Systems approaches to modelling pathways and networks

Thomas Pfau, Nils Christian, Oliver Ebenhöh

August 18, 2011

Abstract

It has become commonly accepted that systems approaches to biology are of outstanding importance to gain understanding from the vast amount of data which is presently being generated by advancing high-throughput technologies. The diversity of methods to model pathways and networks has significantly expanded over the past two decades. Modern and traditional approaches are equally important and recent activities aim at integrating the advantages of both. This integration is of particular importance since most available methods are specialised to particular systems or problems. The rapid progress of the field of theoretical systems biology, however, demonstrates how our fundamental theoretical understanding of biology is gaining momentum. The scientific community has apparently accepted the challenge to truly understand the principles of life.

Keywords systems biology; mathematicalmodel; metabolism; dynamic systems; genome-scale networks; constraint-based modelling

1 Introduction

The advent of high-throughput technologies in the past decades has drastically changed the nature of the biological sciences. It is now possible to monitor thousands of cellular components, such as transcripts, proteins or metabolites simultaneously, allowing to describe the status of a cell in unprecedented precision for a myriad of conditions or genetic backgrounds. This vast amount of new information is contrasted by a lack of theoretical understanding, confronting us with the problem to convert the gathered knowledge into true understanding of the underlying biological processes. The great challenge for biology in the coming decades will be to develop unifying theories of general validity which will allow to place and interpret the data within an overarching theoretical framework.

The first steps towards developing these theories are taken by the emerging field of systems biology. By describing an investigated biological system with a mathematical model, a theoretical framework is established within which data from many experimental conditions can be interpreted. Every model is by definition a simplified representation of reality. The process of model building itself, in which a biological system is simplified to its essential components and cast into the language of mathematics, is of particular importance, because it allows to discover underlying principles according to which a system functions.

Theoretical approaches to biology are too plentiful to be covered in a single review. Because of its central importance for the functioning of any cell, we place our focus here on metabolic systems and we deliberately omit the numerous graph theoretical approaches to study networks (for a recent review, see Amigó et al. [1]). The purpose of this review is to give a summarising overview of several important traditional and modern techniques to model pathways and networks. Instead of aiming at completeness, we present the reader a variety of different approaches and outline their basic concepts and goals and also stress the limitations of their applicability.

Corresponding Author: Oliver Ebenhöh, Department of Physics, University of Aberdeen, Meston Building, Meston Walk, Aberdeen AB24 3UE, UK.
Telephone: +44 1224 272520. Fax: +44 1224 273105. E-mail: ebenhoeh@abdn.ac.uk

Thomas Pfau is a graduate student at the University of Aberdeen with a degree in computer science. His doctoral research aims at a systems understanding of symbiosis between plants and microbes.

Nils Christian is a postdoctoral researcher at the University of Aberdeen. He graduated in physics and received his doctoral degree in Bioinformatics / Systems Biology.

His research activities comprise large-scale network analysis as well as the dynamic study of metabolic pathways with medical applications.

Oliver Ebenhöh is an expert in large-scale metabolic network analysis and the investigation of evolutionary design principles in pathways and networks with optimality principles. Since 2009 he is a reader for Systems Biology at the University of Aberdeen, where he coordinates the Theoretical Systems Biology research programme at the Institute for Complex Systems and Mathematical Biology.

2 Dynamic, differential equations-based models

The description of biological processes by systems of differential equations has a long standing history. This approach is suitable to describe dynamic changes in which the particle numbers are not too low and fluctuations can be neglected. If this is not the case, stochastic simulations, often based on an approach introduced by Gillespie [2], are applied. We will not cover this important field here but rather refer to a recent review by Ullah and Wolkenhauer [3].

Differential equations define the rate of change of the variable components in a system (e.g. chemical species in a metabolic system, mRNA and proteins in a gene regulatory system or population sizes in an ecosystem). These equations are usually nonlinear and highly coupled which means that the expressions defining the rate of change of one variable depend on other variables. In population dynamics, this approach dates back well into the 19^{th} century, with the logistic growth equation derived by Verhulst in 1838 [4] being a prominent example of a simple model explaining the growth of a population (see Fig. 1A). This model is still widely applied to simulate processes as diverse as the change in languages [5] and the increase in fluorescence during a quantitative PCR measurement [6]. The periodic increase and decline of populations in simple predator-prey systems were independently investigated by Lotka [7] and Volterra [8] and the famous Lotka-Volterra equations still form the basis for many ecosystem models (see e.g. [9, 10]). The equations and their typical behaviour are illustrated in Fig. 1B.

The fundamental Michaelis-Menten rate law is an early example how theoretical considerations promote the understanding of underlying mechanisms [11] in biochemical systems and modern dynamic models of metabolism would be unthinkable without this pioneering work. Before the invention of computers, theoretical analyses were restricted to relatively simple systems which are analytically tractable. The complexity of dynamical systems increases greatly with increasing system size and it is therefore evident that only the accessibility to fast computers allowing for numerical integration of a large number of coupled differential equations enabled a theoretical investigation of more complex metabolic systems. A landmark in the computational analysis of biochemical systems was set by the early work of Garfinkel and Hess [12], who developed a detailed model of the glycolytic pathway.

Numerous mathematical models based on differential equations for the simulation of metabolic pathways, gene regulatory circuits and signalling cascades have since been developed. A major goal of this class of models is to verify existing hypotheses and make new quantitative predictions. If a model based on existing knowledge is capable of qualitatively reproducing time resolved data, then this is a good indicator that the assumptions about the underlying molecular mechanisms are correct. The model can then be used to predict the behaviour of the system when subjected to perturbations, such as gene knock-out or overexpression, application of inhibitors or other drugs. Agreement with experiments further consolidates the existing knowledge and falsification leads to the development of new hypotheses which again feed back to improve the model description. An illustrative example how this mutual stepwise improvement of models and experiments led to the discovery of new genes involved in the plant circadian clock is found in Locke et al. [13].

Interestingly, often the most simplistic models provide the most fundamental understanding into the underlying mechanisms of an observed mode of behaviour. Rapoport et al. [14] have explained with a simple model of the glycolytic pathway how demand feedback regulation generates ATP homeostasis. A drastically reduced model is still able to reproduce the most essential feature of the system, namely its ability to maintain approximately constant levels of ATP for a large range of external ATP consumption rates (see Fig. 2). Homeostasis is generated even if all reactions are modelled as simple mass-action kinetics. Homeostasis is an intrinsic feature of the system: Increased ATP consumption directly leads to reduced ATP levels which automatically increases the level of ADP. Because ADP is a substrate of the ATP producing reactions (reaction v_3 in Fig. 2A), the rate of ATP production increases and the increased consumption is counteracted. Other prominent examples of extremely simple but highly insightful models are the Sel'kov oscillator [15], which explains observed metabolic oscillations by a simple positive feedback loop, and the Goodwin oscillator [16], providing an explanation how oscillations can be produced by an inhibitory feedback loop. These two models are still widely employed and form the basis for many dynamic models for oscillatory biological rhythms including the circadian rhythmic gene expression found in many organisms.

