). For a metabolic network of \( r \) metabolic reactions, the rates of changes of its \( n \) metabolic intermediates can be written as,

\[
\frac{d}{dt}m(t, p, e) = N_m \cdot v_m(m(t, p, e), p, e)
\]  

(5.1)

The concentration of metabolic intermediates are represented by the \( n \times 1 \) vector \( m \) and \( t \) denotes time. The parameter vector \( p \) contains environmental parameters (e.g. external substrate concentration, temperature), equilibrium constants, kinetic parameters (e.g. affinity constants, catalytic activities) and initial conditions. The \( n \times r \) stoichiometric matrix is given by \( N_m \) and contains the stoichiometric coefficients of the reactions. The rate \( r \times 1 \) vector \( v_m \) denotes the rate vector containing the kinetic description of all metabolic reactions in terms of rate equations, i.e. mass-action or enzyme kinetics. The vector \( e \) contains the concentrations of the metabolic enzymes.

Generally, the rate of an enzyme catalysed reaction depends linearly on the enzyme concentration, i.e. \( v_i(e_i, m, p) = k_{cat,i} \cdot e_i \cdot \phi_i(m, p) \) where \( \phi_i(m, p) \) describes the saturation of the enzyme with reactants and effectors (Heinrich and Schuster, 1996) and \( k_{cat,i} \) the catalytic rate constant of the enzyme. Here, we are primarily interested in the set of optimal enzyme expression levels that maximise a steady state metabolic reaction rate (flux) in order to maximise fitness. This kind of evolutionary optimisation of fitness will in general occur under constraints, as the cellular protein content is bounded. The origins of those constraints can come from physics, energetics, biochemistry, or the environment and (in-) directly set the maximal cellular protein content (as illustrated in Chapter 3).

A bound on the cellular protein content is a natural constraint for metabolic pathways because every reaction rate depends linearly on the enzyme amount and metabolic fluxes can be enhanced by increasing the enzyme concentrations. However, enzyme synthesis costs nutrients and energy, which are not in excess, and the enzymes themselves require space inside the cell, which is also limited (Molenaar et al., 2009). Thus, a limit exists for the cellular protein content, which will set a bound on the metabolic fluxes that can be achieved. Hence, enhancement of metabolic fluxes can result from proper partitioning of the available resources over metabolic enzyme synthesis through alteration of metabolic gene expression. A question that remains is how the requirements for optimal metabolic functioning constrains the gene networks responsible for the optimal adaptation of enzyme levels.
Before we can consider metabolic pathway optimisation by alterations in enzyme levels and identify the demands for optimal metabolic gene expression, we have to specify the protein constraints in more depth. This we will do in a very nonrestrictive manner. The constraints we will be considering (implicitly) set a limit to the total cellular protein content, \( e_T \), which is defined as the sum of all the protein concentrations, i.e. \( e_T = \sum_{i=1}^{n} e_i \). We denote a constraint function by \( \Phi \), which depends on all the enzyme levels, \( e_i \), and on the specific constraint contribution of each enzyme, denoted by \( \omega_i \); thus we obtain \( \Phi(\omega, e) \). The constraint function is bounded because of physical, energetic or biochemical reasons; in the simplest case we then arrive at,

\[
\Phi(\omega, e) = \sum_{i=1}^{n} \omega_i e_i \leq R
\]

The constraint \( R \) could for instance derive from the available energy, expressed in terms of ATP equivalents, or limited availability of ribosome or RNA polymerases. The value of \( e_T \) depends on the enzyme concentrations. For instance, enzyme concentrations with low \( \omega \)'s are high in concentration then \( e_T \) is higher than in the opposite scenario. Because enzyme reaction rates are generally linear functions of the enzyme concentration, higher enzyme concentrations are favourable for high metabolic reaction rates. But, if those enzymes are expensive then the total cellular enzyme concentration decreases. This indicates that there exists an enzyme expression pattern that maximises the steady state metabolic fluxes. This has been studied in the past by Heinrich and colleagues (Klipp and Heinrich, 1994; Heinrich and Klipp, 1996; Heinrich and Schuster, 1998; Klipp and Heinrich, 1999). Here we will extend these studies by identifying how gene networks can attain optimal levels of metabolic enzymes.

