Chapter 8

Large cell models

Hugo Dourado, Anne Goelzer, Pranas Grigaitis, Wolfram Liebermeister, Elad Noor

Chapter highlights

- Resource balance analysis models are cell models based on three basic constraints: stationary fluxes (balancing production and consumption fluxes, uptake and excretion fluxes, as well as compound dilution by cell growth); capacity constraints relating fluxes to the amounts of catalzing enzymes or other machines; and density constraints, limiting molecule amounts in cell compartments, or molecule concentrations.
- These constraints define a set of possible cell states, on which optimisation can be performed - for example, to maximise protein production at a given growth rate, or to maximise growth rate itself.
- Resource allocation models can be large or small and have been implemented as RBA models in a narrow sense, ME models, and pc-models.

8.1 Introduction: From small models and metabolic models to cell models

8.1.1 How can we build detailed models of entire growing cells?

In the previous chapter on growing cells, we derived possible growth rates and corresponding cell compositions and metabolic strategies based on resource allocation ideas. We asked what amounts of cellular machines will allow a cell to self-replicate, and possibly, to self-replicate at a maximal possible speed. However, the models described were rather simple, with reactions coarse-grained into proteome sectors. Following the same idea, would it be possible to model an entire cell by a one-to-one mapping of all of its reactions, including realistic kinetics and find the configuration for this entire system that would allow the cell to optimally self-replicate? Such models would in principle be capable of describing the detailed resource allocation into each reaction in the cell: they would extend genome-scale models of metabolism (such as in FBA) to models covering all cell subsystems, at the same time accounting for all important constraints on growth, in particular realistic kinetics.

So why don't we directly describe metabolism and macromolecule production in one single, detailed model? If do that and attempt to find the optimal state, we quickly see that the optimality problems for genome-size models are hard to solve, and maybe unsolvable with current means. The reason is simple: due to non-linearities in the rate laws (where rates are typically dependent on substrates and machines, and are therefore at least quadratic), the optimality problem may have multiple local optima. Since the space for optimization is very high-dimensional (with the number of dimensions on the order of the number of compounds and reactions considered) and non-convex (because certain regions of the initially convex solution space are excluded by thermodynamic constraints), the global optimum may be hard (or in practice, impossible) to find.

We can also see this more formally. While kinetic self-replicator models could be formulated, the resulting problems would be non-convex [1, 2, 3] and potentially hard to solve. How do we know this? In the optimal solution, the flux distribution must be an elementary flux mode [2]. Since no better method is known, to find the globally optimal metabolic fluxes, protein and metabolite abundances all elementary flux modes must be enumerated. For large-scale metabolic networks this is hard, and in fact this problem is computationally #P-complete.[4], which prevents its resolution for large-scale models. However, recent attempts of integration of thermodynamics constraints in EFM computation drastically reduced their number [5], which may provide significant help in future developments to solve large self-replicator models.

So in order to model a cell in high resolution (including individual metabolic reactions and their enzymes), what can we do? In previous chapters, we approached this problem from two directions. Flux Balance Analysis (FBA) models, as described in chapter 5, capture metabolic fluxes in detail, but ignore the rest of the cell (such as macromolecule production; and also the fact that all cellular compounds are diluted by cell growth). Some FBA variants account for extra empirical constraints on the total concentration of metabolic enzymes (FBA with molecular crowding, or FBAwMC), or on proteome sectors (Constrained-Allocation FBA, or CAFBA). While these can predict metabolic states more reliably, the empirical constraints come as model assumptions and thus cannot be understood by the models themselves. "Molenaar-type" self-replicator models, as described in chapter **??**, describe a growing cell directly and consider whole-cell aspects such as dilution by growth. However, they contain only very few reactions. Larger cell processes are lumped into a few global variables, ignoring the details. This makes it impossible to model the effects of detailed changes (e.g. the fate of individual metabolites, the inhibition of single proteins, or deletions of single genes).