If simple models are those generating most understanding, then how will we ever be able to bridge the gap between such a reductionist approach and the enormous amount of data produced by our modern experimental equipment? One way to involve a large fraction of the available data into our analysis are metabolic networks. The structure of a metabolic network of arbitrary size is conveniently described by the stoichiometry matrix, denoted $\mathbf{N}$. If there are m metabolites connected by r reactions, this matrix has m rows and r columns with an entry n_{ij} denoting the stoichiometric coefficient of metabolite i in reaction j. n_{ij} is negative if this metabolite is

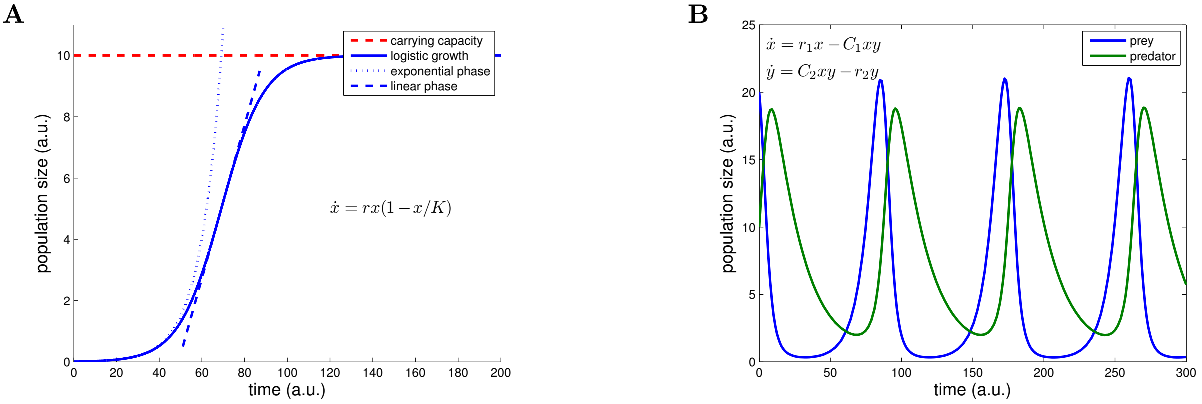


Figure 1: **A** The logistic growth model is determined by two parameters. The maximal growth rate for small population sizes is r and the maximal carrying capacity of the environment is K. The growth exhibits three phases. For small population sizes, growth is approximately exponential (see dotted line). This phase is followed by a period in which growth is approximately linear before growth becomes saturated and the population approaches its maximal size, determined by the parameter K. **B** The Lotka-Volterra equations describe oscillations in a simple ecosystem of one predator and one prey species. The parameter r_1 denotes the relative growth rate of the prey in the absence of predators, r_2 denotes the death rate of the predators. The parameters C_1 and C_2 describe the interaction of predators and prey. C_1 indicates the rate of reduction of prey individuals while C_2 describes the resulting rate of increase of predator individuals.

consumed, positive if it is produced and zero if its concentration does not change through reaction j. The dynamics of the network is characterised by the differential equation system

$$\frac{d\mathbf{S}}{dt} = \mathbf{N} \cdot \mathbf{v}(\mathbf{S}, \mathbf{p}), \qquad (1)$$

where $\mathbf{S}$ is the vector of metabolite concentrations, $\mathbf{v}$ is the vector of reaction rates and $\mathbf{p}$ contains the system parameters.

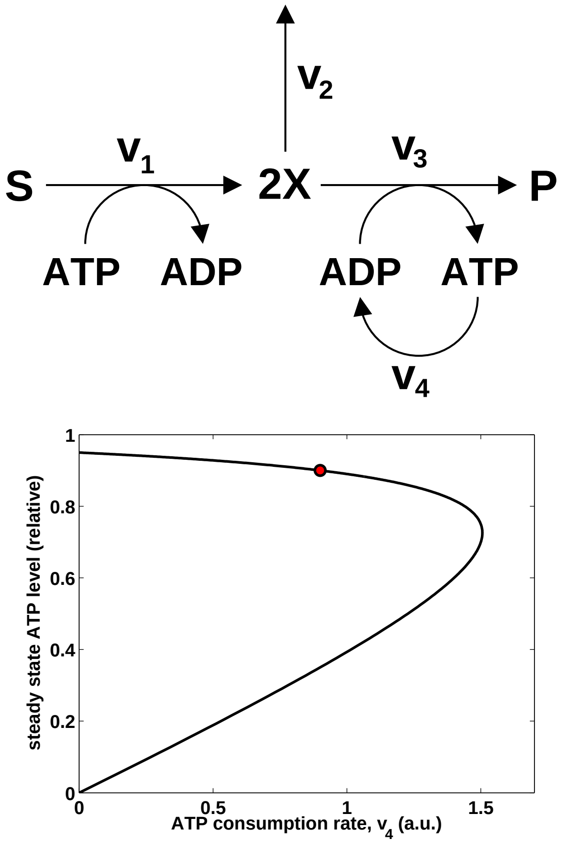
However, a simple up-scaling of this modelling approach to genome-scale is impracticable, despite the fact that modern PCs could easily integrate such systems of thousands of differential equations. The first reason is the lack of knowledge of most enzymatic parameters. But even if for all enzymes these parameters had been determined *in vitro*, uncertainty in enzyme concentrations and their post-translational modifications remains. The second reason lies in the difficulty of simultaneously fitting many parameters to noisy high-throughput data. It is hard to ensure that the model is fitted to the actual data and not to the noise - a process called overfitting. The third, and maybe most fundamental, reason is that a thorough analysis of such a model is hardly possible considering the multitude of degrees of freedom, leaving the unsolved question what would actually be learnt about the biological system with such an immensely complicated model.

3 Metabolic Control Analysis

The analysis of large-scale systems is greatly simplified by only considering stationary conditions. For metabolic systems, stationarity is often a realistic assumption because biochemical reactions are fast compared to changes in protein or gene expression levels. Assuming stationarity simplifies Eq. (1) to

$$\mathbf{N} \cdot \mathbf{v}(\mathbf{S}, \mathbf{p}) = 0. \qquad (2)$$

A systematic analysis of how the possible stationary concentrations and fluxes are restricted by this condition has lead to a theory which has become to be known as *Metabolic Control Analysis* (MCA). The conceptual framework has been independently developed by Kacser and Burns [17] and Heinrich and Rapoport [18]. A central component of this theory are quantities called *Control Coefficients* which describe the relative change of steady-state values upon small perturbation of system parameters (for an illustration see Fig. 3). Unfortunately, control coefficients are difficult to determine experimentally, making a direct application of MCA often hard. Nevertheless, the theory provides fundamental insights which are applicable to systems of arbitrary size. The summation and connectivity theorems are rare examples where properties of biological systems can be rigorously deducted by a mathematical proof (for their derivation, see e.g. Heinrich and Schuster [19]). The fact that flux control coefficients for one particular flux sum up to one explains why often one rate limiting step is observed, namely if one control coefficient is close to one while the others are near zero. However, it also demonstrates that there is no inherent reason why this should be the case and that, more generally, the control is distributed. Concentration control coefficients sum to zero, showing that control on a particular metabolite


S
v1
2X
v2
v3
P
ATP
ADP
ADP
ATP
v4
steady state ATP level (relative)
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
ATP consumption rate, v4 (a.u.)

