# Chapter 2
# Methods and Data

This chapter describes the basics of the fundamental techniques used in this thesis. It is divided in three parts: (1) complex network tools applied to metabolism, (2) description of Flux Balance Analysis (FBA) -used to compute metabolic fluxes at steady state- and of Flux Variability Analysis -a variant of FBA to bound minimum and maximum fluxes for each reaction- and (3) a description of all the genome-scale metabolic reconstructions analysed in this thesis.

Nowadays, the explosion in computational power has allowed us to deal with systems of thousands or even millions of constituents and interactions, boosting the degree of our understanding on how these systems are structured and behave. Complex Network Science comprises a large amount of techniques and models which help us to study these intricate systems as a whole [1, 2]. These methodologies can be applied to any system which can be modelled as a network. Networks can be briefly defined as a set of items that interact, like for example the World Wide Web and the Internet and, in a biological context, metabolic networks [3] or protein–protein interaction networks [4]. Complex Network Science has led to an important advance in the understanding of metabolic networks [3, 5–9] which is in line with the systems-level view of metabolism in fields like Systems Biology [10, 11].

When dealing with metabolic networks, the complex network approach has to be combined with other techniques coming from Systems Biology in order to understand functional or behavioural features, for example why the inability to operate of some reactions leads to cell death, or why some reactions carry a determinate flux given a set of external nutrients. The most widespread mathematical approach used for the systems-level analysis of metabolic networks is Flux Balance Analysis (FBA) [12]. This technique is based on constraint-based analysis and optimization of an objective function, usually the biomass formation function of the cell. In this way, the fluxes through all the biochemical reactions of cell metabolism that maximize the biomass formation rate or, equivalently, the specific growth rate, can be computed. Apart from the mentioned reaction fluxes and growth rate, this technique allows to compute, for instance, the maximum yield of important compounds such as ATP or NADH [12, 13], and the effects of knockouts of genes or reactions [14, 15]. Related to FBA, other related techniques like Flux Variability Analysis (FVA) [16, 17] allow to

© Springer International Publishing AG 2017

O. Güell, *A Network-Based Approach to Cell Metabolism*, Springer Theses,
DOI 10.1007/978-3-319-64000-6_2

identify possible alternate solutions and, in conjunction with FBA, allow to perform a deep study of the flux capabilities of metabolic networks.

Complex network methodologies and constraint-based techniques applied to metabolic reconstructions represent a powerful tool for the analysis and development of new insights into metabolic functions and mechanisms that cells have developed from the earliest stages of life to the current days.

## 2.1 Structural Properties of Metabolic Networks as Complex Networks

Networks are discrete systems of elements that interact. These systems are represented by graphs of *nodes* (or vertices)—which represent elements—connected by *links* (or edges)—which represent interactions. The presence of a large number of nodes interacting in non-trivial connectivity patterns between order and disorder is what gives to networks their intrinsic complexity.

It is important to distinguish between complex and complicated. The main difference between these two words is better explained by a single example: solving a whole metabolic network composed of thousands of reactions is a complex problem in the sense that the large amount of interactions leads to emerging unexpected behaviours, like the effect of the removal of a biochemical reaction on other reactions, which can increase or decrease their fluxes depending on their biological activity. On the contrary, the study of a typical chemical engineering process to obtain a precise output may be a complicated problem, since one needs to draw a flowchart of all the chemical reactions and involved intermediate species that participate in the chemical synthesis. This may require a wide knowledge of the system, implying a large degree of control on all the processes, but the final behaviour of the system will be what is expected in a well-designed process.

### *2.1.1 Basic Representation Frameworks*

Links in networks can have either a defined direction or may lack it. Therefore, when links are directed, they are depicted by arrows, specifying a source and a target. A directed link can represent, for example, a transformation between two metabolites, typically a reactant and a product with the link pointing to the product. When no specification source/target is prescribed, the interaction is mutual, like in a protein–protein interaction,[1] and links without direction are used. Associated to this, networks are classified as *directed*, *undirected*, or *semidirected*. It is worth mentioning that links can also be *bidirectional*, meaning that the interaction allows either the

[1] Protein–protein interactions refer to physical contacts established between two or more proteins as a result of biochemical events and/or electrostatic forces.

forward or backward direction at the same time and this interconnection is thus reciprocal. This is specially important in the context of metabolic networks, where reactions can be either reversible (bidirected links, meaning that both directions of the reaction are possible) or irreversible (directed links, meaning that only one direction is thermodynamically favoured). Moreover, a link can carry a weight, representing the intensity of the interaction. Therefore, networks can be *weighted* or *unweighted*. In the case of metabolic networks, weights usually correspond to fluxes of the biochemical reactions. Metabolic networks typically display a probability distribution of fluxes (or weights) that follows a power law, meaning that fluxes spanning different orders of magnitude coexist in the same metabolic state [18].

Mathematically, unweighted undirected networks are described by the adjacency matrix, a square symmetric matrix $\{a_{ij}\}$ of binary values with an entry of 1 whenever there is a link between nodes $i$ an $j$ and 0 otherwise. In directed networks, the matrix is instead non-symmetric. Weighted networks are encoded by the weighted adjacency matrix $\{\omega_{ij}\}$, in which the values correspond to the weight of the edge between nodes $i$ and $j$.

Furthermore, networks can have different classes of nodes, leading to the so-called multipartite graphs. In multipartite graphs, links happen only between nodes in different categories. Networks with one kind of node are called *unipartite*, whereas networks with two kinds of nodes are called *bipartite* [19]. An important thing to notice is that bipartite networks can be projected into unipartite networks by performing a *one-mode projection*. To do this, one chooses a particular type of node and, in the projected reduction, places a link between two such nodes if there is at least one node of the complementary type connected to both of them.

In the real world, one can find networks combining all the mentioned properties (see Fig. 2.1). Metabolic networks are usually represented as bipartite semidirected networks, with metabolites and reactions belonging to different node categories with no direct connections between any two metabolites or any two reactions [21, 22] (see Fig. 2.1d). Although a bipartite representation is more accurate, it is sometimes preferable and always simpler to work with one-mode projections based on metabolites, which can be either directed or undirected (see Fig. 2.1a, b) depending on the reversibility of reactions, and weighted or unweighted depending on whether fluxes are taken into account. In such a projection, two metabolites get directly connected if there is at last one reaction in which they both participate (see Fig. 2.1f).

### 2.1.2 Degree Distribution

Nodes in networks are locally characterized by the number of their surrounding neighbours. This magnitude is called the *degree* of a node $k$ (see Fig. 2.2). The probability of nodes having a certain degree $k$ is written $P(k)$ and named *degree distribution*, and can be computed from the fraction of nodes in the network that has degree $k$.

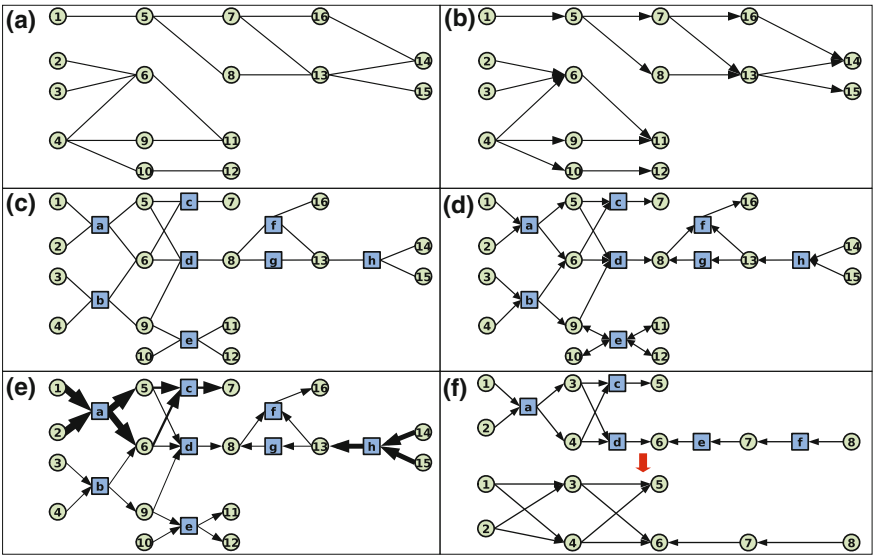



**Fig. 2.1** Examples of different types of networks. **a** Undirected unipartite. **b** Directed unipartite. **c** Undirected bipartite. **d** Semidirected bipartite network. Notice that connections involving node *e* are bidirectional. **e** Semidirected weighted bipartite. The thickness of the links is proportional to their weight. **f** Example of the transformation into a one-mode projected network of metabolites from a semidirected bipartite metabolic network containing metabolites and reactions. Metabolites are represented by *circles* and reactions by *squares*. Parts of this figure have been extracted from Ref. [20] Copyright @ 2014, World Scientific Publishing

Usually, real world networks show degree distributions $P(k)$ that are highly skewed with long tails that reach values far above the mean [23]. In most cases, degree distributions follow a power-law, $P(k) \propto k^{-\gamma}$, where $\gamma$ is the characteristic exponent and it has values in the range $2 < \gamma < 3$. Networks with a degree distribution described by a power-law are called *scale-free*.[2] Networks with power-law degree distributions have attracted much attention and have been studied intensively [24–26]. Notice that, usually, it is useful to work with the complementary cumulative probability distribution function $P(k' \geq k)$ in order to avoid noise effects present for large values of $k$.