So why don't we just combine the two approaches? In the chapters 6 and **??** we already did this, to an extent: we considered a metabolic (FBA-like) model to compute biomass/enzyme productivity, and then used a formula derived from bacterial growth laws (see chapter **??**) to convert biomass/enzyme productivity into cell growth rate. But there were still some disadvantages, most importantly, that the biomass composition has to be known. Biological facts such as biomass composition should not just come as empirical knowledge, but should be explained *ab initio*. Subdividing a system into metabolism and other systems (with "fixed interfaces", like a predefined biomass composition) makes it impossible to understand details of complex behavior. For example, if some proteins are needed to retrieve iron via metabolism and other proteins (or possibly, the same proteins) also contain iron, this will creates a complicated protein allocation response to changing iron supply. Iron depletion is expected lead to an upregulation of the first type of proteins, and a downregulation of the second type. However, FBA can only capture the first effect, because its biomass function is fixed.

8.1.2 Linearization: replacing kinetics by catalytic constraints

So how can large-scale models be made tractable? If we radically linearize all formulae, then instead of a biconvex or convex/concave problem, we obtain a linear problem (a bit like FBA); more precisely, a system of linear equalities and inequalities that define a set of feasible states. This set is a polytope, and linear optimality problems on this set can be solved easily. More specifically, to model metabolism in a growing cell, we need to consider dilution in the growing cell volume. If compounds are described by concentration, dilution can be effectively modeled by "consuming reactions", with a flux given by $v^{\mathrm{dil}} = \mu\, c$, the compound concentration multiplied by the growth rate. By adding these hypothetical dilution reactions to the metabolic network, we obtain a new stationarity condition $\mathbf{N}\,\mathbf{v} = \mu\,\mathbf{c}$ that connects the vectors of fluxes and compound concentrations, and in which the growth rate μ appears as a parameter. For each choice of the parameter μ, we can ask whether a feasible steady growth state – i.e. a feasible combination of $\mathbf{v}$ and $\mathbf{c}$ exists, Furthermore, the feasible combinations $(\mu, \mathbf{v}, \mathbf{c})$ form a convex set, with possible solutions $(\mathbf{v}, \mathbf{c})$ for low values of μ and no solutions above a critial value $\mu^{\max}$. Finding this critical value – the maximal possible growth rate for our model – as well as the corresponding optimal fluxes $\mathbf{v}$ and compound concentrations $\mathbf{c}$ is relatively easy, and can be done by solving a series of Linear Programming problems (checking for potential solutions $(\mathbf{v}, \mathbf{c})$ for different values of μ).

The main downside of this approach is that all relationships between models variables need to be linearized. This concerns, most importantly, all catalyzed processes, in which we assume a linear dependence between catalyzed flux and catalyst (enzyme or machine) concentration, but ignore the dependence on the concentrations of substrates, products, cofactors, or additional regulators. What does this mean in practice? As we know from chapter 3, typical enzymatic rate laws have the form $v = e\ k(\mathbf{c})$: the rate v is proportional to enzyme level e and enzyme efficiency k, which is given by a kinetic rate law $k(\mathbf{c})$, a nonlinear function of the metabolite concentrations. Depending on the context, k is also called catalytic rate or apparent k_{cat}. The kinetic rate laws $k(c)$ have typical shapes, as described in chapter 3. To linearize the expression for v, while keeping the dependence on e, we need to replace k by a fixed number, a model parameter. If the metabolite concentrations were known (experimentally, or from kinetic models under optimality assumptions, see chapter 6), the value of k could be computed. Otherwise, it can also be determined experimentally, by measuring v and e and setting $k = v/e$. Obviously, in reality, neither $\mathbf{c}$ nor k will be fixed and given, but for our linearized model, we need to assume this. This holds both for metabolic reactions (with enzymes as catalysts) and for macromolecular reactions (with molecular machines as catalysts). Under this assumption, we can replace all kinetic constraints by two linar constraints on the enzyme to fulfill its function as $-e\ k^{'} \leq v \leq e\ k$, where coefficients k and $k^{'}$ are the approximation coefficients of the enzyme kinetics in the forward and backward direction respectively. These constraints, which replace the kinetic rate laws, are called capacity constraints and resemble the linear flux/enzyme relation in FBA with molecular crowding. Finally, there is one important exception: like in satFBA. some kinetic rate laws can remain, namely those in which the metabolite concentrations themselves are not variables (or disregarded completely), but are given as model parameters. This concerns, in particular, all transport reactions whose substrate is an extracellular compound with a known concentration in the growth medium. Nonlinearities can be kept as long as they do not depend on decision variables (even if the nonlinear problem can still sometimes be solved for a few nonlinearities if convexity is kept)