**Constraint Optimization of a Linear Metabolic Pathway Flux by Varying Enzyme Levels**

We consider a linear metabolic pathway at steady state, such that all rates of change in Eq. 5.1 equal 0: \( \dot{m}(t, p) = 0 \). This pathway may contain positive and negative feedforward or feedback regulation and include moiety conservation (see Figure 5.1A). In fact, the condition for this theory to work are likely more general than linear metabolic pathways. Actually, we require that in the *optimal state* the resulting network is a linear chain of reactions. We expect this condition to hold quite generally. If a flux of a complex (branched) metabolic network is to be maximised, by optimising enzyme concentrations under constraints that bound the total enzyme concentration, the tendency to use the shortest *linear* pathway to achieve this flux will be quite strong. Since, we do not have a proof of this assertion and do not fully understand its implications and assumptions, we limit ourselves to linear metabolic pathways in this work.
Chapter 5

The optimal enzyme expression levels that optimise the steady state flux, $J$, of a linear metabolic pathway can be found from the following constrained optimisation problem:

$$\begin{align*}
\text{max} \quad & v_i(e) = J(e) \\
\text{subject to:} \quad & e \in \mathbb{R}^{r^+} \\
& N_m \cdot v_m(m, p, e) = 0 \\
& e \cdot \omega = R
\end{align*}$$

(5.2)

This optimisation can be carried out with mathematical optimisation libraries of Mathematica (Mathematica, 2010), Matlab (MATLAB, 2012), Copasi (Hoops et al, 2006) or GAMS (Gams, 2012). Note that this optimisation problem is equivalent to minimising the total amount of resources $R$ as objective given a target flux value, $J$, and the same constraints. Formulating this optimisation problem is not new (Klipp and Heinrich, 1994; Heinrich and Klipp, 1996; Klipp and Heinrich, 1999; Liebermeister et al, 2004), but next we will consider a new dimension to this optimisation problem; i.e. the prediction of enzyme level changes that make sure that the optimal state of the metabolic network is robust to external perturbations.

The outcome of the optimisation problem stated in Eq. 5.2 is a vector of optimal enzyme concentrations $e^o$ that achieves the optimal steady state flux, $J^o$, given the current characterisation of the environment. Our interest is in cellular adaptation over a range of different environmental conditions. Hence, we will write the vector of optimal enzyme levels as a function of an environmental parameter, $s \in p$, i.e. $e^o(s)$ and $s \in S$. The parameter $s$ will often be the extracellular concentration of a substrate of the metabolic network under consideration. Substitution of $e^o(s)$ in the steady state condition for the metabolic network allows for the determination of the metabolite concentration in the optimum state, $m^o(s)$; i.e. by solving the nonlinear system of equations: $N_m \cdot v_m(m^o(s), p, e^o(s)) = 0$ for $m^o(s)$. This procedure gives us for every value of $s$ the following list $(m^o(s), e^o(s))$, which can be viewed as an input-output relationship of the metabolic network with $e^o(s)$ as input and $m^o(s)$ as output. We will exploit this insight below to constrain the gene network controlling the enzyme synthesis of this metabolic network.