In semidirected networks, the degrees of nodes are defined in relation to incoming ($k_{in}$), outgoing ($k_{out}$) and bidirectional ($k_b$) links. Correspondingly, nodes have a total degree expressed as a sum of contributions $k = k_{in} + k_{out} + k_b$. These degrees can present local correlations and so the degrees of nodes are described by the joint probability $P(k_{in}, k_{out}, k_b)$. In addition, for bipartite networks, nodes of each kind have also their own degree distribution.

[2]This name is refers to the scale-invariance that power-laws display: if $f(x) = a(x)^{\gamma}$, then $f(cx) = a(cx)^{\gamma} = c^{\gamma} f(x)$.

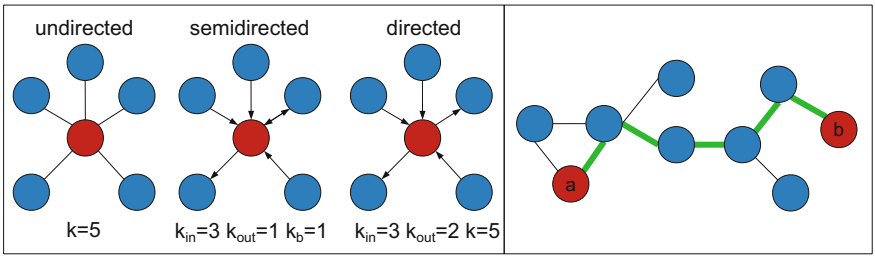



**Fig. 2.2** Schematic example of a degree of a node (*left*) and a path between two nodes (*right*). *Left* Example of the degrees in a undirected, semidirected, and directed networks. *Right* Path between node $a$ and $b$, highlighted in *green*. In this case, the shortest path length between nodes $a$ and $b$ is $\ell_{ab} = 5$ (color figure online)

Regarding specifically metabolic networks, the total degree of metabolites $k_M$ in bipartite representations follows a power-law degree distribution $P(k_M) \propto k_M^{-\gamma}$ [1, 7]. In Ref. [3] it is found that in the organism *Escherichia coli*, the probability $P(k_{in})$ that a metabolite participates as a product and the probability $P(k_{out})$ that a metabolite participates as a reactant have both a value of $\gamma$ of 2.2. Similarly, Ref. [27] shows that for the organism *Helicobacter pylori*, the exponent has a value of 2.32. The fact that metabolites display a scale-free degree distribution means that there is a high diversity in the number of reactions in which metabolites participate. The largest part of metabolites have a few connections, whereas a few metabolites, generically called hubs, have many of them. Examples of these highly-connected metabolites are ATP, $H_2O$, or $H^+$, which can participate in up to 50% of the total number of reactions for the case of $H^+$ in the organism *E. coli* [28]. On the contrary, reactions show a peaked distribution of total degree, the peak being located at an average degree $< k_R > \sim 4$. The bounded form of the distribution arises from the fact that reactions have a limited number of participants, typically from 2 to 12.

In Fig. 2.3a, the bipartite cumulative probability distribution function $P(k'_M \geq k_M)$ of metabolites and the bipartite probability distribution function $P(k_R)$ for reactions of the three organisms analysed in this thesis, *E. coli* [29–31], *Mycoplasma pneumoniae* [32, 33], and *Staphylococcus aureus* [34] are shown. Clearly, metabolites show a power-law degree distribution and reactions a peaked distribution, as mentioned above. In fact, all networks studied here have similar tendencies for both distribution functions, showing that metabolic networks, in spite of corresponding to quite different microorganisms, display often universal properties [3].

### 2.1.3 Average Path Length

Another common feature of complex networks, and in particular of metabolic networks, is the fact that any two nodes are connected by *paths* of links that are typically very short in the number of intermediate steps [7]. This is called the *small-world*

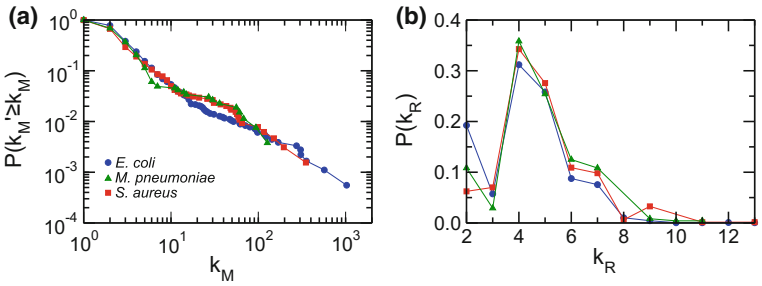



**Fig. 2.3** Features of the networks of *Escherichia coli* *i*JO1366 (see Sect. 2.3.1), *Mycoplasma pneumoniae* *i*JW145 (see Sect. 2.3.2), and *Staphylococcus aureus* *i*SB619 (see Sect. 2.3.3). **a** Complementary cumulative probability distribution function of metabolites. **b** Degree distribution of reactions. Extracted from Ref. [35]

property. In technical terms, the distance $\ell$ between two nodes is defined as the number of jumps or hops along the shortest path that connects them (see Fig. 2.2). Hence, it is possible to define the average shortest path length $< \ell >$, which is the average of all the shortest distances between pairs of nodes. The small-world property is stated in the fact that $< \ell >$ increases as the logarithm of the network size $N$ (number of nodes) [23, 25].

Small average path lengths indicate that the network contains highly-connected nodes that act as shortcuts, reducing the average number of steps needed to go from one node to another. This is crucial in many real contexts, and in particular for cell metabolism. In Ref. [3], the authors measured the average path length for 43 organisms and found a similar value for all of them, $< \ell > \sim 3.2$. This value was explained by the role of hubs, which decrease dramatically the number of steps needed to travel from one node to another. When hubs are not taken into account, longer and variable path lengths are obtained [36, 37], depending on the biological domain where organisms belong to. Typical values are 9.57, 8.50, and 7.22 for eukaryotes, archaea, and bacteria, respectively, with the differences due to evolutionary processes. Nevertheless, there remains some controversy about the small-world property in metabolic networks. In Ref. [38], it is stated that usually paths are computed by directly linking metabolites through reactions and that this is not adequate, since pathways computed in this way do not conserve their structural moieties[3] and thus they do not correspond to pathways on a traditional metabolic map. Therefore, in Ref. [38] metabolites are linked depending on the conserved structural moieties in the adjacent reactions and, as a result, it is stated that the average path length of *E. coli* metabolism is longer than it was previously thought and, consequently, the

[3] According to the IUPAC, a moiety is a part of a molecule that may include either whole functional groups or parts of functional groups as substructures.

*E. coli* metabolic network is not small in terms of biosynthesis and degradation of metabolites. However, it is generally accepted that metabolic networks show indeed the small-world property at the structural level. In this thesis, path lengths will be computed in Chap. 4.

### *2.1.4 Communities at the Mesoscale*

It is thought that biological networks are composed by subsets of nodes that are functionally separable called *modules* [25, 39]. In general, this idea corresponds to the concept of communities in networks. The organization of a network into communities does not imply fragmentation. Instead, communities are subsets of a network which contain a dense interconnection pattern between nodes inside the community and lower interconnection levels with nodes outside. This can be related with the presence of a large clustering (see Sect. 2.1.6) between nodes inside the community.

Community detection [40] represents an active field in Complex Network Science motivated by the potential identification of communities with functional or operational units. Several methods, based on different exploratory techniques, have been proposed. Among the most successful community detection methods one finds, for instance, algorithms that use random walkers to partition the network into communities, like Infomap [41]. Other methods are based on the optimization of modularity. Modularity is a measure of the quality of a community structure [42]. It measures the internal connectivity of identified communities with reference to a randomized null model with the same degree distribution. Algorithms based on modularity optimization try to find the best community structure in terms of the modularity measure. Examples of successful algorithms based on this measure are SpinGlass [43] or Louvain [44]. On what follows, the three methods used in Chap. 3 of this thesis to detect communities are explained.

- *Distance hierarchical clustering*: this method starts by defining a distance between pairs of nodes in the network. Then, once the pairs of nodes have a defined distance, one groups similar nodes into communities according to this distance. There are different schemes based on distances to group nodes intro communities. The two simplest methods are single-linkage clustering, in which two sets of nodes are considered separate communities if and only if all pairs of nodes in the different sets have distance larger than a given threshold, and complete linkage clustering, in which all nodes of a community have a distance smaller than a threshold [45] (see Fig. 2.4).
- *Infomap algorithm* [41]: the main idea of this algorithm is that a random walker will tend to flow at different paces within a network, spending more time inside communities and less time to pass between them (see Fig. 2.4). The way in which the random walker moves around communities can be compared to the flow of messages between individuals. In this way, there is a strong current of messages between

individuals inside a community, and a weaker current of messages between individuals of different communities.