To obtain an ideal detailed cell model, the self-replicator models from the previous chapter would have to be extended with explicit kinetics and thermodynamic constraints. Since we know that such problems are hard to solve , simplifications are currently necessary to keep linearity (and convexity) to solve RBA optimization problems for large-scale models. The main simplification is performed at the level of the enzyme efficiencies by ignoring kinetics or thermodynamics relationships. However, RBA is able to predict realistic complex adaptations (e.g. some proteins contain iron; under iron starvation, the cell may *increase* the import of iron, but also *avoid* using proteins that contain iron, and the pathways in which they operate), even in complex growth conditions [6, 7]. These effects cannot be captured by standard CBM methods in an autonomous way, i.e. without the addition of specific constraints on fluxes. Altogether, RBA provides a reasonable compromise between numerical tractability, cell phenotype prediction accuracy and model complexity.

8.2 Describing cells through constraints - the four constraints of fine-grained resource allocation models

The canonical FBA/genome-scale metabolic models (GEMs), as described above (and in previous chapters), hold a number of assumptions which allow metabolic states irregardless of the corresponding resource allocation pattern. Therefore, we need to impose sets of additional constraints on metabolic networks. The general description of these constraints in fact is the same as for coarse-grained small (e.g. Molenaar-type) models, only the number of individual constraints increases. Although the precise formulations vary, resource allocation models build on three principal types of constraints:

- Mass-conservation constraints
- Flux coupling constraints
- Compartment capacity constraints

Alongside these three major classes there is another set of constraints, which we could call "environment" constraints - these correspond to, e.g. the composition of growth medium, biomass composition at at given growth rate μ, etc. The fine-grained models build on the genome-scale metabolic models to encompass all the reactions that potentially could happen in a metabolic network, and different extensions of GEMs were proposed to predict optimal resource allocation in different microorganisms [8]. We will now expand on the three constraints listed above. Note that the description is not exhaustive and peculiarities may vary among different formulations.

First, the *mass-conservation* constraints define the metabolic network (stoichiometry and relation between fluxes). The initial building blocks of these extended models are GEMs, and thus the metabolic network stoichiometry is already there; what remains to be defined are the protein turnover processes. We consider 4 types of protein turnover reactions in fine-grained resource allocation models: protein synthesis, folding, degradation and dilution-by-growth. So, for every protein present in such a model, we add these four reactions: two of them, translation and degradation, include the stoichiometry of amino acids needed for its translation and released during degradation based on the protein sequence. Protein folding and dilution are modeled by the same types of reactions for different proteins (a series of two steps, "unfolded" $\rightarrow$ "folded" $\rightarrow \emptyset$).

Next, the flux *coupling* constraints couple the [metabolic] fluxes with protein usage: usually, the usage scales with the catalytic turnover value k_{cat} of the enzyme. In this step we have to collect the kinetic information (most usually, catalytic constant/k_{cat} values), which are used as model input. We establish the coupling between fluxes and protein synthesis by setting $v = k_{\text{cat}}\ e\ f$, where e is the enzyme concentration and $0 < f \leq 1$ is an efficiency term summarizing the effects of reaction thermodynamics, enzyme saturation, and possibly small-molecule regulation. Coupling constraints are introduced to couple both (i) the metabolic reactions with enzyme usage (as described above) and (ii) protein turnover reactions with the respective macromolecular machinery (e.g. sum demand of ribosomes for protein translation, $v_{\text{translation}} = [\text{Ribosome}] \times k_{\text{cat,ribosome}}$). The sheer number of the kinetic parameters needed for formulating the coupling constraints in the fine-grained models requires the modeller to consider different assumptions and simplifications when building and parametrizing these models, as briefly discussed below.