**Optimal Adjustment of Enzyme Levels upon Environmental Perturbation**

The constrained optimisation problem introduced above can be expressed in terms of Lagrange multipliers; this analysis indicates that in the optimal state the following
Figure 5.1. Illustration of a network that restores to an optimal state upon a parameter perturbation. (A) Shown is a metabolic network that consist of four enzymes that convert substrate $S$ into product $P$. Metabolite $m_3$ is involved in a negative feedback loop, inhibiting the first enzyme. ATP and ADP are used as co-substrates in reactions 2 and 4, forming a conserved moiety. We consider the metabolic network in steady state, with some flux $J$ (underlined species indicate that their concentration is fixed). The system is perturbed by a perturbation $\delta p$ and we assume that the new steady state of the metabolic network is again optimal. (B) A sketch of the optimal response of a reaction flux in metabolism upon a change in a parameter, $p$. Shown is an enzyme, parameter flux solution space. For a given set of parameters $p$ and enzyme concentrations $e$, the flux is maximal (indicated by the open square). Upon a change in parameter $p$ of $\delta p$ the system moves to a new steady state. As long as the enzyme levels are not adjusted this new state is not optimal (open circle). Optimal changes in enzyme levels, induced by gene expression regulation and post-transcriptional processes, then restore the metabolic flux to an optimal state again at $(p + \delta p, e_i + \delta e_i, \text{open triangle})$. 

$J(p+\delta p, e_i) \rightarrow J(p, e_i)$

$\frac{\partial J}{\partial p_\delta p}$

$\delta p$

Optimal solution

($p+\delta p, e_i$)

Non-optimal solution

($p, e_i$)
Chapter 5

A relationship exists (Heinrich and Schuster, 1998; Klipp and Heinrich, 1999),

$$\forall i : \quad f_i(p, e^o) = \frac{\partial J(p, e^o)}{\partial e_i} - \frac{\omega_i J(p, e^o)}{R} = 0 \quad (5.3)$$

Note that we are considering a linear metabolic pathway and therefore: $\forall i : \nu_i = J$. As a consequence of Eq. 5.3, the flux control coefficient of enzymes in the optimal state equals their fractional resource requirement, i.e. $C_i^f = \frac{\omega_i}{R}$, which is known from earlier studies (Heinrich and Schuster, 1998; Klipp and Heinrich, 1999, see also Chapter 3).

Our interest is the identification of the constraints for a gene network that is capable of optimally adjusting the enzyme levels of a metabolic network such that it maintains optimal functioning upon environmental perturbation. Then, both the reference and the new steady state, attained after the parameter perturbation, are optimal states of metabolic activity. This is displayed schematically in Figure 5.1B. In addition, we demand that the system spends an equal amount of resources, $R$, before and after the parameter perturbation. Since, we are optimising the metabolic flux and the reaction rates all depend linearly on all the enzyme concentrations, we know that at the optimal state all resources will be spent. Hence, the above requirements demand that upon perturbation $\delta p_k$ of parameter $p_k$,

$$\delta R = R(p_k + \delta p_k) - R(p_k) = \sum_{i=1}^{r} \omega_i \delta e_i = 0$$

As a consequence, we only need to determine $r - 1$ changes in enzyme levels as the change in the $r$-th enzyme level follows from the previous relationship. The $r - 1$ enzyme levels we will refer to as the independent enzyme levels and denote them by $\hat{e}_j$ (with $1 \leq j \leq r - 1$). The relationship between all the enzyme levels and the independent metabolites can be written as $\delta e = L\hat{e}$ with $L$ as an $r \times (r - 1)$ matrix composed of an $(r - 1) \times (r - 1)$ identity matrix stacked on top of the $(r - 1) \times 1$ (row) vector with as its $i^{th}$ entry, $\omega_i/\omega_r$.