- *Recursive percolation*: this method has been developed in a work related to this thesis [35]. Recursive percolation identifies components in which the network is fragmented just below the percolation threshold (see Sect. 2.1.5), where the connected network disaggregates into smaller components. To find them, links are removed sequentially from lower to higher weights until the percolation transition is detected. Then, clusters are identified using a burning algorithm [46]. This procedure is applied to each component until the distribution of sizes of the obtained communities reaches some thresholds, for instance, to be similar to those given by the distance hierarchical clustering technique and Infomap. A schematic example of this process is shown in Fig. 2.4.

### 2.1.5 Large-Scale Connected Components

Global connectedness is one of the most fundamental properties of complex systems. The theory that describes the behaviour of network connected components is *percolation theory*. Briefly, percolation theory states that there exists a critical point, called *percolation threshold* and denoted as $p_c$, where a transition in the global connectedness of the network occurs, from a state where the network is formed by small isolated components to the emergence of a *giant connected component* (GCC) spanning a macroscopic fraction of the network. This means that it is always possible to find a path connecting every pairs of nodes inside the GCC.

This concept can be extended to networks with directed links. The connectivity of directed networks presents special features since the path between two nodes $i$ and $j$ can be different when going from $i$ to $j$ or vice versa. This fact leads to the existence of a bow-tie structure inside the GCC [22, 47, 48]. The main feature of the bow-tie structure of a GCC in a directed network is that one can detect the presence of a *strongly connected component* (SCC), which is a region of the network where any node is reachable from any other by a directed path. It can happen that directed networks contain more than one SCC.

Apart from the SCC, one of the other significant regions that can be found in the bow-tie structure of directed networks is called *IN component*, with nodes that can reach the SCC but that cannot be reached from the SCC. Analogously, the *OUT component* contains nodes that can be reached from the SCC but that cannot return to it. *Tubes* are sequences of nodes that connect the IN with the OUT component without going through the SCC. Finally, *tendrils* are composed by nodes that have no access to the SCC and that are not reachable from it, similarly to tubes. They go out from the IN component and come in from the OUT component. A visual scheme of the bow-tie structure of directed networks is shown in Fig. 2.5a. The bow-tie structure of *E. coli* and *Mycoplasma pneumoniae* will be explicitly considered in Chap. 5.

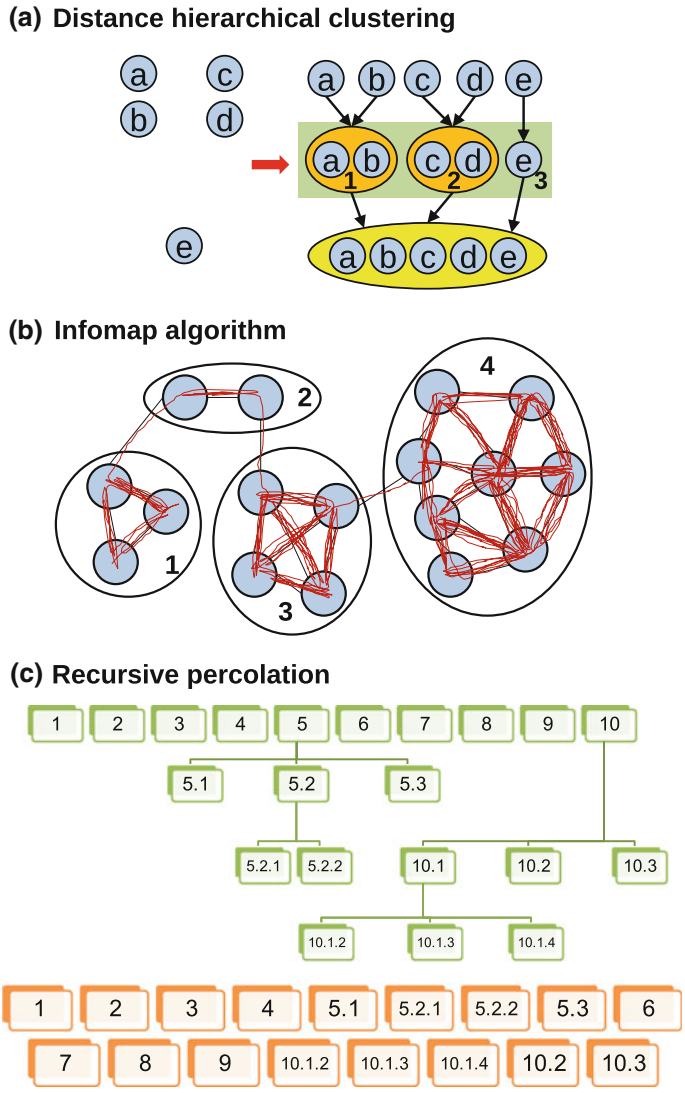



**Fig. 2.4** Examples of the clustering methods. **a** Example of the distance hierarchical clustering method. Modules are formed by nodes that are nearer. Notice that with this method it is necessary to apply a threshold depending on the distances. In this example, the threshold is represented by the *green rectangle*. At this level, three communities are detected. **b** Example of the Infomap algorithm. Clusters are found with a random walker. Communities are found depending on the frequency of times that each random walker visits a set of nodes. **c** Example of the application of Recursive percolation. The first step leads to 10 clusters. Among these 10 clusters, the largest are fragmented, leading to more clusters. This partition is iterated until the distribution of sizes is similar to that in other methods. Parts of this figure have been extracted from Ref. [35] (color figure online)

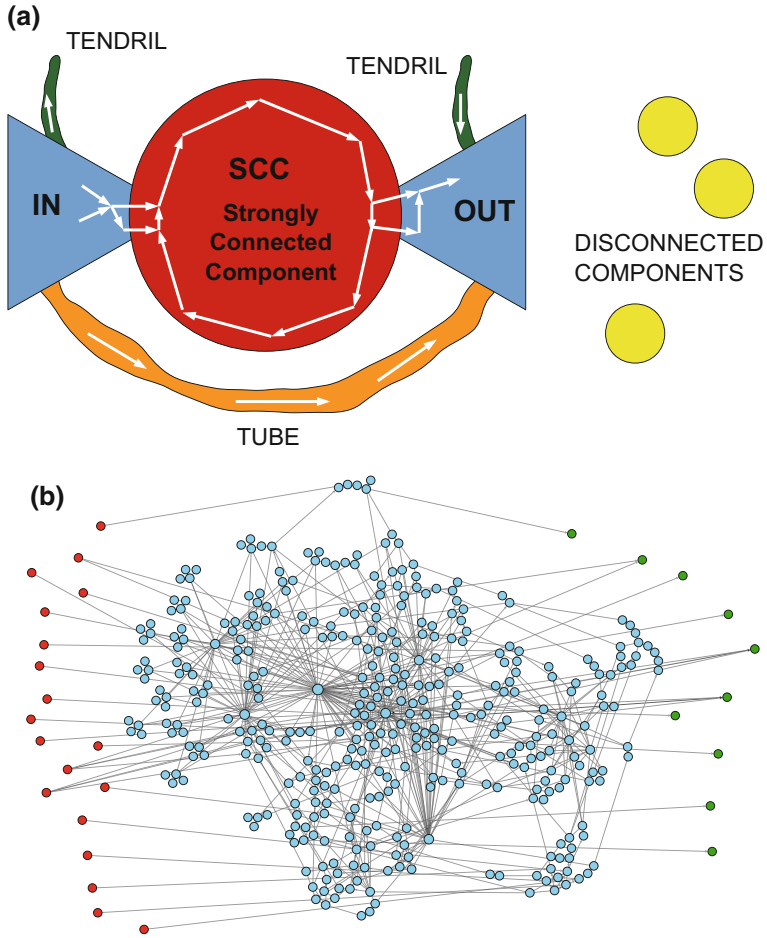



**Fig. 2.5** Examples of connected components. **a** Schematic example of a bow-tie structure. **b** Example of the bow-tie structure of *Mycoplasma pneumoniae* [32], an organism studied in this thesis. *Blue nodes* compose the SCC, *red nodes* compose the IN component, and *green nodes* compose the OUT component (color figure online)

Metabolic networks show a bow-tie structure typically with a large SCC connected to non-structured IN and OUT components (see Fig. 2.5b) [47, 49]. The SCC contains the largest part of metabolites and reactions composing the network, representing thus the entire metabolic machinery of cells. IN and OUT components are formed of, respectively, nutrients and waste products directly connected to the SCC (see Fig. 2.5b).

### 2.1.6 Other Structural Properties of Complex Networks

Real networks exhibit also the presence of non trivial correlations in their connectivity. At the level of two nodes, it is convenient to characterize degree correlations with the average nearest neighbour degree $\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k)$, where $P(k'|k)$ is the probability of having a node with degree $k'$ given that it is connected to a node with degree $k$. It basically considers the mean degree of the neighbours of a node as a function of its degree $k$. If $\bar{k}_{nn}(k)$ increases with $k$, it is said that the network is *assortative*, with nodes that connect preferentially to other nodes of similar degree. If $\bar{k}_{nn}(k)$ decreases with $k$, the network is named *disassortative*, with high-degree nodes attached preferentially to nodes with low degrees. Biological networks, and in particular metabolic networks, usually show a disassortative pattern [7].