Stoichiometry matrix:

$$\mathbf{N} = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{cccc} R_1 & R_2 & R_3 & R_4 \end{array}$$

$$\mathbf{N} = \left(\begin{array}{cccc} +2 & -1 & -1 & 0 \\ -1 & 0 & +1 & -1 \\ +1 & 0 & -1 & +1 \end{array}\right) \begin{array}{c} X \\ ATP \\ ADP \end{array}$$

Dynamic equations:

$$\frac{d}{dt}\begin{pmatrix} X \\ ATP \\ ADP \end{pmatrix} = \mathbf{N}\cdot\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

Conserved moieties: $\mathbf{C}\cdot\mathbf{N} = \mathbf{0}$

$$\mathbf{C} = (0\,1\,1) \Leftrightarrow \frac{d}{dt}(ATP + ADP) = 0$$

$$\Leftrightarrow ATP + ADP = A = \text{const.}$$

Steady state flux distributions: $\mathbf{N}\cdot\mathbf{K} = 0$

$$\mathbf{K} = (\mathbf{v_1}\,\mathbf{v_2}),\ \mathbf{v_1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix},\ \mathbf{v_2} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 1 \end{pmatrix}$$

Figure 2: A simple model of glycolysis can explain ATP homeostasis. **A** Schematic representation of the model. The upper part of glycolysis is represented by a single reaction (v_1) in which one molecule of ATP is consumed per substrate molecule (S, corresponding to glucose). Reaction v_2 takes into account consumption of the intermediate X (corresponding to triose phosphates) for biosynthesis pathways. Reaction v_3 lumps the lower part of glycolysis into one reaction producing one molecule of ATP per intermediate X. The reaction v_4 comprises all ATP consuming processes. **B** The stationary ATP concentration varies only slightly for changed ATP consumption rates. All reaction rates have been assumed to follow mass-action kinetics: $v_1 = k_1 \cdot ATP$, $v2 = k_2 X$, $v_3 = k_3 X \cdot ADP$, $v_4 = k_4 \cdot ATP$. The circle indicates the operation point of many tissues *in vivo*. In this region, the curve has a small negative slope, meaning that increased ATP consumption will lead to a small reduction of the stationary ATP level while reduced consumption will lead to a small increase. **C** A stoichiometric analysis of this model illustrates how the dynamic equations and stoichiometry matrix **N** are connected. Conserved moieties can be calculated from the left-sided kernel of **N**, the possible flux distributions are determined by the nullspace of **N**.

concentration *must* be distributed with positive and negative contributions balancing each other. Most importantly, control is determined by the network as a whole and can often not easily be explained by considering the single isolated steps. For example, the observation that upon perturbation a substrate of a particular enzyme is downregulated while its product is simultaneously upregulated may intuitively seem to entail that this enzyme must be under allosteric regulation. However, metabolic control theory shows us that this view is too simplistic. Numerous counter-examples for which this so-called *Crossover Theorem* is invalid are given in Heinrich and Rapoport [20]. Remarkably, and possibly because control coefficients are not directly measurable, this over 30 year-old knowledge is not yet established in experimental research and the straightforward but flawed conjecture is still often found when metabolomics data are interpreted.

4 Nullspace Analysis

Eq. (2) implicitly determines the steady-state concentrations as functions of the parameters. Determination of **S** requires detailed knowledge of all enzymatic rate laws and closed expressions can usually not be derived. However, possible solutions for the vector **v** of flux distributions are simply obtained by solving the equation

$$\mathbf{N}\cdot\mathbf{v} = \mathbf{0}\ , \tag{3}$$

In steady state: $\mathbf{N} \cdot \mathbf{v}(\mathbf{S}, \mathbf{p})$. This condition implicitly defines the dependence of steady-state concentrations $\mathbf{S} = \mathbf{S}(\mathbf{p})$ and fluxes $\mathbf{J} = \mathbf{J}(\mathbf{p}) = \mathbf{v}(\mathbf{S}(\mathbf{p}), \mathbf{p})$ on the parameter values. *Control coefficients* quantify how these values change upon parameter perturbation. Concentration and flux control coefficients are defined as

$$C^S_{ij} = \frac{v_j}{S_i} \frac{\partial S_i}{\partial v_j} \text{ and } C^J_{ik} = \frac{v_k}{J_i} \frac{\partial J_i}{\partial v_k}.$$

Elasticities are enzyme properties and independent on the network. They quantify how changes in metabolite concentrations affect enzymatic rates:

$$\varepsilon_{ij} = \frac{S_j}{v_i} \frac{\partial v_i}{\partial S_j}.$$

The *Summation Theorems* assert that

$$\sum_j C^S_{ij} = 0 \text{ and } \sum_k C^J_{ik} = 1.$$

The *Connectivity Theorems* connect control coefficients and elasticities by

$$\sum_j C^S_{ij} \varepsilon_{jk} = -\delta_{ik} \text{ and } \sum_j C^J_{ij} \varepsilon_{jk} = 0,$$

where $\delta_{ik} = 1$ for $i = k$ and zero otherwise.

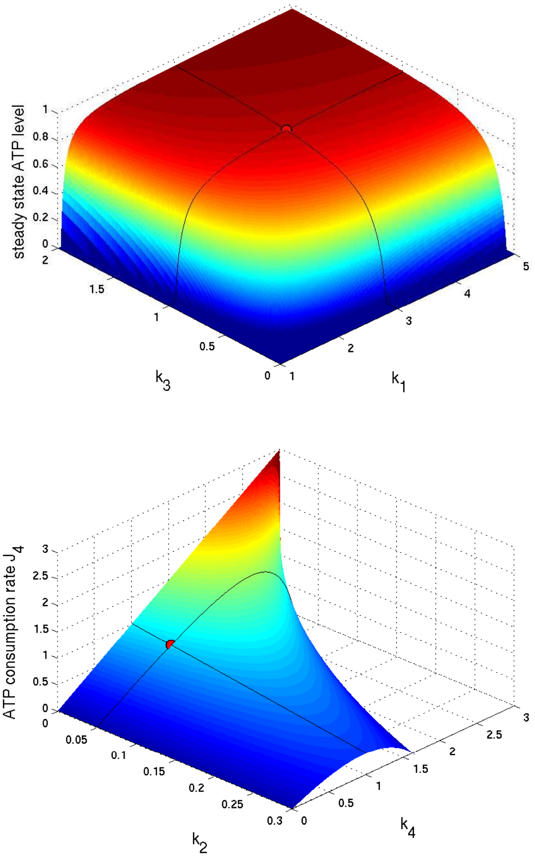


Figure 3: **A** Metabolic Control Analysis in a nutshell. **B** The stationary ATP level of the simple glycolytic model in Fig. 2 is plotted in dependence of the parameters k_1 and k_3. The dot indicates the reference state from Fig. 2B. The black lines indicate the change of a single parameter. The control coefficients can be interpreted as the slope of these lines at the reference point. Here: $C^{ATP}_{v_1} = 1/12$ and $C^{ATP}_{v_3} = 1/9$. **C** The stationary flux through reaction 4 (ATP consumption rate) is plotted as function of the parameters k_2 and k_4. The same reference state as in panel B is indicated. Again, the control coefficients correspond to the slopes of the black lines at this point. Here: $C^{J_4}_{v_2} = -1/9$ and $C^{J_4}_{v_4} = 11/12$.

which depends only on the structure of the metabolic network, encoded by the stoichiometry matrix, and not on any additional information. As illustrated in the Box of Fig. 2, the (right) nullspace of $\mathbf{N}$ contains all solution vectors $\mathbf{v}$. Similarly, by calculating the left nullspace all conserved moieties can be determined [21]. For these calculations efficient algorithms are abundant, making them feasible even for genome-scale networks.

Mathematically, the nullspace is characterised by the kernel matrix $\mathbf{K}$ containing as columns all linearly independent solution vectors $\mathbf{v}$ satisfying condition (3). If the kernel matrix contains (up to a scalar factor) identical rows (each row corresponds to a particular reaction) then this means that the fluxes through these reactions always have a fixed ratio, regardless of the exact flux distribution. Such sets of reactions (or the set of enzymes catalysing these reactrions) are called *Enzyme Subsets* [22].