When a parameter perturbation, $\delta p_k$, is applied to the metabolic network, e.g. a change in the pathway substrate is made, then the change in Eq. 5.3 can be expressed as,

$$\forall i \in \{1, 2, 3, ..., r - 1\} : \quad f_i(e + \delta e, p_k + \delta p_k) = f_i(e, p_k) + \frac{\partial f_i}{\partial p_k} \delta p_k + \left( \frac{\partial f_i}{\partial e} L \right)_{1 \times (r - 1)} \delta \hat{e} \quad (5.4)$$

In addition, we observe that the response of the independent enzyme concentrations are due to the parameter change, $\delta \hat{e} = \frac{\partial \hat{e}}{\partial p_k} \delta p_k$. If the reference state at $p_k$ and the new state $p_k + \delta p_k$ are both optimal then $f_j(e + \delta e, p_k + \delta p_k) = 0$ and $f_j(e, p_k) = 0$ such that
the Eq. 5.4 reduces to,

$$\forall i : \frac{\partial f_i}{\partial p_k} = - \left( \frac{\partial f_i}{\partial e} \right) \frac{\partial \hat{e}}{\partial p_k}$$ (5.5)

Eq. 5.5 indicates that the parameter perturbation leads to a deviation from the optimality conditions, which is compensated for by a response in enzyme levels. We have $r - 1$ such relationships, as many as the unknown optimal changes in enzyme levels, $\delta \hat{e}$. Rewriting the previous expression in matrix form, we obtain the optimal enzyme responses equated in terms of information about the metabolic network only,

$$\left( \frac{\partial \hat{e}}{\partial p_k} \right) = - \left( \frac{\partial f_i}{\partial e} \right)^{-1} \left( \frac{\partial f_i}{\partial p_k} \right)$$ (5.6)

Here we assume that the $r - 1 \times r - 1$ matrix $\frac{\partial f_i}{\partial e}$ is invertible.

**Response of Metabolic Intermediates upon Optimal Change in Enzyme Levels**

Eq. 5.6 gives us the optimal response of the enzyme levels that compensates for the deviation from the optimality condition (Eq. 5.3) upon a parameter perturbation, $\delta p_k$. The change in the enzyme levels and parameter value both cause a change in the steady state metabolite concentrations. To determine this change we have to implicitly differentiate Eq. 5.1 at steady state condition,

$$N_m v_m(e(p_k), m(p_k), p_k) = 0,$$

with respect to $p_k$; we obtain,

$$\left( N_m \frac{\partial v_m}{\partial m} \frac{\partial m}{\partial p_k} + N_m \frac{\partial v_m}{\partial e} \frac{\partial e}{\partial p_k} + N_m \frac{\partial v_m}{\partial p_k} \right) \delta p_k = 0$$

Here we have assumed that no linear dependencies exist amongst the rows of $N_m$, which means that we consider the differential equation of the metabolic network in terms of its independent variables. All molecular reaction systems can be written in this format (Reder, 1988; Heinrich and Reder, 1991). The $n \times n$ matrix $N_m \frac{\partial v_m}{\partial m}$ is the so-called Jacobian matrix, which we assume to have eigenvalues with negative real parts — the stability condition for the metabolic network. All these derivatives are to be evaluated at the optimal state prior to perturbation of parameter, $p_k$. The response of the metabolite levels to a change in the parameter is given by,

$$\delta m = \frac{\partial m}{\partial p_k} \delta p_k = - \left( N_m \frac{\partial v_m}{\partial m} \right)^{-1} N_m \left( \frac{\partial v_m}{\partial e} \frac{\partial e}{\partial p_k} + \frac{\partial v_m}{\partial p_k} \right) \delta p_k$$ (5.7)
Chapter 5

In Eq. 5.7 the response of the enzyme concentration is given by Eq. 5.6,

\[
\frac{\partial e}{\partial p_k} = L \left( \frac{\partial \hat{e}}{\partial p_k} \right) = -L \left( \frac{\partial f}{\partial e} \right)^{-1} \left( \frac{\partial \mathcal{A}}{\partial p_k} \right)
\] (5.8)