Correlations among three nodes can be measured by means of the concept of *clustering*, which refers to the tendency to form triangles between the neighbours of a vertex. Watts and Strogatz [50] proposed a measure known as *clustering coefficient*, $c_i = \frac{2E_i}{k_i(k_i-1)}$, where $E_i$ is the number of edges that exist between neighbours of the node $i$ and $k_i$ denotes the degree of the node $i$. Although this measure is helpful as a first indication for clustering, it is more informative to work with quantities which depend explicitly on the degree $k$. Therefore, a degree-dependent clustering coefficient $\bar{c}(k)$ is calculated as the clustering coefficient of nodes averaged for each degree class $k$. Metabolic networks tend to display high levels of clustering [5, 25] with $\bar{c}(k)$ having a decreasing dependence on $k$ [6].

A final mention is deserved to structures called *motifs* [11]. Motifs are small subsets of connected nodes that are found in networks more often than expected at random. They are considered as elementary functional units, and each real network has its own set of distinct motifs. Their identification provides useful insights into the typical local connectivity patterns in the network.

### 2.1.7 Null Model Networks and Randomization Methods

Null models in Complex Network Science serve to study fundamental properties of complex networks and to asses the statistical significance of a property, first measuring it in the real network and then comparing the original results to the ones obtained in the randomized versions. These models can be used to prove the existence of graphs satisfying various properties, or to provide a rigorous definition of what it means for a property to hold for almost all graphs or, finally, to act as a benchmark for specific features of real networks.

One of the most known models was the graph structure proposed by Paul Erdös and Alfréd Rényi. The *Erdös–Rényi* model [51, 52] consists on generating realizations of random networks given the total number of nodes $N$ and a total number of links $L$, and connecting every pair of them with probability $p$. This leads to a binomial degree distribution, that can be approximated by a Poisson distribution for realizations with a large number of nodes.

Another important method to construct random networks is the *Configuration model*, an algorithm to construct random networks with a degree sequence or degree distribution $P(k)$ settled a priori [53, 54]. The total number of nodes $N$ remains constant. For each node, a random number $k$ is drawn from the probability distribution $P(k)$ and it is assigned to the node in the form of half-edges. The network is then constructed by connecting pairs of these link ends chosen uniformly at random. These realizations, like the Erdös–Rényi networks described above, are uncorrelated and have no clusters in the thermodynamic limit $N \to \infty$.

Instead of comparing real networks with null models as those described above, it is sometimes preferable to randomize a network obtained from real data by rewiring, i.e., by picking two links at random and swapping their end [55]. While randomizing, one can preserve different properties, for instance the degrees of all nodes. Two rewiring randomization methods have been used in this thesis, one that preserves the degrees of all nodes—similar to comparing with the Configuration model—called *degree-preserving* randomization, and another that generates randomized versions taking into account that new reactions must be stoichiometrically balanced, called *mass-balanced* randomization.

#### 2.1.7.1 Degree-Preserving Randomization

In metabolic networks, the degree-preserving randomization method is similar to the Configuration model in bipartite networks. Degree-preserving randomization works by choosing two pairs of connected nodes (metabolites and reactions) of the bipartite network at random and swapping their ends, unless this would lead to a repeated metabolite in a reaction (see Fig. 2.6, left). The steps of the algorithm are:

1. Pick two links at random: $m_1 \to r_1$ and $m_2 \to r_2$ or $r_1 \to m_1$ and $r_2 \to m_2$, where $m$ are metabolites and $r$ reactions.
2. Swap the end of the links avoiding repeated links and self-production: ($m_1 \to r_2$ and $m_2 \to r_1$ or $r_1 \to m_2$ and $r_2 \to m_1$).
3. Repeat until $L^2$ swappings are performed, where $L$ is the total number of links in the network.
4. Make several realizations of the randomized metabolic network following the three previous steps.

Reversible reactions are rewired independently of the irreversible ones in order to preserve the degrees of metabolites which correspond to reversible and irreversible reactions. This method gives networks which preserve the degrees of metabolites and reactions and it is useful, for instance, to determine the role of the degree distribution in large failure cascades in bacterial organisms, which may have evolved towards reducing the probability of having large cascades that produce metabolic damage, increasing thus robustness [56]. This method will be used in Chap. 3.

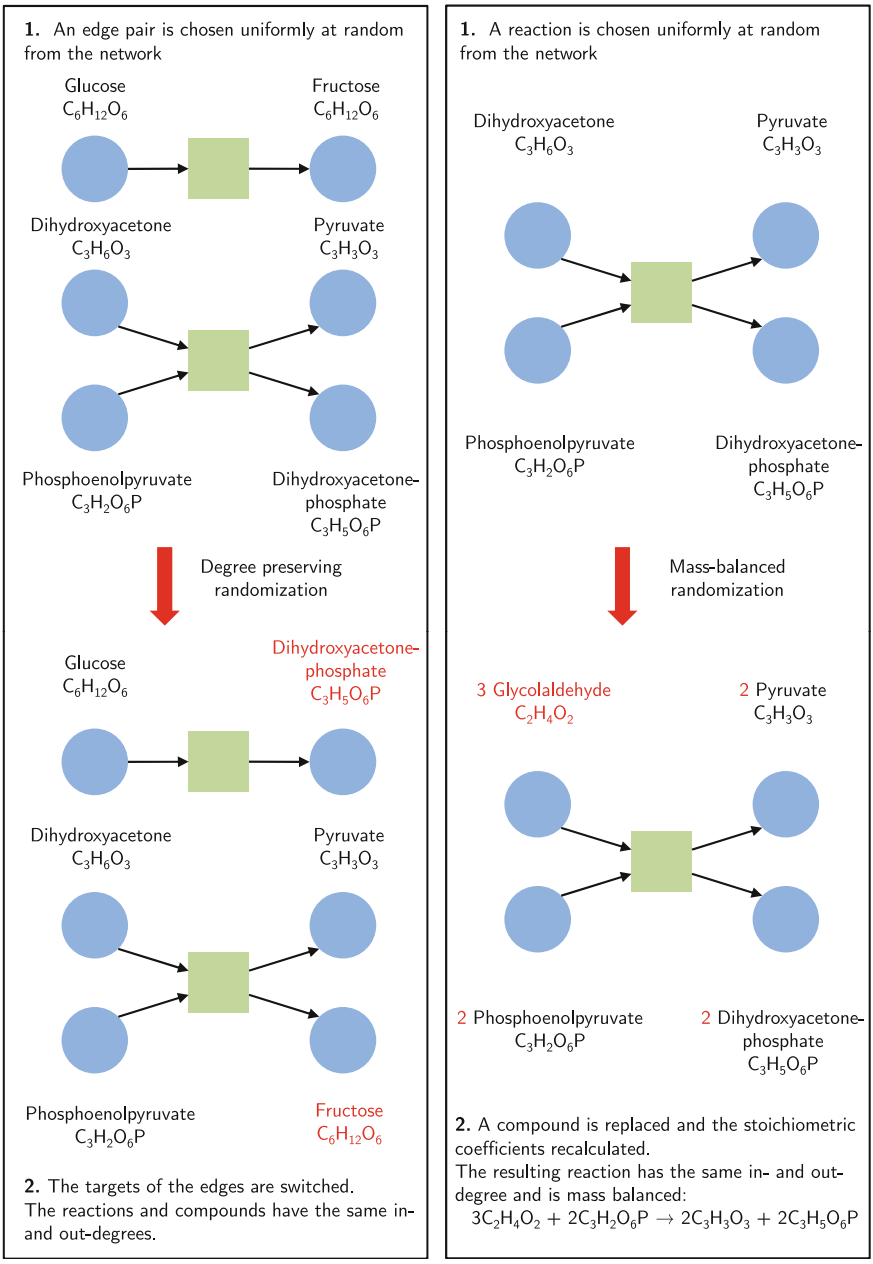



**Fig. 2.6** *Left* Scheme of the degree-preserving randomization algorithm. IN and OUT degrees are conserved, but mass balance is not satisfied. *Right* Scheme of the mass-balanced randomization. In this case metabolites are switched only if the new reaction is mass balanced; while reaction degrees are kept constant, the degrees of metabolites are not preserved. Extracted from Ref. [57] Copyright @ 2012, PACIS-JCIS

#### 2.1.7.2 Mass-Balanced Randomization

Mass-balanced randomization generates randomized networks by rewiring the links corresponding to substrate-reaction or product-reaction relationships, while preserving atomic mass balance of the reactions [58]. Given a reaction $r$, its atomic mass balance is given by:

$$\sum_{e \in E_r} s_{e,r} \cdot m_e = \sum_{p \in P_r} s_{p,r} \cdot m_p \tag{2.1}$$

where $E_r$ denotes the set of substrates and $P_r$ the set of products in $r$, and $m_e, m_p$ are the mass vectors ($m_{H_2O} = (0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$ as an example) of $e$ and $p$, respectively. Finally, $s_{e,r}$, $s_{p,r}$ are their stoichiometric coefficients. For instance, consider the reaction $A \rightarrow B$, with $m_A = m_B = C_6H_{12}O_6$. Then, $A$ may be substituted by a compound $C$ with $m_C = C_3H_6O_3$ from within the network, resulting in the randomized reaction $2\ C \rightarrow B$, which satisfies Eq. 2.1 since $2\ C_3H_6O_3 = C_6H_{12}O_6$ (see Fig. 2.6, right). In addition to substituting individual substrates or products, the method also allows more complex substitutions involving pairs of substrates or products, yielding a large number of possible substitutions.