The number of processes described in a fine-grained manner directly translates to the number of reactions and metabolites in the model. For instance, transcription is modelled explicitly in the ME-models [9]. The modellers' decision is key here: under assumption that transcription and translation form a linear pathway with fixed scaling factors (i.e. there is a fixed ratio of peptides translated per mRNA transcribed), the flux through mRNA translation reaction can be computed post-optimization based on the flux through the protein translation reaction. Explicit modelling of transcription would require describing processes of mRNA transcription, processing, export from nucleus, and then cytosolic degradation after the mRNA is translated – for each of the transcripts, with precise stoichiometry and a new set of coupling constraints.

The next issue is kinetic parametrization of these fine-grained models. We currently can use only very simplified kinetics in the models (flux coupling $v = k_{\text{cat}}\ e\ f$), and simplify such factors as enzyme saturation and thermodynamic driving force into a single value of factor f. Two approaches are used to deal with this, as a large fraction of parameters are not even available. First, condition-dependent kinetic parameters ("apparent catalytic constants", k_{app}) are fitted from experimental (mostly quantitative proteomics) data (setting $k_{\text{eff}} = k_{\text{cat}}\ \alpha$, where $0 < \alpha \leq 1$) with a value α chosen to match predicted enzyme abundance and experimental measurements. Otherwise, for the enzymes with measured k_{cat} values, we can assume that enzymes work at their maximal rate, i.e. the saturation function $f = 1$. Then the model computes the *minimal* protein requirement to sustain the flux through the metabolic reactions. The comparison of *minimal predicted* vs. *observed* protein abundance can represent the "apparent saturation", or "overcapacity" of enzymes. For instance, it is common in yeast *S. cerevisiae* that the flux and not protein expression varies across conditions, and the relationship between predicted and measured expression can suggest the nature of the observed protein expression [10].

The final layer of information in the fine-grained resource allocation models is a set of *capacity* constraints. These constraints describe the [upper] limit of cellular process(es), e.g. maximal protein capacity of a compartment. The capacity constraints are formulated as weighted sums of protein abundance, and usual weighing multipliers are proportional to the molecular weight of the protein. Usually, the capacity constraints are expressed in terms of [maximal] *mass*, *area*, and *volume* of the compartment (e.g. "what is the maximal mass the mitochondrial proteins can take up in gDW of cells?"). Based on the biological interpretation of the constraints, we formulate the weighing multipliers to represent either of the metrics (mass/area/volume) that every protein occupies.

The capacity constraints can be both *equality* and *inequality* constraints: more frequent are the latter (usually defining the "upper limit" of, e.g. amount of protein targeted to mitochondria). However, some cell properties should be described through equality constraints: one of these is the protein density of biomass, defining the "target" protein translation per gram dry cell biomass.

Protein *abundance* versus *concentration* in fine-grained resource allocation models Here we would like to include a relevant note for interpretation of the output of the fine-grained resource allocation models. Both the classical FBA and these extensions do not consider "metabolite concentration" as a concept: optimization variables are all *fluxes*. Frameworks discussed in this chapter model protein synthesis from amino acids and energy equivalents explicitly, with a typical flux dimension of $mmol\ gDW^{-1}\ h^{-1}$ (as for any other fluxes). To compute the amount of protein that has to be produced in the steady-state growth, we should consider the flux balance for the protein e: $v_{\text{synthesis,e}} = v_{\text{degradation,e}} + v_{\text{dilution,e}}$, or, rewritten with the respective parameters, $v_{\text{synthesis,e}} = (k_{\text{deg,e}} + \mu)\ e$. Here, $k_{\text{deg,e}}$ is the degradation rate for the protein e, and μ is the specific growth ($=$ dilution-by-growth) rate. The $[e]$ in the rewritten equation holds dimension of mmol gDW^{-1}, which is protein *abundance*, rather than *concentration*.