From Eq. 5.6 and 5.8 we can determine the change in the input-output relationship of the metabolic network required to restore optimal metabolic functioning upon a parameter change

\[
(\delta \mathbf{m}^0, \delta \mathbf{e}^0) = \left( -\left( \mathbf{N}_m \frac{\partial \mathbf{v}_m}{\partial \mathbf{m}} \right)^{-1} \mathbf{N}_m \left( \frac{\partial \mathbf{v}_m}{\partial \mathbf{e}} \frac{\partial \mathbf{e}}{\partial p_k} + \frac{\partial \mathbf{v}_m}{\partial p_k} \right), -L \left( \frac{\partial f}{\partial e} \right)^{-1} \left( \frac{\partial \mathcal{A}}{\partial p_k} \right) \right) \delta p_k
\] (5.9)

All the derivatives on the rhs are evaluated at the optimal reference state \((\mathbf{m}^0(p_k), \mathbf{e}^0(p_k))\) at \(p_k\) and the lhs gives the required changes in the metabolite and enzyme concentrations to guarantee optimality in the new steady state at \(p_k + \delta p_k\).

Constraining the Gene Regulatory Network of the Metabolic System to Guarantee Optimal Responses

In the previous section, we have derived the response of the metabolite concentrations in a linear metabolic pathway upon a change in a parameter change and the levels of its metabolic enzymes to restore optimal functioning of the metabolic pathway. These requirements guarantee robustness of metabolic optimal flux by way of optimal gene expression. We have not yet made the link to the gene network. The metabolic network has as its inputs, the value of the perturbation parameter and the enzyme levels. As its output it has the metabolite concentrations. Typically some of those metabolite concentrations act as regulators that bind to a transcription factor to modulate gene transcription of metabolic enzymes (Figure 5.2, see also Chapter 4). As a consequence, the gene network will have metabolites as its input and metabolic enzyme levels as its output; thus, if we know the optimal input-output response of the metabolic network (Eq. 5.9), we also know the change in the gene network. Therefore, any gene network that receives \(\delta \mathbf{m}^0\) as its input, while operating at a steady state at \(\mathbf{m}^0\), and that has as output \(\delta \mathbf{e}^0\) will restore optimal metabolic functioning of a metabolic network that receives a parameter change \(\delta p_k\) while functioning optimally at steady state \((\mathbf{m}^0(p_k), \mathbf{e}^0(p_k), p_k)\) with maximal metabolic flux \(J\) given the resource abundance \(R\) and costs \((\omega_i)\) for all enzymes.

We will now constrain the gene network by the demand of the optimal input-output relationship. We collect the subset of the metabolites that acts as the input for the gene network in the vector \(\mathbf{\hat{m}}\) and the set of optimal concentration and optimal perturbations in concentrations respectively in \(\mathbf{\hat{m}}^0\) and \(\delta \mathbf{\hat{m}}^0\) and recall that these vectors are elements of \(\mathbf{m}^0\) and \(\delta \mathbf{m}^0\), respectively. Thus the gene network, \(G\), operates as the map
The question is: can we say something about the gene network structure, using this metabolic information only? To solve this problem we exploit the fact that modular response analysis (MRA) expresses the response of gene networks in terms of their interaction map (Kholodenko et al., 2002; Bruggeman et al., 2002, 2008). This is possible for gene networks because they essentially only contain regulatory interactions between variables; the variables are not consumed while acting as regulators. For instance, a transcription factor modulates the synthesis rate of a mRNA without being consumed and the mRNA in turn modulates the translation rate while not being consumed in the process. Finally, the protein catalyses a metabolic reaction while not being consumed in the process. Networks with these particular properties are called hierarchical networks and have been studied within metabolic control analysis for quite some time (Kahn and Westerhoff, 1991; Hofmeyr and Westerhoff, 2001; Bruggeman et al., 2002, 2008). In MRA, the strength of molecular interactions within gene regulatory networks, i.e. transcription factor influences on mRNA levels and mRNA level influences on protein levels, are quantified by so-called local response coefficients denoted by $r^X_Y$ for the influence of $X$ on $Y$. These coefficients are defined as the changes in the steady state level of $Y$ upon a change in the steady state level $X$ while all other species are held at their reference steady state concentrations. Thus, all indirect interactions between variables in MRA have a local response coefficient of zero.