The motivation for preserving atomic mass balance of reactions, a fundamental physico-chemical constraint, is that the resulting null model allows estimating the importance of network properties with respect to evolutionary pressure. As biological systems and their properties evolve under physical constraints and evolutionary pressure, a null model which satisfies physical principles but does not account for evolutionary pressure differs from a metabolic network only in the properties which are affected by evolutionary pressure. Thus, a property deemed statistically significant following mass-balanced randomization is beyond basic physical constraints and likely to be a result of evolutionary pressure [59]. The method preserves mass balance and reaction degrees but not the degrees of metabolites, since the stoichiometric coefficients and metabolite degrees are changed. This method will be used in Chap. 3.

## 2.2 Flux Balance Analysis

A general aim of the study of a metabolic network is to characterize and understand the configuration of fluxes of the reactions constituting the network in connection to phenotype and behaviour. The study of fluxes in metabolic networks deserves a special treatment more biochemically focused than in usual chemical kinetics schemes. With the knowledge of the kinetic constants of the reactions, it would be possible to solve the equations associated to the fluxes of reaction and the concentrations of metabolites in the metabolic network using proper mathematical methods. However, there is a lack in the availability of kinetic parameters [60] due to the difficulty in measuring them experimentally. As an alternative, computational techniques have been proposed in order to estimate fluxes through reactions of metabolic networks at steady-state.

*Flux Balance Analysis* is maybe the most successful and widely used approach to compute the fluxes through metabolic reactions of an organism. In addition, FBA also estimates its growth rate by maximizing the flux through the biomass reaction of the network. This technique will be used in Chaps. 4, 5, and 6.

To be more specific, metabolic reactions can be represented in terms of a *stoichiometric matrix*, this being the fundamental basis in FBA and other modelling approaches [12, 17, 61, 62]. To construct a stoichiometric matrix [63–65], one must first write the typical kinetic equations which describe the temporal variation of the concentration of metabolites, which are derived from the mass conservation principle,

$$\frac{dc_i}{dt} = \sum_{j=1}^{N_R} S_{i,j}\nu_j \tag{2.2}$$

The concentration of metabolite $i$ is denoted by $c_i$, $N_R$ is the total number of reactions, $S_{i,j}$ is the stoichiometric coefficient of metabolite $i$ in reaction $j$, and $\nu_j$ stands for the flux of reaction $j$. Note that, typically, reaction fluxes have units of mmol $\text{gDW}^{-1}$ $\text{h}^{-1}$, where gDW means grams Dry Weight. Notice that the values of the **S** matrix correspond to the stoichiometric coefficients of each metabolite in each reaction. Thus, each row represents a metabolite, whereas each column represents a reaction. Therefore, if a metabolite $i$ does not participate in a reaction $j$, its stoichiometric coefficient will be 0, $S_{ij} = 0$. Otherwise, if the metabolite is a reactant, the stoichiometric coefficient will be negative, $S_{ij} < 0$, and if it is a product, it will be positive, $S_{ij} > 0$ (see Fig. 2.7).

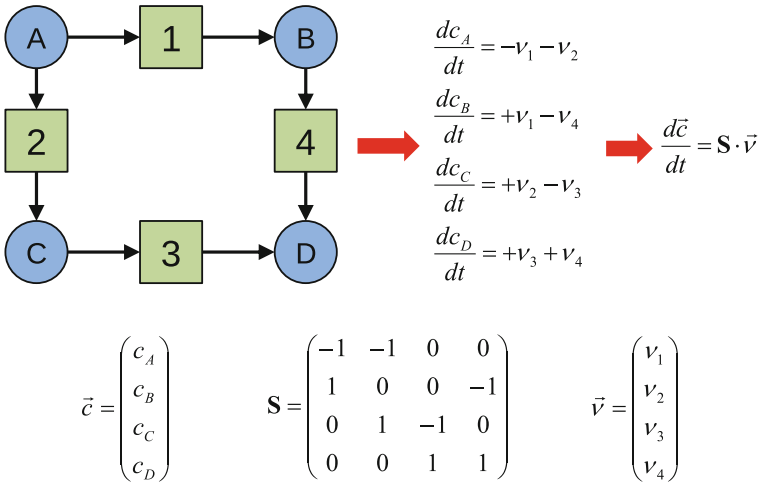



**Fig. 2.7** Equations derived from mass-balance associated to a simple metabolic network. Matrix **S** is the so-called stoichiometric matrix, $\vec{\nu}$ is a vector containing all the fluxes of the metabolic network, and $\vec{c}$ denotes the vector with concentrations of metabolites

Metabolic networks are open-systems, which implies that some metabolites can leave or enter the organism. Therefore, it is not possible to arrive to a thermodynamic equilibrium state. However, it is possible to attain a non-equilibrium steady state, where the concentrations of metabolites do not change with time, forcing the system to exchange metabolites with the environment. This steady-state condition simplifies the system of coupled differential Equation 2.2 derived from mass balance into an ordinary linear system of equations, which can be written as a product of the stoichiometric matrix $\mathbf{S}$ by the vector of fluxes $\vec{\nu}$,

$$\mathbf{S} \cdot \vec{\nu} = \vec{0} \tag{2.3}$$

This is the typical form of the equation to be solved by the FBA technique. As mentioned before, it is important to notice that no kinetic parameters [66, 67] appear explicitly in Eq. 2.3 and, thus, they are not needed in relation to FBA applications.

It is important to precise that, apart from the intrinsic constraints imposed by the steady-state condition, other bounds of the form $\alpha_i \leq \nu_i \leq \beta_i$ may be imposed on the values of the fluxes to render the whole scheme both chemically and biologically realistic. These upper and lower bounds may depend on the thermodynamics of reactions, more precisely on their reversibility. If reactions are reversible, fluxes can have positive or negative fluxes, whereas for the case of irreversible reactions, reactions must have only positive fluxes. Further, since the steady-state condition forces the system to exchange metabolites with the environment, constraints on *exchange* fluxes are imposed for metabolites that can either enter or leave the organism. These exchange fluxes are taken positive from the system to the environment. Notice that fluxes obtained using FBA will depend on the particular chosen external medium.

In metabolic networks, there are usually more reactions than metabolites. The system of Eq. 2.3 is thus underdetermined, i.e., there are multiple solutions even after imposing the mentioned constraints. Therefore, a biological objective function is introduced to restrict the solution space to a single biologically meaningful solution. Technically, this means that FBA selects the state in the solution space that maximizes the value of the objective function (see Fig. 2.8). This objective function depends on the biological information that one wants to extract, but usually one chooses to optimize biomass formation adjusted to be equivalent to maximize the specific growth rate of the organism. To do this, a *biomass reaction* is added to the network which simulates the biomass production. Other possible objective functions are ATP or NADH production or yield.

Often, other auxiliary reactions are needed apart from exchange and the biomass formation reactions. The first category includes physiological requirements, like the *ATP maintenance* reaction, which is a reaction which consumes ATP in order to simulate biological energetic costs for the organism which are not associated to growth. A second category are the so-called *sink* reactions, which are reactions that have not been identified yet and that consume some metabolites to avoid accumulation. A generic sink reaction has the simple form $A \rightarrow \emptyset$.

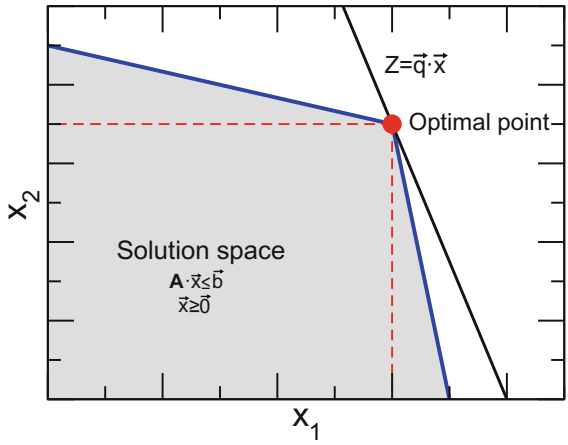



**Fig. 2.8** Example of the optimization of an objective function on a system of two variables

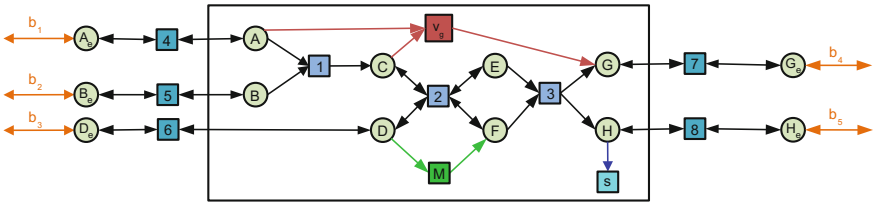



**Fig. 2.9** Example of a FBA calculation in a metabolic network. Reactions are denoted by *squares* and metabolites by *circles*. The biomass production reaction (*red square*) is labelled as $\nu_g$. Exchanges fluxes for interactions with the environment (*orange arrows*) are denoted with *b* labels. A sink reaction (*cyan square*) is shown with a *s* label. The ATP maintenance reaction (*green square*) is also shown denoted with a *M* label. Parts of this figure have been extracted from Ref. [20] Copyright @ 2014, World Scientific Publishing (color figure online)

In this way, a consistent system of equations representing the whole cell metabolism is obtained and one tries to find a solution that optimizes the value of an objective function (see Fig. 2.9 for a schematic picture of a FBA computation). If no solution exists for optimization of biomass production in a particular medium condition, one can assume that the system is not able to grow and therefore one can conclude that the organism is not able to survive in this medium.