This concept is generalised by the *Reaction Correlation Coefficients*, introduced by Poolman et al. in 2007. Instead of identifying only strictly identical row vectors (up to a multiplicative constant), similarity of vectors is determined. This similarity can reveal highly correlated reactions which could be missed if only enzyme subsets are considered. The problem that the kernel matrix is not unique is solved by replacing the column vectors $\mathbf{v}$ by vectors of an orthogonal basis of the nullspace. This basis is also not unique but the angle between the row vectors is independent on the specific choice. The reaction correlation coefficient of two reactions is defined as the cosine of the angle between the corresponding two row vectors. If this value is 1 or -1 the reactions belong to an enzyme subset, if it is zero, the fluxes are maximally uncorrelated. This method was employed to search for closely related reactions in *Escherichia coli* and *Streptomyces coelicolor* [23].

5 Pathway Identification

The analysis outlined above does not take into account that due to thermodynamic constraints some reactions are practically irreversible. Mathematically the irreversibility of reactions imposes a further constraint on the solution vectors $\mathbf{v}$ of the steady state condition (3). Only those vectors for which the entries corresponding to irreversible reactions are non-negative are of biological relevance,

$$\mathbf{N} \cdot \mathbf{v} = \mathbf{0} \text{ with } v_i \geq 0 \text{ for } i \in I, \tag{4}$$

where I denotes the subset of irreversible reactions. A systematic description of this more complicated solution space is provided by the concept of *Elementary Flux Modes* (EFM) [24, 25]. An EFM is a flux distribution which fulfils the steady-state condition (4) and is minimal in the sense that no reaction carrying a flux can be removed without violating this condition. Furthermore the elementarity refers to the property that no EFM may contain another EFM. EFMs can be interpreted as the most elementary pathways of a metabolic system. They provide concise information about the metabolic network because they describe the possible modes of operation of the system.

The sometimes counter-intuitive results of EFM analysis have been shown by de Figueiredo et al. [26]. A model of the glycolytic pathway and the Krebs cycle has been tested for the ability to produce sugars from fatty acids. A stripped-down version of this pathway, together with its EFM are shown in Fig. 4 and Tab. 1. Intuitively one might guess that acetyl-CoA (AcCoA), the breakdown product of fatty acids, can be converted by this pathway to Glucose 6-phosphate (G6P). EFM analysis demonstrates that this is only possible if the glyoxylate shunt is present, explaining why plants can convert fat to sugars, while animals cannot. The glyoxylate shunt bypasses the CO_2-releasing steps in the citric acid cycle. Without it, both carbons of the acetyl group are lost as CO_2 and thus no net-flux to sugars is possible.

A related concept to EFMs are *Extreme Pathways* (EP) [27, 28]. For the calculation of EPs it is assumed that all reactions are irreversible. To account for reversible reactions, these are represented as two reactions operating in opposite directions. The mathematical advantage is that the solution space containing all feasible flux vectors is now restricted to vectors with non-negative entries only, the disadvantage is an increased number of reactions. EPs represent a convex linearly independent generating set of the solution space, meaning that all feasible flux distributions can uniquely be described as a non-negative linear combination of EPs. A good overview over mathematical properties of both EFMs and EPs was recently published by Jevremovic et al. [29].

A complementary concept to EFMs are *Minimal Cut Sets* (MCS) introduced by Klamt and Gilles [30]. Whereas EFMs are minimal pathways, an MCS is a minimal set of reactions that need to be removed to inactivate a specified target reaction. An MCS is minimal in the sense that removing any subset of it from the network is not sufficient to inactivate the target reaction. MCSs are useful to predict sets of genes which should be knocked out in order to inactivate a particular metabolic route. Klamt [31] could show that MCS are a dual representation of EFMs, i.e. all MCS can be uniquely determined from the EFMs and vice versa. Both theories are thus representing the same mathematical problem with different applications in biology.

EFMs, EPs and MCSs provide a useful and illustrative way to characterise the properties of a metabolic network, based on the stoichiometry of the network alone. The main limitation is a combinatorial explosion of their numbers for large network sizes [32]. For genome-scale metabolic networks millions of EFMs would be expected. This leads to a computational problem for algorithms calculating all EFMs and EPs simultaneously from the stoichiometry matrix [27, 22, 33]. To overcome this, de Figueiredo et al. recently developed an alternative method using constraint-based linear programming to determine “k-shortest” EFMs [34]. This method calculates EFMs in an iterative fashion to compute subsets of the full set of EFMs. The generated sets can be restricted to include only EFMs containing specified reactions.

Table 1: Elementary flux modes of the model illustrated in Fig. 4 with glyoxylate shunt.