The predicted amount of protein in cells can be compared to experimental measurements in two ways. First option is to convert abundance to concentration using the relationship between the cell volume and dry weight (e.g. $V_{gDW} = 1.7\, mL\ gDW^{-1}$ in *Saccharomyces cerevisiae*, [11]). Alternatively, proteome mass fractions are a popular unit in label-free mass spectrometry-based protein quantification, a popular method in quantitative microbiology. Respectively, predicted proteome mass fractions can be inferred by converting protein abundance in $mmol$ to g, and scaling to the protein content in dry cell biomass. Here, it is important to consider the conversion factors (protein content in dry biomass). Fast-growing bacteria, e.g. *E. coli*, maintains rather constant protein content in dry weight across growth rates (ca. 0.55 (g protein) gDW^{-1}) [12, 13]. On the contrary, the protein content is known to vary in *S. cerevisiae* as a function of growth rate [11].

8.3 Resource balance analysis

Resource Balance Analysis (RBA) has been developed as (and is considered) a flexible and generic modeling framework which describes the functioning of an organism using the most relevant set of linear equality and inequality constraints. Importantly, RBA does not refer to a *specific* unique set of constraints, but it provides the framework in which such constraints can be taken into account. Here we distinguish the general RBA framework from a specific type of RBA model, i.e. the model of a specific organism, which contains by definition a set of specific constraints. The general structure of such models, as well as the typical constraints, are shown in Figure 8.1.

8.3.1 Mathematical description

Notation. Below A^T refers to the transpose of the matrix A. $\mathbb{R}^n_{>0} \triangleq \{x \in \mathbb{R}^n\, | x_i > 0 \text{ for all } i \in \{1, \cdots, n\}\}$, $\mathbb{R}_{>0} \triangleq \mathbb{R}^1_{>0}$, $\mathbb{R}^n_{\geq 0} \triangleq \{x \in \mathbb{R}^n\, |\, x_i \geq 0 \text{ for all } i \in \{1, \cdots, n\}\}$ and $\mathbb{R}_{\geq 0} \triangleq \mathbb{R}^1_{\geq 0}$.

In a standard RBA model, we consider balanced growth , that is, the average state of a cell in a cell bacterial population growing exponentially at the specific (constant) growth rate $\mu \geq 0$, i.e. the amount of produced biomass per biomass per cell per unit of time. Our simulated average cell is composed of different molecule species:

(i) n_y types of molecular machines, which can be subdivided further into n_e enzymes and transporters involved in the metabolic network $\mathbb{E} \triangleq (\mathrm{E}_1, \ldots, \mathrm{E}_{n_e})$ at the concentrations $\mathbf{e} \triangleq (e_1, \ldots, e_{n_e})^T$ and metabolic fluxes $\nu \triangleq (\nu_1, \ldots, \nu_{n_e})^T$; and n_m macromolecular machines $\mathbb{M} \triangleq (\mathrm{M}_1, \ldots, \mathrm{M}_{n_m})$ involved in non-metabolic cellular processes, such as the translation apparatus, at the concentrations $\mathbf{m} \triangleq (m_1, \ldots, m_{n_m})^T$;