The mathematical notation to denote a standard FBA problem choosing to optimize the specific growth rate is

$$\begin{aligned} &\textbf{maximize} && \nu_g \\ &\textbf{subject to} && \mathbf{S} \cdot \vec{\nu} = \vec{0} \\ &\textbf{and} && \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \end{aligned}$$

where $\vec{\alpha}$ and $\vec{\beta}$ represent the vectors determining the lower and upper bounds of the reactions, and $\nu_g$ denotes the specific growth rate.

The software used to perform these calculations in this thesis is GNU Linear Programming Kit (GLPK) [68–71], through its associated solver GLPSol. This solver uses a dual *simplex* algorithm to compute the solutions. It is a variant of the normal simplex algorithm [72]. The latter is an iterative algorithm which is based on finding first feasible solutions and then finding the most optimal solution based on these feasible solutions. On the contrary, dual simplex works by first finding optimal solutions and then finding a feasible solution, again, if it exists.

### 2.2.1 Formulation of the Biomass Reaction

FBA problems are usually solved by maximizing the flux through the biomass reaction [29, 31, 33]. This typically gives a particular flux state of the metabolic network compatible with the constraints. However, the solution obtained by FBA is often not unique. In some cases, the metabolic network is able to achieve the same specific growth rate by using alternate reactions and pathways. Therefore, phenotypically different solutions that optimize the specific growth rate are possible, implying that FBA solutions can be degenerate [12].

Technically, the biomass reactions is modelled as a reaction, $aA + bB + cC + dD... \longrightarrow xX + yY + zZ$, which produces and consumes some specific metabolites (see Fig. 2.9). These metabolites are known biosynthetic precursors present in the metabolic network under consideration. The key point is given by their stoichiometric coefficients in the biomass reaction, which are experimentally measured proportions in the biomass of the organism measured in dry weight conditions. The stoichiometric coefficients of the metabolites participating in the biomass reaction have units of mmol $\text{gDW}^{-1}$, and the biomass reaction has units of $h^{-1}$. It is worth stressing that this reaction simulates the growth of an organism given a set of external nutrients and that its coefficients are adjusted so that its flux is equivalent to the specific growth rate of the organism.

FBA can also maximize the biomass yield, which is the equivalent to maximize the specific growth rate but taking into account that the maximum uptake of the carbon source, for example glucose, must be set to 1 mmol $\text{gDW}^{-1}\ h^{-1}$ to set the maximum amount of biomass that can be produced per 1 mol of nutrient.

### 2.2.2 Simulation of Different Environments

It is important to make explicit the way to simulate changes in the environment using FBA. To do this, one must tune the upper and lower bounds of the values of the exchange reactions of the metabolites that are present in the environment. As an example, suppose that one wants to model that glucose is present in the environment and that, therefore, the organism consumes it in order to obtain energy. The explicit

form of the constraint of the exchange flux of glucose will be $-10 \leq \nu_{glucose}^{exchange} \leq \infty$, which means that the organism can expel as much as glucose as it wants but that it can eat glucose with a maximum uptake of 10 mmol $gDW^{-1}$ $h^{-1}$.

Notice that nutrients have a negative lower bound and an unlimited upper bound, whereas waste products have a value of the lower bound of 0 and unlimited upper bound, which means that the organism cannot uptake it but, if the compound is generated inside the organism, it can be expelled to the exterior as waste. As an example, this would be the case for $CO_2$ in *E. coli*, which is not eaten by the organism but that is expelled due to respiration.

To summarize, an environment is simulated by choosing a set of nutrients and assigning a lower bound $-\alpha_i$ to each nutrient, which is the maximum uptake of each nutrient, and assigning a lower bound of 0 to components not present in the environment. For all external metabolites, the upper bound is set to $\infty$. Therefore, for nutrients one has $-\alpha \leq \nu_{nutrient}^{exchange} \leq \infty$, whereas for waste products one has $0 \leq \nu_{waste}^{exchange} \leq \infty$. The rest of reactions are modelled as told in the previous section.

#### 2.2.2.1 Construction of Minimal Media

A minimal medium is the minimal set of metabolites which ensure the viability of an organism. The modelling of these media can be made as in Ref. [31]. Minimal media consist of a set of mineral salts, and one source of carbon, of nitrogen, of sulphur and of phosphorus, from four families representing carbon, nitrogen, phosphorus, and sulphur compounds, respectively. To construct different minimal media, the set of mineral salts is always the same—which contains, for example, magnesium sulphate, iron chloride, and calcium chloride-, but each source family is browsed while the other three sources are fixed to the standard metabolites of each kind (C*: glucose, N*: ammonia, P*: phosphate, S*: sulphate) (see Table 2.1).

**Table 2.1** Examples of the construction of minimal media. Asterisks denote the standard metabolite of each kind. To construct carbon media, the sources of nitrogen, phosphorus and sulphur are set to the standard components of each kind whereas the carbon sources are varied. The same procedure applies to construct nitrogen, phosphorus and sulphur media

| Variation of carbon sources | | | | | Variation of phosphorous sources | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Medium 1 | $C_1$ | N* | P* | S* | Medium 1 | C* | N* | $P_1$ | S* |
| Medium 2 | $C_2$ | N* | P* | S* | Medium 2 | C* | N* | $P_2$ | S* |
| Medium 3 | $C_3$ | N* | P* | S* | Medium 3 | C* | N* | $P_3$ | S* |
| Variation of nitrogen sources | | | | | Variation of sulphur sources | | | | |
| Medium 1 | C* | $N_1$ | P* | S* | Medium 1 | C* | N* | P* | $S_1$ |
| Medium 2 | C* | $N_2$ | P* | S* | Medium 2 | C* | N* | P* | $S_2$ |
| Medium 3 | C* | $N_3$ | P* | S* | Medium 3 | C* | N* | P* | $S_3$ |

#### 2.2.2.2 Construction of Rich Media

Sometimes it can be useful to perform FBA computations in a medium with more components that the ones present in a minimal medium. These media containing more nutrients than a minimal medium are called rich media. One of this rich media is an amino acid-enriched medium. This medium can be constructed from a minimal medium with the standard metabolites explained in Sect. 2.2.2.1 (glucose, ammonia, phosphate, and sulfate), by adding the following set of amino acids: D-Alanine, L-Alanine, L-Arginine, L-Asparagine, L-Aspartate, D-Cysteine, L-Cysteine, L-Glutamine, L-Glutamate, Glycine, L-Histidine, L-Homoserine, L-Isoleucine, L-L-Leucine, L-Lysine, L-Methionine, L-Phenylalanine, L-Proline, D-Serine, L-Serine, L-Threonine, L-Tryptophan, L-Tyrosine, L-Valine. This set of amino acids enriches the minimal medium allowing the organism to take them as nutrients. Otherwise the organism would have to synthesize them, resulting in a more stringent environment for the organism. To simulate the presence of this set of amino acids in the medium, the exchange constraints bounds of these amino acids are set to $-10$ mmol/(gDW h).

Another famous rich medium is called Luria-Bertani Broth [73]. The Luria-Bertani Broth used in this thesis contains all the nutrients present in the amino acid-enriched medium, but it contains as additional compounds purines and pyrimidines, vitamins (namely biotin, pyridoxine, and thiamin), and also the nucleotide nicotinamide monocleotide [74]. The exchange constraints bounds of these compounds are usually set to $-10$ mmol/(gDW·h) ($\nu^{exchange}_{compound} \geq -10$) for *E. coli*.

### 2.2.3 *Activity and Essentialify of Genes and Reactions*

An important application of FBA is to compute the *activity* and *essentiality* of reactions in a network. These concepts can be applied either to genes or reactions, since a reaction is catalysed by an enzyme which at the same time is codified by a gene or a set of genes. Both concepts will be analysed in Chap. 4.

The concept of activity is quite simple. A reaction is said to be active when, given an external environment, the chosen reaction carries a non-zero flux. The concept of essentiality is more subtle. It refers to how a network, and thus the growth rate, is affected when one reaction is forced to be non-operative through the knockout of a reaction or of the corresponding gene.

To calculate the effect of the knockout of a reaction, the selected reaction is removed from the network, which is equivalent to force the chosen reaction to have a null flux. The new system is usually called a mutant. In terms of the notation used before, this is modelled as $\nu_i = 0$ with $i$ the removed reaction and $\nu_i = 0$ its flux. Thus, a FBA problem with a reaction $i$ constrained to be non-active is

$$\begin{array}{ll} \textbf{maximize} & \nu'_g \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = \vec{0} \\ \textbf{and} & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \\ & \nu_i = 0 \end{array}$$

where $\nu_g'$ denotes the growth rate of the mutant. As a consequence, the system can respond in three different ways as compared to the non-perturbed case $\nu_g$:

1. The growth rate is unaltered $\nu_g' = \nu_g$.
2. The growth rate is decreased $0 < \nu_g' < \nu_g$, which means that the biomass formation of the organism is reduced but the organism is still alive at the expense of losing some performance.
3. The growth rate takes a null value $\nu_g' = 0$, meaning that the performed knockout is lethal for the organism. This is the signature of essentiality.