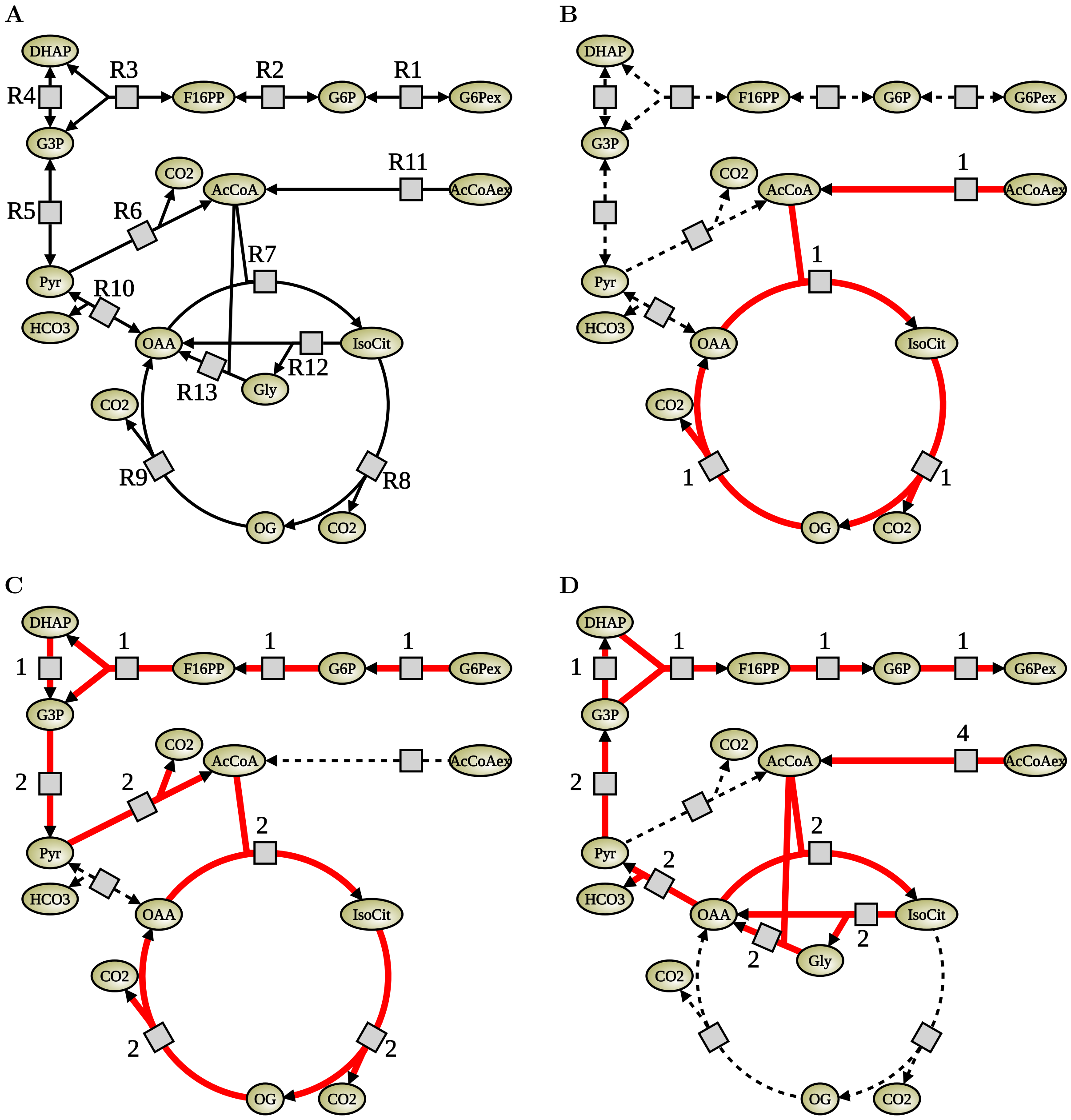


Figure 4: Glycolysis and Krebs cycle. **A** A visual inspection of the model might intuitively suggest that acetyl-CoA can be transformed to glucose 6-phosphate even in the absence of the glyoxylate shunt (R12, R13). **B** The only elementary flux mode of this system with net consumption of acetyl-CoA: Two molecules CO_2 are released, no sugars can be produced. **C** The only other EFM respires sugars, converting one molecule glucose 6-phosphate into 6 CO_2. **D** With the glyoxylate shunt present, there exists one EFM with a net consumption of 4 acetyl-CoA, releasing two bicarbonates and producing one molecule glucose 6-phosphate. Conversion of fat to sugars is possible.

Reaction	EFM					Enzyme subset
	1	2	3	4	5	
R1	0	-1	-1	1	0	1
R2	0	1	1	-1	0	
R3	0	-1	-1	1	0	
R4	0	-1	-1	1	0	
R5	0	2	2	-2	0	
R10	0	0	-2	-2	-1	2
R12	0	0	2	2	1	
R13	0	0	2	2	1	
R8	1	2	0	0	0	3
R9	1	2	0	0	0	
R6	0	2	4	0	1	none
R7	1	2	2	2	1	

The reactions are grouped by enzyme subsets. The ratios of the fluxes through reactions in an enzyme subset are always fixed. The lower part of the Table indicates the net turnover of external metabolites for each EFM.

6 Constraint-based Models

Only a small fraction of all possible flux distributions will actually be realised *in vivo.* Instead of aiming at a complete characterisation of the solution space, it is plausible to search for those flux distributions which perform optimally with respect to a certain criterion. This goal is pursued by *Constraint-Based Modelling* (CBM), also called *Flux Balance Analysis* (FBA), a general approach in which sensible constraints to reduce the number of possible flux distributions are defined and those solutions optimising a specified objective function are identified. The irreversibility conditions in Eq. 4 are generalised to include lower and upper bounds for every involved reaction:

$$\mathbf{N} \cdot \mathbf{v} = 0 \text{ with } {v_i}^{\text{low}} \leq v_i \leq {v_i}^{\text{up}}. \tag{5}$$

These bounds may represent minimal and maximal enzymatic rates or can be used to restrict the flux solutions to those which comply with experimentally measured fluxes, such as nutrient uptake rates. Further, some objective function $Z = Z(\mathbf{v})$ is defined and the optimisation problem

$$\text{maximise/minimise } Z \text{ under the constraints (5)} \tag{6}$$

is solved. Since this strategy has recently been thoroughly reviewed by Gianchandani et al. [35] and Ruppin et al. [36], we will only give a brief account on the most essential ideas.

If the function Z is a linear function of all fluxes $\mathbf{v}$, the optimal solution can efficiently be identified with linear programming [37]. While the computation of an optimal flux distribution is rather straight-forward, it is by no means apparent which objective function should be employed, because it is unknown according to which optimality principles fluxes are arranged *in vivo.* Often those fluxes are identified which optimise biomass yield [38]. This approach leads to a remarkably accurate prediction of fluxes in *Escherichia coli* for various external constraints and can distinguish between genes essential and not essential for growth in most growth conditions [38]. While *E. coli* might indeed have optimised its growth rate, this assumption is apparently unrealistic for other, especially multi-cellular, organisms (for a review on the biomass objective function, see [39]). Other plausible objective functions consider finding fluxes maximising ATP production [40, 41, 42], minimising the overall flux through enzymatic reactions [43] or minimising a set of fluxes given partial experimental flux data [44, 45]. Every particular objective represents a different assumption about the metabolic network and different objective functions will in general yield different optimal flux distributions. However, optimised flux distributions are very informative to characterise the general abilities of the underlying network. Further, if a certain objective function produces realistic flux distributions, the hypothesis is supported that the corresponding organism has indeed evolved to optimise this function. Solving the optimisation problem (6) results in a single solution, but several solutions may exist which exhibit identical or almost identical values of the objective function Z. To identify alternative optimal solutions, approaches based on mixed integer linear programming (MILP) [46] have been developed [47, 48].

A conceptual problem with CBM is that constraints (5) do not consider cellular growth. Finding a flux distribution balancing all intermediates does not guarantee that the intermediates can actually be replenished when diluted during growth. This problem was first addressed by Kruse and Ebenhöh [49] and the notion of *sustainable* metabolites was introduced to describe that a net production of a metabolite is possible under constant dilution. To account for dilution, Benyamini et al. [50] proposed *Metabolic Dilution Flux Balance Analysis* (MD-FBA). In MD-FBA each metabolite produced in any flux carrying reaction will have an associated positive dilution value which is incorporated into the steady state assumption. The method was shown to improve gene essentiality prediction especially because it identifies genes involved in the biosynthesis of cofactors as essential, which are not predicted by normal FBA approaches.

An interesting application of CBMs is the attempt to predict changes in flux distributions upon gene knock-out. A gene knock-out is simulated simply by forcing all reactions which are catalysed by proteins coded by the respective gene to zero. Segrè et al. [51] propose the method of *Minimisation Of Metabolic Adjustment* (MOMA), which is based on the assumption that an organism will initially try to adjust to a knockout with the minimum possible effort and that consequently the metabolic fluxes in the perturbed system are as close as possible to the original flux distribution. *Regulatory On/Off Minimisation* (ROOM) [52] pursues a similar goal by assuming that the number of significant flux changes (on/off) is minimal. Interestingly, MOMA is very successful in predicting the adapted fluxes a short time after genetic perturbations while ROOM yields better predictions for long-term adaptations.

The ability to predict knockouts maximising the production rates of particular metabolites is of interest for biotechnological applications. The OptKnock algorithm [53] assumes that after gene knock-out metabolic fluxes are rearranged to optimise biomass accumulation under the perturbed conditions. It then employs MILP to identify sets of genes whose knock-out results in a maximal production rate of a product of interest. Tepper and Shlomi argued that the assumption of maximal biomass production for modified organisms is not necessarily justified and proposed the RobustKnock strategy [54] which identifies sets of genes that upon knock-out lead to a solution space containing no solutions without a production of the target metabolite. In this way it is ensured that the reduced networks will always produce some amounts of this target, regardless of the actual flux distribution that is realised.