The power of MRA derives from the fact that the integrated changes in all the variables of the gene network system can be equated in terms of the local response coefficients through a very simple linear algebraic relation (Bruggeman et al., 2002, 2008),

$$R^X_m = -(r^X_m)^{-1} r^X_m$$

Eq. 5.10 is usually written in terms of normalised derivatives but here we consider them unnormalised. The vector $x$ denotes all the variables of the gene regulatory networks, i.e. the transcription factors, mRNA’s and proteins. The vector $m$ denotes the inputs – the perturbation parameters – of the gene network, i.e. in our case the gene regulatory metabolites deriving from the metabolic network. The vector $r^X_m$ contains the local response coefficients of the gene regulatory network to the regulatory metabolites; those will often be transcription factor and metabolite binding events. The matrix $r^X$ denotes the interaction map, which is essentially a weighted adjacency matrix (see Figure 5.2C for an illustration of this matrix for the example network), with the weights corresponding to the local response coefficients. Eq. 5.10 relates the output and the input of the gene network,

$$\delta x = R^X_m \delta m$$
The metabolic enzyme concentrations are an element of the state of the gene network, i.e. $e \in x$. Hence, we can focus on those variables alone,

$$\delta e = R_m^e \delta \hat{m}$$

If the following holds,

$$\delta e^o = R_m^e \delta \hat{m}^o$$

then the gene network generates the optimal input-output relationship and the metabolic system restores its optimal functioning after the parameter perturbation and the change in its metabolic enzyme levels.

**Finding the Minimal Structure of an Optimal Gene Network Using Mixed Integer Linear Programming**

We now turn to the problem of finding the minimal gene regulatory network capable of steering the system back to an optimal state after a perturbation. This is interesting to know for a number of reasons. In synthetic biology, it can be helpful designing as simple as possible circuits that optimally steer a target network. Alternatively, if a regulatory network topology is known, it can be used to check if that network is, in principle, capable of keeping the system in an optimal state. If not, it might be that there are still
unknown interactions, or that the objective of the system should be reconsidered. If an interaction network is more complicated than strictly necessary, it might imply that there are additional tasks performed by that system.

We will discuss how linear programming can be used to find the minimal gene network structure by considering the example network depicted in Figure 5.3A. Two substrates $S_1$ and $S_2$ are converted in two intermediate metabolites, $m_1$ and $m_2$; these are the precursors of a third intermediate $m_3$, which is finally transformed in a product $P$. In principle, all metabolites can also function as signals to three transcription factors, $TF_1$, $TF_2$ and $TF_3$, which influence the expression of the four enzymes ($e_1$, $e_2$, $e_3$, $e_4$) that catalyse the metabolic interactions. The effect of the transcription factors on the enzymes is mediated by their effect on the particular mRNA, which is not explicitly modelled.

As discussed above, a kinetic model of the system can be used to compute the optimal input-output relation of the regulatory network. Assuming there are no interactions between transcription factors, we can define the gene network interactions as the product of the influence of the metabolites on the transcription factors and the influence of the transcription factors on the enzyme levels:

$$R_{m}^{TF} = r_{TF}^{e} \cdot r_{TF}^{m}$$

For any optimal input-output relation, there is a large (if not infinite) number of matrices $R_{m}^{TF}$ that satisfies Eq. 5.11. The minimal network is defined as the "simplest" topology of the regulatory network that solves Eq. 5.11 for given $\delta e^0$ and $\delta m^0$. Here, by simplest we mean the network with the least number of interactions, i.e. with the least non-zero entries in the matrices $r_{TF}^{e}$ and $r_{TF}^{m}$.