It has been shown that FBA predicts gene essentiality with an accuracy of 90% [29] in *E. coli* under glucose aerobic conditions, which means that FBA is a reliable tool to predict whether a knockout will be lethal or not in this particular condition.

### *2.2.4 Flux Variability Analysis*

Sometimes it is useful to identify which are the minimum and maximum bounds that each reaction can take independently of the growth optimality condition. In this way, one can have an idea of the flux space for a particular environmental condition, and in particular which reactions can have a non-zero flux in a given environment, since some reactions may be active for low values of the growth rate but the same reactions must have a zero flux in order to ensure growth optimality. This may happen due to the fact that some reactions can compete with the growth reaction by consuming metabolites needed to grow and therefore this would reduce the flux through the biomass reaction. As a consequence, when one optimizes the flux through the biomass reaction, all reactions whose activation competes with the flux of the biomass reaction will have a null value in order to assure maximum growth conditions.

In addition to identifying those reactions that can compete with the growth rate, reactions whose minimum and maximum values are close indicate that they may be important for the organism since those reactions are allowed only to have a low variability in their fluxes. To know the minimum and maximum flux values of a reaction, one applies the technique called Flux Variability Analysis (FVA) [16, 17, 75].

In most applications of FVA, the biomass reaction is imposed to have a minimum value $\nu_g \geq \nu_g^{min}$ to ensure viability. Hence, one can consider that the limiting fluxes correspond to states where the organism is alive, even if the growth rate is not the maximum value that the organism can achieve. Using the mathematical notation used in Linear Programming computations, FVA for each flux of a metabolic reaction can be written as follows

$$\begin{array}{llll} \textbf{minimize} & \nu_i & \textbf{maximize} & \nu_i \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 & \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 \\ & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} & & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} \\ & \nu_g \geq \nu_g^{min} & & \nu_g \geq \nu_g^{min} \end{array}$$

However, it may happen that one is interested in capturing all the possible scenarios independently of the value of the flux of the biomass reaction, since in this way non-optimal/low-growth scenarios can be taken also into account. Therefore, FVA can be modified to compute the minimum and maximum possible values of the flux of each reaction regardless of the value of the biomass formation rate. To this end, the value of the flux of the biomass reaction is not constrained and any positive value is allowed, $\nu_g \geq 0$. Under this condition, one will obtain the maximal set of reactions that can be active in the considered medium independently of the rate of biomass formation. This variation of FVA [76, 77] will be used in Chaps. 4 and 5. Using the previous notation, this version of FVA, which we call *Biomass unconstrained Flux Variability Analysis*, can be written as

$$\begin{array}{lll} \textbf{minimize} & \nu_i & \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 & \\ & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} & \\ & \nu_g \geq 0 & \end{array} \qquad \begin{array}{lll} \textbf{maximize} & \nu_i & \\ \textbf{subject to} & \mathbf{S} \cdot \vec{\nu} = 0 & \\ & \vec{\alpha} \leq \vec{\nu} \leq \vec{\beta} & \\ & \nu_g \geq 0 & \end{array}$$

## 2.3 Model Organisms

Information about metabolism of specific organisms [31, 33, 34, 78–85]—most single cell—are gathered in databases, like the BiGG database [86], Kyoto Encyclopedia of Genes and Genomes (KEGG) [87], BioCyc/EcoCyc/MetaCyc [88], BRENDA [89], etc. The BiGG database deserves special attention in this thesis, since it has been extensively used as it contains full reconstructions of metabolic networks for specific organisms including all the biochemical reactions and the biomass formation function in order to compute FBA solutions for different organisms.

The BiGG database provides high-quality curated information. Network reconstructions coming from this database are structured in compartments like cytosol inside cells or periplasm—the space bordered by the inner and the outer membranes in Gram-negative bacteria. Therefore, metabolites present in different compartments of the organisms are treated as different nodes. Using different compartments allows the inclusion of transport systems in both the inner and outer membrane and thus the metabolic machinery of organisms is more accurately represented. As an example, water in the periplasm will be a different metabolite than water in the cytosol. In addition, a directed bipartite representation of the metabolic network can be constructed since in the databases reactants, products, reversible, and irreversible reactions are distinguished. Further, the BiGG database specifies which enzyme catalyses each reaction and also which gene or set of genes codifies each enzyme. However, reactions are also listed which have neither associated enzymes nor genes. It may be that, for these particular reactions, enzymes have not been identified yet or that some reactions are spontaneous and they can take place without the need of an enzyme.

It is important to notice that there exist different versions for each metabolic network of each organism. This happens due to the fact that the reconstructions of metabolic networks are constantly improved and, therefore, versions are constantly updated. As an example, the first version of *E. coli* [90] contained 660 genes, 627 reactions, and 438 metabolites, while the last version of *E. coli* [31] contains 1366 genes, 2250 reactions, and 1805 metabolites.

### 2.3.1 *Escherichia coli*

*Escherichia coli*, abbreviated as *E. coli*, is the most studied prokaryotic organism and it is the bacterial model that is most frequently used in experiments due to the ease of its manipulation. More precisely, the strain studied in this thesis is K-12 MG1655. This strain colonizes the lower gut of animals. Moreover, it has been maintained as a laboratory strain with minimal genetic manipulation.

Three versions of this strain have been used in this thesis. The first one is *i*AF1260, which can be obtained either from Ref. [29] or directly from the BiGG database. This version is based on an earlier reconstruction called *i*JR904 [91], on the annotation of the genome of *E. coli* from Ref. [92], on contents from the EcoCyc (an *E. coli* version of BioCyc) database [93] and on specific biochemical characterization studies from Ref. [29]. The *i*AF1260 version contains 2077 reactions, 1669 metabolites, and 1260 genes [29] (see Table 2.2). Metabolites are located in three compartments: exterior, periplasm and cytosol. Notice that although the exterior is not a real compartment, it is treated in this way in order to be able to use the exchange reactions explained before.

The most recent version of *E. coli* is *i*JO1366 [31]. It is an update of the *i*AF1260 version. EcoCyc [94] and the KEGG database [95] were used in order to improve the *i*AF1260 version, in addition to experimental techniques [31]. It contains 2250 biochemical reactions, 1805 metabolites, and 1366 genes [31] (see Table 2.2). Like in *i*AF1260, metabolites are located in cytosol, periplasm and exterior.

A simplified version called core *E. coli* metabolic model is also used, which can be obtained either from Refs. [12, 30] or the BiGG database. It is a condensed version of the genome-scale metabolic reconstruction *i*AF1260 that contains 73 metabolic reactions in central metabolism, 72 metabolites, and 136 genes (see Table 2.2). This network is complemented with a biomass formation reaction and an ATP maintenance reaction.

### 2.3.2 *Mycoplasma Pneumoniae*

*Mycoplasma pneumoniae*, abbreviated as *M. pneumoniae*, is a human pathogen of primary atypical pneumonia that has recently been proposed as a genome-reduced model organism for bacterial and archaeal systems biology [32, 33, 96, 97]. Interest

in this organism has grown recently since it lacks many anabolic processes and rescue pathways compared to more complex organisms. This in turn translates into a highly linear metabolism singularly suited to study basic metabolic functions [33]. This property will be again mentioned in Chaps. 3 and 4.

The first version of the metabolic network of *M. pneumoniae* used in this thesis was published in Ref. [32], where the authors integrated biochemical and computational studies, complementing the information using the KeGG database. Its metabolic reconstruction contains 187 reactions taking place in cytosol and in exterior, the number of metabolites is 228, and the number of genes is 140 (see Table 2.2).

The *i*JW145 version of *M. pneumoniae* is the last update [33]. This network was constructed by determining the behavior of the organism under different nutrition conditions, using literature information and experimental data. It contains 240 biochemical reactions, 266 metabolites, and 145 genes (see Table 2.2). Metabolites can be located in cytosol and exterior.

### 2.3.3 *Staphylococcus aureus*

*Staphylococcus aureus*, abbreviated as *S. aureus*, is found in the human respiratory tract and on the skin. It is an anaerobic bacterium which is present world-wide, and it is a common cause of skin infections, respiratory disease, and food poisoning. The strain used in this thesis is N315, a major pathogen which is able to acquire antibiotic-resistance [98].

The *i*SB619 version of *S. aureus* can be obtained either from the BiGG database or from Ref. [34]. To construct this model, the authors used the KeGG database and the Comprehensive Microbial Resource (CMR) at The Institute for Genomic Research (TIGR) website [99]. Missing functions were annotated based on reported evidence from this organism, as well as for *Bacillus subtilis* and *E. coli*. The number of reactions is 642 and the number of metabolites is 644 (see Table 2.2). Like in *M. pneumoniae*, there are only cytosol and exterior compartments.