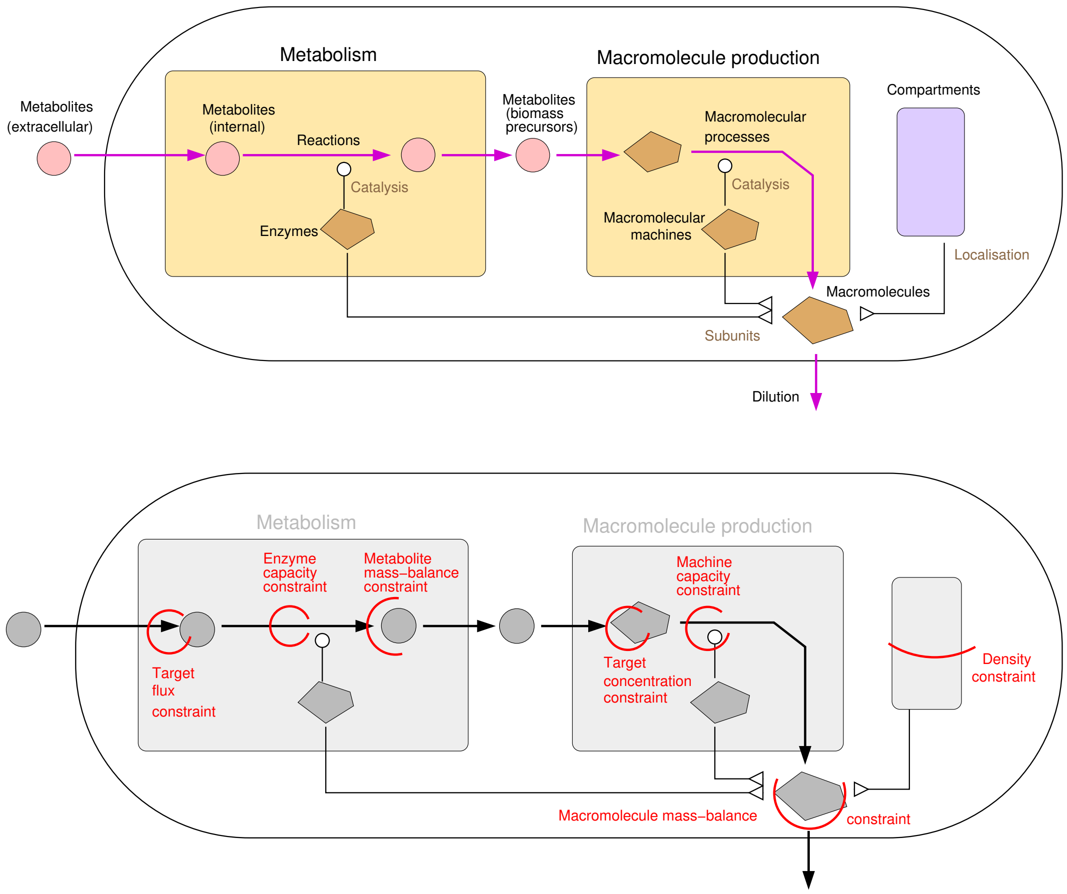

Figure 8.1: Resource balance analysis model: overview of biological components and mathematical constraints. (a) Typically, an RBA model describes metabolisms and macromolecule production in a growing cell (yellow blocks). Precursors from metabolism are needed to produce macromolecules, and some macromolecules serve as enzymes to catalyze metabolic reactions. In addition, macromolecules are diluted and are localized in cell compartments. (b) Mathematical constraints. The variables and processes described by an RBA model must satisfy a number of constraints, include mass-balance constraints (between production, degradation, and dilution of compounds); capacity constraints (relating process velocities to the concentrations of catalysts); density constraints (on the total amount of compounds in a cell compartment); and possibly empirical physiological constraints on any types of "target variables", to ensure realistic models.

(ii) n_p proteins $\mathbb{P} \triangleq \{\mathrm{P}_1, \ldots, \mathrm{P}_{n_p}\}$ belonging to unspecified cellular processes. $\mathbf{p} \triangleq (p_1, \ldots, p_{n_p})^T$ denotes the set of concentrations of $\mathbb{P}$;

(iii) n_s intracellular and mass-balanced metabolites $\mathbb{S} \triangleq (\mathrm{S}_1, \ldots, \mathrm{S}_{n_s})$. Within the set $\mathbb{S}$, we distinguish a subset $\mathbb{B} \triangleq (\mathrm{B}_1, \ldots, \mathrm{B}_{n_b})$ of abundant metabolites which have fixed growth-independent concentrations $\bar{\mathbf{b}} \triangleq (\bar{b}_1, \ldots, \bar{b}_{n_b})^T$ (and usually coincide with biomass macro-components such as DNA, cell wall or plasmic membrane). We also consider a set of extracellular metabolites $\mathbb{S}_{\text{ext}} \triangleq (\mathrm{S}_{\text{ext},1}, \ldots, \mathrm{S}_{\text{ext},\mathrm{n}_{\text{ext}}})$ of concentrations $\mathbf{s}_{\text{ext}} \triangleq (s_{\text{ext},1}, \ldots, s_{\text{ext},\mathrm{n}_{\text{ext}}})^T$ that are not mass-balanced.