Typically, the optimal response of the regulation network will depend on the perturbation applied and on the external condition, e.g. the external substrate concentration. This means that each condition will in principle give different values for $(\delta e^0, \delta m^0)$, and thus gives a different set of equations. Since environments are typically dynamic, a regulatory network that is only capable of attaining optimal outputs in one particular condition is not of much use. Furthermore, the response of any given regulatory network, optimal or not, will also be condition dependent. However, the structure, or topology, of the network is typically independent of the conditions. Hence, we are interested in the minimal network structure that is able achieve optimal responses for a range of conditions.

In order to find the minimal network topology, we define the interaction topology maps $I_{m}^{TF}$ and $I_{TF}^{e}$, where each entry in these matrices represents a potential interaction between a metabolite and a transcription factor or between a transcription factor and an enzyme, respectively. The elements in the interaction topology map can be either
1, for an activating interaction, -1 for inhibition, or 0 if the interaction is absent. All the entries in the matrices $\mathbf{r}^e_{\text{TF}}$ and $\mathbf{r}^{TF}_m$ are then required to be positive. So, the entries in the interactions topology maps define the type of interaction (inhibiting, activating or not present) whereas the entries in the matrices $\mathbf{r}^e_{\text{TF}}$ and $\mathbf{r}^{TF}_m$ define the strength of these interactions.

We want to formulate this as a mixed integer linear programming (MILP) problem. However, as it is stated here, this is not a linear programming problem, for a number of reasons. Firstly, there will appear nonlinear $r^e_{\text{TF}} \cdot r^{TF}_m$-terms in the optimisation problem. We can reformulate the problem by considering all "paths" between inputs and outputs, and treating them as single variables instead of a combination of two. Each entry in $\mathbf{R}^e_m$ is than a sum of all possible paths between an input and an output.

$$\{\mathbf{R}^e_m\}_{ij} \rightarrow \sum_{\xi \in \text{all TFs}} r^\xi_{ij}$$

where

$$r^\xi_{ij} = \{r^e_{\text{TF}}\}_{ij} \times \{r^{TF}_m\}_{ij}$$

If any of the interactions in the path is not present (i.e. has a zero entry in the interaction topology map), the total response of that path is zero. The simplest way to implement such an absence or presence in a MILP problem is by using indicator constraints.

$$\min : \sum_{\text{all elements}} |\mathbf{I}^e_{\text{TF}}| + |\mathbf{I}^{TF}_m|$$

subject to:

$$\delta e^e_k = \mathbf{R}^e_m \delta m^e_k \quad k \in \text{all conditions}$$

$$\{\mathbf{I}^{TF}_m\}_{ij} = 0 \Rightarrow r^e_{ij} = 0 \quad \forall \xi$$

$$\{\mathbf{I}^e_{\text{TF}}\}_{ij} = 0 \Rightarrow r^{TF}_{ij} = 0 \quad \forall \xi$$

$$0 \leq r^\xi_{ij} \leq r_{\text{max}}$$

$$\{I^\xi\}_{ij} \in \{-1, 0, 1\}$$

The objective function indicates a minimisation of the total number of interactions. The first constraint indicates that the response of the regulatory network should be optimal for all conditions. The second and third constraint indicate that if any interaction in a path is absent, that path will not contribute to a response. The fourth constraint restricts the magnitude of a response. As it is defined here, this is strictly speaking still not a MILP problem with indicator constraints. The reason for this is that the elements of the interaction topology map should be binary, and the objective function should contain real values rather than absolute values. However, this problem can be straightforwardly solved by splitting each potential interaction in an activating and an inhibiting one. We refer to the Methods section for more details and a more formal formulation of the MILP-
problem.