7 Thermodynamic constraints

Flux predictions based on the structure of the metabolic network alone ignore that biochemical fluxes must obey fundamental laws of thermodynamics. As a consequence, the identified solution spaces contain fluxes which are thermodynamically unfeasible. An attempt to incorporate thermodynamic constraints was introduced by Beard et al. [55] as the concept of *Energy Balance Analysis* (EBA). While the mass-balance constraint as defined by Eq. (3) is a chemical analogon to Kirchhoff's current law in electric circuits, EBA imposes additional constraints which essentially follow from the second law of thermodynamics and are a chemical analogon to Kirchhoff's loop law. This approach identifies flux distributions which are thermodynamically unfeasible as a result of the network structure alone and in particular rules out series of cyclic reactions without net conversion. A disadvantage is that the additional constraints are non-linear and therefore the calculation becomes computationally challenging for large networks.

To restrict the solution space even further, additional information is required. The idea of *Thermodynamics-Based Metabolic Flux Analysis* (TMFA) [56] incorporates thermodynamic constraints as additional linear constraints and depends on the knowledge of standard enthalpies of reactions. The formulation of the problem as MILP includes the metabolic activities as new variables which underlie the optimisation procedure. Essentially, metabolite ranges are identified which are in accordance with thermodynamic constraints under the condition of maximal bacterial growth. Similar thermodynamic constraints are applied in Hoppe et al. [57] with the important difference that, instead of maximising biomass accumulation, a feasible flux distribution is determined for which the metabolite concentrations deviate minimally from set-point values defined by biochemical knowledge and simultaneously the overall flux is minimal. This approach outlines a possible route how experimentally determined metabolite concentrations can be incorporated in CBM: If absolute concentrations for a large number of metabolites are known, these can be used to define further constraints.

A promising approach which integrates steady-state mass conservation, energy conservation, the second law of thermodynamics and reversible enzyme kinetics into a single set of equalities and inequalities has been proposed by Fleming et al. [58]. Here, the changes of chemical potentials are derived from the elementary steps of the enzymatic kinetics, and forward and backward reactions are treated individually. The method includes in a very general approach fluxes, kinetic parameters, enzyme and metabolite concentrations. It has been shown that in principle this *Integrated stoichiometric, thermodynamic and kinetic constraint-based modelling* approach is applicable to networks of moderate size. However, the application to larger systems is still in its infancy.

8 Integrative approaches

The discussion above of the existing methods to investigate genome-scale metabolic networks and to predict and understand their metabolic functions reveals the difficulty in integrating different sources of data. Some of the thermodynamic approaches clearly show the potential to include metabolite data to further constrain the solution space containing all possible flux distributions. But how can the multitude of other high-throughput data, in particular transcriptomic and proteomic profiles be included to increase our understanding of metabolic networks?

An innovative approach to tackle this problem was taken by Yizhak et al. [59]. The introduced method, termed *Integrative Omics-Metabolic Analysis* (IOMA), represents a further development of the thermodynamic approaches discussed above and includes available proteomics and metabolomics data, while allowing for missing information. It is formulated as a quadratic programming (QP) problem in which steady-state flux distributions are identified that display the highest possible agreement with experimental proteomics and metabolomics data and kinetically derived flux estimations. It has been applied to predict the metabolic state of human erythrocytes upon genetic alterations and the comparison with kinetic model results has shown excellent agreement. The application to a genome-scale network of *E. coli* knock-out lines for which proteomic and metabolomic data were available has shown a considerable improvement of the quality of the predictions when compared to MOMA or the straightforward flux balance approach which assumes optimised growth rates.

For most applications of CBM, a genome-scale metabolic network model can be used and is commonly derived from annotations of the genome sequence. However, it is obvious that under different conditions different parts of this network are activated. Especially in a multicellular organism, the active parts of the network may be drastically different for different tissues. A possibility to exploit transcriptomic and proteomic data to infer the active subnetworks has recently been proposed in Jerby et al. [60]. Their constraint-based method is formulated as a MILP which optimises the activity pattern of the network such that it resembles the experimental observations as accurate as possible, while obeying the constraint that the active network is consistent in the sense that biomass can be produced and all reactions may be activated under steady-state conditions. By this strategy, the algorithm was able to automatically generate tissue-specific subnetworks based on high-throughput data for different tissues of the human body.

9 Network Expansion

The success of the computational methods for the investigation of genome-scale metabolic network models is apparent. For a review on over 60 papers published on genome-scale models of *E. coli* alone, see Feist and Palsson [61]. A bottleneck for the application to a wider spectrum of organisms is that all constraint-based approaches require highly accurate and well-curated networks. This process still involves a large amount of manual labour and the construction of a consistent network model from the genome sequence easily takes several months despite the fact that many steps can be performed in a semi-automated fashion [62]. This labour-intensive generation of network models explains why the numbers of sequenced genomes, with over 1600 completed and over 8000 ongoing [63], is in stark contrast with an estimated 50 published genome-scale models [59].

An alternative approach to study functional and structural properties of large-scale metabolic networks is the so-called *Method of Network Expansion* [64, 65]. The method determines which metabolites are in principle producible by the metabolic network. Given a predefined set of metabolites, the seed, the underlying algorithm generates a series of expanding networks by stepwise adding all reactions for which all substrates are either present in the seed or are products of previously added reactions. The process ends when no further reactions can be added. The set of metabolites contained in the final network is called the *scope* of the seed. Due to the simplicity of the underlying algorithm, this method is computationally highly efficient. In contrast to constraint-based models this method can only produce qualitative predictions regarding the principle capacity of a network to produce metabolites. Another disadvantage of this method is the fact that metabolites that are required for their own production (such as ATP in glycolysis) cannot be handled in a straight-forward way. This problem is overcome by the introduction of an additional mechanism based on the observation that many of these self-depending compounds are cofactors. These cofactors mostly occur in pairs on both sides of a reaction equation and donate or take up a chemical group (e.g. ATP - ADP for the transfer of phosphate groups or NADH - NAD$^+$ for the transfer of electrons). These pairs are either known by biochemical knowledge or can be identified with heuristics. These cofactors are then added to the seed in a way which ensures that they can only act in their role as cofactors and that no metabolite can be created from nothing. The strength of this method is that it is far more error-tolerant against inaccuracies of the underlying network and yields stable results for networks which are retrieved from databases such as KEGG [66] or MetaCyc [67] and subsequently curated only moderately in an automated process [68]. The applicability to database-derived networks makes this method a good choice to systematically compare hundreds of networks. In Ebenhöh and Handorf [69], the concept of *Carbon Utilisation Spectra* was introduced to to characterise the biosynthetic capabilities for over 400 organism-specific networks if only one carbon source but abundant inorganic material is available. This analysis has revealed that it is to a certain extent possible to infer the lifestyle of an organism from its metabolic network structure alone.

The fast computational power can be exploited to perform thousands or even millions of scope calculations on a simple PC. In Handorf et al. [70], a greedy search algorithm was employed to identify minimal combinations of nutrients which must be supplied to a network such that it can produce all essential biomass precursors. The large number of solutions were restricted to biologically relevant combinations by applying heuristics which ensure that small metabolites and those for which transporters have been characterised are preferentially included in a solution. The systematic comparison of predicted nutrient types revealed that it is possible to distinguish between generalist and specialist species.

An approach to integrate genomic, proteomic and metabolomic data was developed in Christian et al. [71] to identify and fill gaps in metabolic networks. For this, a draft network derived from a genome sequence is embedded in a reference network derived from MetaCyc. Subsequently, minimal sets of reactions are identified which have to be added to the draft network to make it consistent with experimental observations. Consistency is assumed if the network is capable of producing all required biomass precursors as well as all experimentally identified metabolites from the applied growth medium. This approach led to an improved genome annotation of the green alga *Chlamydomonas reinhardtii* [72]. Similar attempts to automatically identify missing reactions were also developed based on CBM, see e.g. Reed et al. [73]. However, for their successful application both the draft and the reference network need already be cleaned from all stoichiometric inconsistencies.