**Table 2.2** Summary of the properties of all metabolic reconstructions used in this thesis. $N_R$, $N_M$, and $N_G$ stand for the number of reactions, metabolites, and metabolic genes respectively. Metabolites in different compartments are treated as different metabolites

| Organism | $N_R$ | $N_M$ | $N_G$ | Source |
|---|---|---|---|---|
| *E. coli i*AF1260 | 2077 | 1669 | 1260 | Ref. [29], BiGG |
| *E. coli i*JO1366 | 2250 | 1805 | 1366 | Ref. [31] |
| *E. coli* core model | 73 | 72 | 136 | Refs. [12, 30], BiGG |
| *M. pneumoniae* | 187 | 228 | 140 | Ref. [32] |
| *M. pneumoniae i*JW145 | 240 | 266 | 145 | Ref. [33] |
| *S. aureus i*SB619 | 642 | 644 | 619 | Ref. [34], BiGG |

# References


1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Newman M (2010) Networks: an introduction. Oxford University Press, New York
3. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. Nature 407:651–654
4. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42
5. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc R Soc Lond B 268:1803–1810
6. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
7. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cells functional organization. Nat Rev Genet 5:101–113
8. Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. BioEssays 27:176–188
9. Motter AE, Gulbahce N, Almaas E, Barabási AL (2008) Predicting synthetic rescues in metabolic networks. Mol Syst Biol 4:168
10. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, New York
11. Alon U (2006) An introduction to systems biology: design principles of biological circuits. CRC Press, Boca Raton
12. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28:245–248
13. Varma A, Palsson BØ (1993) Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165(4):477–502
14. Suthers PF, Zomorrodi A, Maranas CD (2009) Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol 5:301
15. Barve A, Rodrigues JFM, Wagner A (2012) Superessential reactions in metabolic networks. Proc Natl Acad Sci USA 1091:E1121–E1130
16. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
17. Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. BMC Bioinform 11:489
18. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427(6977):839–843
19. Guillaume JL, Latapy M (2006) Bipartite graphs as models of complex networks. Phyics A 371:795–813
20. Güell O, Serrano MÁ, Sagués F (2014) Environmental dependence of the activity and essentiality of reactions in the metabolism of *Escherichia coli*. In: Engineering of Chemical Complexity II. World Scientific Publishing, Singapore, pp 39–56. ISBN 978-981-4616-12-6
21. Holme P, Liljeros F, Edling CR, Kim BJ (2003) Network bipartivity. Phys Rev E 68(5):056107
22. Ma HW, Zeng AP (2003a) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19:270–277
23. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
24. Barabási AL, Bonabeau E (2003) Scale-free networks. Sci Am 288(5):50–59
25. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(21):4947–4957
26. Keller EF (2005) Revisiting "scale-free" networks. BioEssays 27(10):1060–1068
27. Tanaka R (2005) Scale-rich metabolic networks. Phys Rev Lett 94(16):168101
28. Kim P et al (2007) Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. Proc Natl Acad Sci USA 104(34):13638–13642
29. Feist AM et al (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121

30. Orth JD, Fleming RM, Palsson BØ (2009) EcoSal—*Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, DC
31. Orth JD et al (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. Mol Syst Biol 7:535
32. Yus E et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268
33. Wodke JAH et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol Syst Biol 9:653
34. Becker SA, Palsson BØ (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5:8
35. Güell O, Sagués F, Serrano MÁ (2012) Predicting effects of structural stress in a genome-reduced model bacterial metabolism. Sci Rep 2:621
36. Ma HW, Zeng AP (2005) Reconstruction of metabolic networks from genome information and its structural. Computational systems biology. Academic Press, New York
37. Kriete A, Eils R (2005) Computational systems biology. Academic Press, New York
38. Arita M (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101(6):1543–1547
39. Gao JT, Guimerà R, Li H, Pinto IM, Sales-Pardo M, Wai SC, Rubinstein B, Li R (2011) Modular coherence of protein dynamics in yeast cell polarity system. Proc Natl Acad Sci USA 108(18):7647–7652
40. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174
41. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105:1118–1123
42. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
43. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110
44. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 10:P10008
45. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
46. Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, Englewood Cliffs
47. Ma HW, Zeng AP (2003b) The connectivity stucture, giant strong component and centrality of metabolic networks. Bioinformatics 19:1423–1430
48. Boguñá M, Ángeles M (2005) Generalized percolation in random directed networks. Phys Rev E 72:016106
49. Serrano MÁ, De Los P (2008) Structural efficiency of percolated landscapes in flow networks. PLoS ONE 3:e3654
50. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
51. Erdös P, Rényi A (1959) On random graphs I. Publ Math Debr 6:290–297
52. Erdös P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
53. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithm 6:161–179
54. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:026118
55. Milo R et al (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827
56. Smart AG, Amaral LAN, Ottino J (2008) Cascading failure and robustness in metabolic networks. Proc Natl Acad Sci USA 105:13223–13228

57. Güell O, Sagués F, Basler G, Nikoloski Z, Serrano MÁ (2012) Assessing the significance of knockout cascades in metabolic networks. J Comp Int Sci 3(1–2):45–53
58. Basler G, Ebenhöh O, Selbig J, Nikoloski Z (2011) Mass-balanced randomization of metabolic networks. Bioinformatics 27:1397–1403
59. Basler G, Grimbs S, Ebenhöh O, Selbig J, Nikoloski Z (2012) Evolutionary significance of metabolic network properties. J R Soc Interface 9:1168–1176
60. Costa RS, Machado D, Rocha I, Ferreira EC (2011) Critical perspective on the consequences of the limited availability of kinetic data in metabolic dynamic modelling. IET Syst Biol 5(3):157–163
61. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17(2):53–60
62. Price N, Reed J, Papin J, Wiback S, Palsson BØ (2003) Network-based analysis of metabolic regulation in the human red blood cell. J Theor Biol 225(2):185–194
63. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5(1):320
64. Terzer M, Maynard ND, Covert MW, Stelling J (2009) Genome-scale metabolic networks. Wiley Interdiscip. Rev. Syst. Biol. Med. 1(3):285–297
65. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol Syst Biol 9(1):661
66. Schilling CH, Palsson BØ (1998) The underlying pathway structure of biochemical reaction networks. Proc Natl Acad Sci USA 95:4193–4198
67. Schilling CH, Edwards JS, Letscher D, Palsson BØ (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnol Bioeng 71:286–306
68. Makhorin A (2001) GNU linear programming kit. Moscow Aviation Institute, Moscow
69. Ceron R (2006) The GNU linear programming kit, Part 1: introduction to linear optimization. IBM, Raleigh
70. Ceron R (2006b) The GNU linear programming kit, Part 2: intermediate problems in linear programming. IBM, Raleigh
71. Ceron R (2006c) The GNU linear programming kit, Part 3: advanced problems and elegant solutions. IBM, Raleigh
72. Murty KG (1983) Linear programming, vol 57. Wiley, New York
73. Sezonov G, Joseleau-Petit D, D'Ari R (2007) Escherichia coli physiology in Luria-Bertani broth. J Bacteriol 189:8746–8749
74. Wunderlich Z, Mirny LA (2006) Using the topology of metabolic networks to predict viability of mutant straints. Biophys J 91:2304–2311
75. Müller AC, Bockmayr A (2013) Fast thermodynamically constrained flux variability analysis. Bioinformatics 29:903–909
76. Reed JL, Palsson BØ (2004) Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14(9):1797–1805
77. Güell O, Sagués F, Serrano MÁ (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol 10(5):e1003637
78. Duarte NC, Herrgard MJ, Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14:1298–1309
79. Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T (2004) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol Syst Biol 2:2006
80. Duarte NC, Becker SS, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104(6):1777–1782

81. Jamshidi N, Palsson BØ (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26
82. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2008) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143
83. Senger RS, Papoutsakis ET (2008) Genome-scale model for *Clostridium acetobutylicum*: part I. Metabolic network resolution and analysis. Biotechnol Bioeng 101(5):1036–1052
84. Raghunathan A, Reed J, Shin S, Palsson BØ, Daefler S (2009) Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. BMC Syst Biol 3(1):38
85. Thiele I et al (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31(5):419–425
86. Schellenberger J, Park JO, Conrad TC, Palsson BØ (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinform 11:213
87. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
88. Caspi R et al (2012) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 40(D1):D742–D753
89. Schomburg I et al (2012) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res 41:D764–D772
90. Edwards JS, Ibarra RU, Palsson BØ (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97:5528–5533
91. Reed JL, Vo TD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54
92. Riley M et al (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. Nucleic Acids Res 34(1):1–9
93. Keseler IM et al (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 33(suppl 1):D334–D337
94. Keseler IM et al (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res 37(suppl 1):D464–D470
95. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(suppl 1):D355–D360
96. Kühner S et al (2009) Proteome organization in a genome-reduced bacterium. Science 326:1235–1240
97. Güell M et al (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326:1268–1271
98. Kuroda M et al (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357(9264):1225–1240
99. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. Nucleic Acids Res 29(1):123–125

http://www.springer.com/978-3-319-63999-4

A Network-Based Approach to Cell Metabolism
From Structure to Flux Balances
Güell, O.
2017, XX, 151 p. 49 illus., Hardcover
ISBN: 978-3-319-63999-4