Figure 5.3. Identification of a minimal gene regulatory network capable of optimal regulation of metabolic system. (A) At the metabolic network two substrates $S_1$ and $S_2$ are converted in two intermediate metabolites, $x_1$ and $x_2$; these are the precursors of a third intermediate $x_3$, which is transformed in a product $P$. In principle, all the metabolites could function as signals to three transcription factors, $TF_1$, $TF_2$ and $TF_3$, which influence the expression of the four metabolic enzymes, these interactions are shown by the grey lines. Details about the mathematical model are described in the Methods. (B) Formulation of the full MILP problem used to find the minimal structure of the gene network. (C) Example of the resulting interaction topology maps for the network as depicted in (A). A value of minus and plus one, indicate an inhibitory or activating interaction, respectively. A zero indicate that the interaction is not present. (D) Example of a minimal gene network structure. Green lines indicate activations and red lines represent inhibitory interactions.

Generally, a solution to this problem will not be unique. Most linear programming solvers have functions to generate a set of solutions, rather than just a single solution. The freedom in the solution space can also be used to perform further optimization. An example of an outcome of the MILP optimization problem in terms of the interaction topology maps for the example network is shown in Figure 5.3C. This solution thus corresponds to the minimal number of interactions that can lead to a solution to Eq. 5.11. These interaction topology maps define (i) the number of interactions between metabolites and transcription factors and between transcription factors mRNA’s; (ii) whether
these interactions are inhibiting or activating. The corresponding gene network can then straightforwardly be drawn as shown in Figure 5.3D.

Discussion

The rapid acquisition of experimental data requires quantitative approaches that allow understanding and predictive power of the networks involved. Here, we provide such a theoretical framework that allows for the identification of regulatory gene networks that are able to steer a metabolic network to an optimal state. Our approach is inspired by two concepts that have been frequently used to study the regulation of molecular networks. The first one is MCA. We extended previous findings with respect to optimal enzyme distributions from Heinrich and colleagues (Klipp and Heinrich, 1994; Heinrich and Klipp, 1996; Heinrich and Schuster, 1998; Klipp and Heinrich, 1999) to identify required changes in enzyme and metabolite levels when the optimal metabolic network is perturbed. These changes ensure robustness of the optimal metabolic flux. Subsequently, these changes are exploited to characterise a gene network that is capable of perfect adaptation in varying environments. Using the (second) concept of an interaction map from MRA, we use linear programming to find the minimal structure of such a robust gene network.

Using LP to identify minimal (gene) networks bears several advantages. First of all, since it is a mixed integer linear programming problem, it should in principle be possible to scale it to fairly large systems without running into serious computational problems. Secondly, it is very simple to put prior knowledge about interactions in the model. This can be done by constraining the appropriate interaction term in the LP-problem formulation. There are also a number of issues that are worth pointing out. Since in the linear optimization we only considered the “paths” between inputs and outputs, we can only put bounds on the paths. It would be of particular interest to extend the current framework to include individual reactions such that, for instance, interactions between transcription factors can be taken into account. Furthermore, we consider networks that are capable to maintain an optimal state upon a perturbation. We do not show how the system is brought to an optimal state in the first place. Finally, the perturbations and responses are assumed to be infinitesimal. Specifically the latter two poses a serious challenges before this technique can be used for the rational design of gene regulatory networks.

The strategy to use an input-output relationship to infer network topologies has also been applied by others. However, many of the published studies focus on signalling networks. Becker and colleagues, for instance, have used the erythropoietin receptor network to resolve different modes of information processing through this receptor for a broad range of ligand concentrations (Becker et al, 2010). In another study, the authors aimed to identify topological organising principles that facilitate robust control of intra-
cellular concentrations in the face of multifarious perturbations (Steuer et al, 2011). In this work, we used similar reasoning but applied it to the gene or regulatory networks underlying metabolic networks.

In summary, we have provided a first attempt to identify robust gene networks based on metabolic information only. Although we have illustrated our approach using small illustrative networks, given the successes obtained with genome-scale models using LP in FBA, we expect that our approach should also be applicable to larger networks. Furthermore, the networks that are considered in synthetic biology are typically of similar complexity as the networks considered in this article, therefore we expect this framework to be a valuable tool to find possible structures of robust gene networks that can optimally regulate a metabolic network.