The method of network expansion has also been applied to gain insights about the evolution of metabolic pathways. In Raymond and Segrè [74] the effect of the appearance of oxygen in earth's atmosphere on the complexity of biochemical networks was investigated and it was found that all organisms can be classified into four groups of increasing metabolic complexity. A modified expansion algorithm was employed in Schütte et al. [75]. Here, the expansion proceeds by single reactions only and preferably those were attached to the network for which the protein sequences show the greatest similarity to sequences already present in the network. This study provided computational evidence for a punctuated equilibrium in molecular evolution, indicating that the evolution of new enzymes rather occurred in short periods in which the rate of invention was high with intermittent long silent intervals than in a gradual fashion.

10 Future Challenges

We have discussed a number of methods to analyse and interpret pathways and networks and demonstrated how the application of traditional as well as newly developed theoretical methods can considerably advance our understanding of biological systems. Evidently, we are still far from the goal of reaching a truly systemic understanding of a cellular system as a whole. However, the concepts discussed here may well be the seed for the development of unifying biological theories to answer the question 'What is life?' [76].

The grand challenge for the coming years and decades will be to unify these methods in order to allow for a seamless description across scales, from molecules to whole-systemic behaviour. A prerequisite which needs to be addressed urgently is the definition of standardised descriptors within models. While the Systems Biology Markup Language (SBML) [77] was developed as a standard model format with the aim to allow for an easy exchange of models, it does not solve the problem of unique identifiers for metabolites or proteins. Even though there has been effort to generate links between different databases by defining minimal information that annotations of biochemical models should contain [e.g. MIRIAM, 78], it can still be laborious to combine different models based on different databases. In particular models which contain non standardized compounds or proteins inferred from local experiments are often hard to handle.

A conceptual challenge will be the integration of different levels of biological networks. Obviously, the metabolic state of a cell is controlled through gene regulatory networks and, in turn, the metabolic state influences the activities of genes. Recent publications show that this problem is addressed on various scales. A classical, dynamical systems approach is followed by Baldazzi et al. [79], where metabolism is treated as a quasi stationary subsystem and, for small systems, the signs of indirect regulations could be analytically calculated, but the applicability to large-scale systems is difficult. Simulating the gene regulatory network of *E. coli* metabolism by a Boolean network [80] indicated a high degree of flexibility and the method seems suitable to be integrated with network expansion to investigate biosynthetic capabilities for different regulatory states. The strict boolean on/off conditions are relaxed in the recently presented method called *Probabilistic Regulation of Metabolism* [81] by which transcription regulatory networks and genome-scale metabolic models can be integrated with high-throughput data.

Despite the impressive existing repertoire of mathematical methods, it is clear that existing theories are still far from providing a unifying description of the phenomenon 'life'. The rapid recent development of modelling approaches demonstrates, however, that the systems biology research community has accepted the great challenge and has embarked on the quest to understand the principles of life.

Summary key points:

1. Theoretical deliberations are the key to convert biological knowledge in understanding.

2. For different system sizes and different biological problems, numerous modelling approaches exist. Dynamic differential equations-based models are useful for the investigation of systems of relatively small size. To study genome-scale metabolic networks, constraint-based models have been developed which omit the dynamic component by assuming a stationary state.

3. It will be a key challenge to integrate these approaches to arrive at a seamless description of biological processes across scales. The rapid development of systems approaches to biology give rise to optimisim that the scientific community is on the right track to develop a thorough understanding of the phenomenon life.

References

[1] Amigó JM, Glvez J, and Villar VM. A review on molecular topology: applying graph theory to drug discovery and design. *Naturwissenschaften* (2009), **96**(7), 749–761.

[2] Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* (1977), **81**(25), 2340–2361.

[3] Ullah M and Wolkenhauer O. Stochastic approaches in systems biology. *Wiley Interdiscip Rev Syst Biol Med* (2010), **2**(4), 385–397.

[4] Verhulst PF. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathmatique et physique* (1838), **10**, 113–121.

[5] Zuraw K (2003). *Probabilistic Linguistics*, chapter 5: Probability in Language Change, pages 139–176. The MIT Press, Cambridge, MA, USA.

[6] Boggy GJ and Woolf PJ. A mechanistic model of PCR for accurate quantification of quantitative PCR data. *PLoS One* (2010), **5**(8), e12355.

[7] Lotka AJ. Analytical note on certain rhythmic relations in organic systems. *Proc Natl Acad Sci U S A* (1920), **6**(7), 410–415.

[8] Volterra V. Variazioni e fluttuazioni del numero dindividui in specie animali conviventi. *Mem. Acad. Lincei Roma* (1926), **2**, 31–113.

[9] Berryman AA. The origin and evolutoin of predator-prey theory. *Ecology* (1992), **73**, 1530–1535.

[10] Brauer F and Castillo-Chavez C. (2000). *Mathematical Models in Population Biology and Epidemiology.* Springer, Heidelberg, Germany.

[11] Michaelis L and Menten M. Kinetik der Invertinwirkung. *Biochem. Z.* (1913), **49**, 333369.

[12] Garfinkel D and Hess B. Metabolic control mechanisms. VII.A detailed computer model of the glycolytic pathway in Ascites cells. *J Biol Chem* (1964), **239**, 971–983.

[13] Locke JCW, Kozma-Bognr L, Gould PD, *et al.* Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Mol Syst Biol* (2006), **2**, 59.

[14] Rapoport TA, Heinrich R, and Rapoport SM. The regulatory principles of glycolysis in erythrocytes in vivo and in vitro. *Biochem J* (1976), **154**(2), 449–469.

[15] Sel'kov EE. Self-oscillations in glycolysis. 1. A simple kinetic model. *Eur J Biochem* (1968), **4**(1), 79–86.

[16] Goodwin BC. Oscillatory behavior in enzymatic control processes. *Adv Enzyme Regul* (1965), **3**, 425–438.

[17] Kacser H and Burns JA. The control of flux. *Symp Soc Exp Biol* (1973), **27**, 65–104.

[18] Heinrich R and Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem* (1974), **42**(1), 89–95.

[19] Heinrich R and Schuster S. (1996). *The Regulation of Cellular Systems.* Chapman & Hall, London, UK.

[20] Heinrich R and Rapoport TA. A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector. *Eur J Biochem* (1974), **42**(1), 97–105.

[21] Schuster S and Höfer T. Determining all extreme semi-positive conservation relations in chemical reaction systems: a test criterion for conservativity. *J. Chem. Soc., Faraday Trans.* (1991), **87**, 2561–2566.

[22] Pfeiffer T, Sánchez-Valdenebro I, Nuo JC, *et al.* Metatool: for studying metabolic networks. *Bioinformatics* (1999), **15**(3), 251–257.

[23] Poolman MG, Sebu C, Pidcock MK, *et al.* Modular decomposition of metabolic systems via null-space analysis. *J Theor Biol* (2007), **249**(4), 691–705.

[24] Schuster S and Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems* (1994), **2**(2), 165–182.

[25] Schuster S, Dandekar T, and Fell DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* (1999), **17**(2), 53–60.

[26] de Figueiredo LF, Schuster S, Kaleta C, *et al.* Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics* (2009), **25**(1), 152–158.

[27] Schilling CH, Letscher D, and Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology* (2000), **203**(3), 229–248.

[28] Schilling CH, Covert MW, Famili I, *et al.* Genome-scale metabolic model of Helicobacter pylori 26695. *J Bacteriol* (2002), **184**(16), 4582–4593.

[29] Jevremovic D, Trinh CT, Srienc F, *et al.* On algebraic properties of extreme pathways in metabolic networks. *J Comput Biol* (2010), **17**(2), 107–119.

[30] Klamt S and Gilles ED. Minimal cut sets in biochemical reaction networks. *Bioinformatics* (2004), **20**(2), 226–234.

[31] Klamt S. Generalized concept of minimal cut sets in biochemical networks. *Biosystems* (2006), **83**(2-3), 233–247.

[32] Klamt S and Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* (2002), **29**(1-2), 233–236.

[33] Schuster R and Schuster S. Refined algorithm and computer program for calculating all non-negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Comput Appl Biosci* (1993), **9**(1), 79–85.

[34] de Figueiredo LF, Podhorski A, Rubio A, *et al.* Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* (2009), **25**(23), 3158–3165.

[35] Gianchandani EP, Chavali AK, and Papin JA. The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med* (2010), **2**(3), 372–382.

[36] Ruppin E, Papin JA, de Figueiredo LF, *et al.* Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr Opin Biotechnol* (2010), **21**(4), 502–510.

[37] Dantzig GB, Orden A, and Wolfe PS. (1954). *Notes on Linear Programming: Part I: The Generalized Simplex Method for Minimizing a Linear Form Under Linear Inequality Restraints.* RAND Corporation, Santa Monica, CA, USA.

[38] Edwards JS and Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* (2000), **97**(10), 5528–5533.

[39] Feist AM and Palsson BO. The biomass objective function. *Curr Opin Microbiol* (2010), **13**(3), 344–349.

[40] Ramakrishna R, Edwards JS, McCulloch A, *et al.* Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *Am J Physiol Regul Integr Comp Physiol* (2001), **280**(3), R695–R704.

[41] Dauner M and Sauer U. Stoichiometric growth model for riboflavin-producing Bacillus subtilis. *Biotechnol Bioeng* (2001), **76**(2), 132–143.

[42] Ebenhöh O and Heinrich R. Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bull Math Biol* (2001), **63**(1), 21–55.

[43] Poolman MG, Miguet L, Sweetlove LJ, *et al.* A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol* (2009), **151**(3), 1570–1581.

[44] Blank LM, Kuepfer L, and Sauer U. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol* (2005), **6**(6), R49.

[45] Williams TCR, Poolman MG, Howden AJM, *et al.* A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiol* (2010), **154**(1), 311–323.

[46] Grötschel M, Lovász L, and Schrijver A. (1993). *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, Heidelberg, Germany, second corrected edition.

[47] Lee S, Phalakornkule C, Domach MM, *et al.* Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Computers & Chemical Engineering* (2000), **24**(2-7), 711–716.

[48] Murabito E, Simeonidis E, Smallbone K, *et al.* Capturing the essence of a metabolic network: A flux balance analysis approach. *Journal of Theoretical Biology* (2009), **260**(3), 445–452.

[49] Kruse K and Ebenhöh O. Comparing flux balance analysis to network expansion: producibility, sustainability and the scope of compounds. *Genome Inform* (2008), **20**, 91–101.

[50] Benyamini T, Folger O, Ruppin E, *et al.* Flux balance analysis accounting for metabolite dilution. *Genome Biol* (2010), **11**(4), R43.

[51] Segrè D, Vitkup D, and Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America* (2002), **99**(23), 15112–15117.

[52] Shlomi T, Berkman O, and Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* (2005), **102**(21), 7695–7700.

[53] Burgard AP, Pharkya P, and Maranas CD. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* (2003), **84**(6), 647–657.

[54] Tepper N and Shlomi T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* (2010), **26**(4), 536–543.

[55] Beard DA, dan Liang S, and Qian H. Energy balance for analysis of complex metabolic networks. *Biophys J* (2002), **83**(1), 79–86.

[56] Henry CS, Broadbelt LJ, and Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. *Biophys J* (2007), **92**(5), 1792–1805.

[57] Hoppe A, Hoffmann S, and Holzhütter HG. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol* (2007), **1**, 23.

[58] Fleming RMT, Thiele I, Provan G, *et al.* Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *J Theor Biol* (2010), **264**(3), 683–692.

[59] Yizhak K, Benyamini T, Liebermeister W, *et al.* Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* (2010), **26**(12), i255–i260.

[60] Jerby L, Shlomi T, and Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* (2010), **6**, 401.

[61] Feist AM and Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. *Nat Biotechnol* (2008), **26**(6), 659–667.

[62] Fell DA, Poolman MG, and Gevorgyan A. Building and analysing genome-scale metabolic models. *Biochem Soc Trans* (2010), **38**(5), 1197–1201.

[63] Liolios K, Chen IMA, Mavromatis K, *et al.* The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* (2010), **38**(Database issue), D346–D354.

[64] Ebenhöh O, Handorf T, and Heinrich R. Structural analysis of expanding metabolic networks. *Genome Inform* (2004), **15**(1), 35–45.

[65] Handorf T, Ebenhöh O, and Heinrich R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol* (2005), **61**(4), 498–512.

[66] Kanehisa M, Goto S, Furumichi M, *et al.* KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* (2010), **38**(Database issue), D355–D360.

[67] Caspi R, Altman T, Dale JM, *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* (2010), **38**(Database issue), D473–D479.

[68] Handorf T and Ebenhöh O. MetaPath Online: a web server implementation of the network expansion algorithm. *Nucleic Acids Res* (2007), **35**(Web Server issue), W613–W618.

[69] Ebenhöh O and Handorf T. Functional classification of genome-scale metabolic networks. *EURASIP J Bioinform Syst Biol* (2009), page 570456.

[70] Handorf T, Christian N, Ebenhöh O, *et al.* An environmental perspective on metabolism. *J Theor Biol* (2008), **252**(3), 530–537.

[71] Christian N, May P, Kempa S, *et al.* An integrative approach towards completing genome-scale metabolic networks. *Mol Biosyst* (2009), **5**(12), 1889–1903.

[72] May P, Wienkoop S, Kempa S, *et al.* Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics* (2008), **179**(1), 157–166.

[73] Reed JL, Patel TR, Chen KH, *et al.* Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* (2006), **103**(46), 17480–17484.

[74] Raymond J and Segrè D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* (2006), **311**(5768), 1764–1767.

[75] Schütte M, Skupin A, Segrè D, *et al.* Modeling the complex dynamics of enzyme-pathway coevolution. *Chaos* (2010), **20**(4), 045115.

[76] Schrödinger E (1944). *What is life?* Cambridge University Press, Cambridge, UK.

[77] Hucka M, Finney A, Sauro HM, *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* (2003), **19**(4), 524–531.

[78] Novère NL, Finney A, Hucka M, *et al.* Minimum information requested in the annotation of biochemical models (miriam). *Nat Biotechnol* (2005), **23**(12), 1509–1515.

[79] Baldazzi V, Ropers D, Markowicz Y, *et al.* The carbon assimilation network in escherichia coli is densely connected and largely sign-determined by directions of metabolic fluxes. *PLoS Comput Biol* (2010), **6**(6), e1000812.

[80] Samal A and Jain S. The regulatory network of E. coli metabolism as a boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst Biol* (2008), **2**, 21.

[81] Chandrasekaran S and Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* (2010), **107**(41), 17845–17850.