Finally, let us introduce the vector $\mathbf{y}^T \triangleq (\mathbf{e}^T, \mathbf{m}^T)$ of concentrations of molecular machines of size n_y. Typical units of concentrations $\mathbf{e}$, $\mathbf{m}$ and $\mathbf{p}$ are in millimoles per gram of cell dry weight, and fluxes ν in millimoles per gram of cell dry weight per unit of time.

For a given cell growth rate $\mu \geq 0$, the RBA optimization problem (named $\mathcal{P}_{\text{rba}}(\mu)$) can be formalized mathematically as follows.

For a fixed vector of concentrations $\mathbf{p} \in \mathbb{R}^{Nn_p}_{>0}$ and the given growth rate $\mu \geq 0$,

$$\begin{array}{ll}
\text{find possible cell states} & \mathbf{y} \in \mathbb{R}^{n_y}_{\geq 0}, \nu \in \mathbb{R}^{n_e}, \\
\text{subject to} & \\
(C_1) & -\boldsymbol{\Omega}\nu + \mu(\mathbf{C}^{\mathrm{S}}_{\mathrm{Y}}\,\mathbf{y} + \mathbf{C}^{\mathrm{S}}_{\mathrm{B}}\,\bar{\mathbf{b}} + \mathbf{C}^{\mathrm{S}}_{\mathrm{P}}\,\mathbf{p}) = 0 \\
(C_{2a}) & \mu(\mathbf{C}^{\mathrm{M}}_{\mathrm{Y}}\,\mathbf{y} + \mathbf{C}^{\mathrm{M}}_{\mathrm{P}}\,\mathbf{p}) - \mathbf{K}_{\mathrm{T}}\,\mathbf{y} \leq 0 \\
(C_{2b}) & -\mathbf{K}^{'}_{\mathrm{E}}\,\mathbf{y} \leq \nu \leq \mathbf{K}_{\mathrm{E}}\,\mathbf{y} \\
(C_3) & \mathbf{C}^{\mathrm{D}}_{\mathrm{Y}}\,\mathbf{y} + \mathbf{C}^{\mathrm{D}}_{\mathrm{P}}\,\mathbf{p} - \bar{\mathbf{d}} \leq 0
\end{array}$$

where all the inequalities are defined componentwise and:

- $\boldsymbol{\Omega}$ is the stoichiometry matrix of the metabolic network of size $n_s \times n_e$, where Ω_{ij} corresponds to the stoichiometry of metabolite $\mathrm{S_i}$ in the j-th enzymatic reaction;
- $\mathbf{C}^{\mathrm{S}}_{\mathrm{Y}}$ (resp. $\mathbf{C}^{\mathrm{S}}_{\mathrm{P}}$) is an n_s n_y (resp. n_s n_p) matrix where each coefficient $\mathrm{C}^{\mathrm{S}}_{\mathrm{Y_{ij}}}$ corresponds to the number of metabolite $\mathrm{S_i}$ consumed (or produced) for the synthesis of one machine $\mathrm{Y_j}$ (resp. $\mathrm{P_j}$); $\mathrm{C}^{\mathrm{S}}_{\mathrm{Y_{ij}}}$ is then positive, negative or null if $\mathrm{S_i}$ is produced, consumed or not involved in the the synthesis of one machine $\mathrm{Y_j}$ (resp. $\mathrm{P_j}$);
- $\mathbf{C}^{\mathrm{S}}_{\mathrm{B}}$ is an $n_s \times n_b$ matrix in which each coefficient $\mathrm{C}^{\mathrm{S}}_{\mathrm{B_{ij}}}$ corresponds to a metabolite $\mathrm{S_i}$ consumed (or produced) for the synthesis of one $\mathrm{B_j}$;
- $\mathbf{K}_{\mathrm{T}}$ ($\mathbf{K}_{\mathrm{E}}$ and $\mathbf{K}^{'}_{\mathrm{E}}$, respectively) are matrices of size $n_m \times n_y$ ($n_e \times n_y$, respectively) in which each coefficient $\mathrm{k_{T_i}}$ ($\mathrm{k_{E_i}}$ and $\mathrm{k}^{'}_{\mathrm{E_i}}$, respectively) is positive and corresponds to the efficiency of molecular machine $\mathrm{M_i}$, i.e. the rate of the process per amount of the catalyzing molecular machine, (the efficiency of the enzyme $\mathrm{E_i}$ in forward and backward sense, respectively);
- $\mathbf{C}^{\mathrm{M}}_{\mathrm{Y}}$ (resp. $\mathbf{C}^{\mathrm{M}}_{\mathrm{P}}$) is an $n_m \times n_y$ (resp. $n_m \times n_p$) matrix in which each coefficient $\mathrm{C}^{\mathrm{M}}_{\mathrm{Y_{ij}}}$ typically corresponds to the length in amino acids of the machine $\mathrm{Y_j}$ (resp. $\mathrm{P_j}$). In some cases (for instance for the constraints on protein chaperoning), the length in amino acids can be multiplied by a coefficient, such as the fraction of the whole proteome that necessitates chaperoning;
- $\bar{\mathbf{d}}$ is a vector of size n_c, where n_c is the number of compartments (compartment membrane and/or compartment interior for which density constraints are considered. $\bar{\mathrm{d}}^{\mathrm{i}}$ is the density of molecular entities within the volume or surface area. Densities are typically expressed as a number of amino-acid residues by volume or surface area.
- $\mathbf{C}^{\mathrm{D}}_{\mathrm{Y}}$ (resp. $\mathbf{C}^{\mathrm{D}}_{\mathrm{P}}$) is an $n_c \times N_y$ (resp. $n_c \times n_p$) matrix in which each coefficient $\mathrm{C}^{\mathrm{D}}_{\mathrm{Y_{ij}}}$ corresponds to the density of one machine $\mathrm{Y_j}$ (resp. $\mathrm{P_j}$) in the compartment i. By construction, we have one unique localization per machine.

At given μ, $\mathcal{P}_{\mathrm{rba}}(\mu)$ is a feasibility optimization problem, where constraints (C_1-C_3) define the feasibility domain. The feasibility domain can be empty or non-empty. If there exists a solution $(\mathbf{y}, \nu)$ to $\mathcal{P}_{\mathrm{rba}}(\mu)$ -i.e. the feasibility domain is non-empty-, then there exists a feasible resource distribution compatible with the given growth rate. In other words, the cell can grow at this growth rate value.

Remarks:

1. In practice, the vector $\bar{\mathbf{b}}$ contains non-zero values only for the concentrations of macro-components such as DNA, cell wall, and lipid membranes, and for a few set of metabolites. These values are usually extracted from the biomass formation reaction used in FBA models (see chapter).
2. To model reversible enzymes, we introduced two diagonal matrices containing the enzyme efficiencies, i.e. $\mathbf{K}_{\mathrm{E}}$ and $\mathbf{K}^{'}_{\mathrm{E}}$, describing the capacity constraints of enzymes in both directions. If an enzyme $\mathrm{E_i}$ is considered irreversible, $\mathrm{k}^{'}_{\mathrm{E_i}}$ is set to 0.

3. In [14, 6], an RBA model was built for *Bacillus subtilis*. It integrates two macromolecular processes in constraint C_{2a}, the translation and chaperoning of proteins, and two density constraints, the limitation of the cytosolic density and of the membrane occupancy. An RBA model can be refined by integrating for instance other cellualr processes and molecular machines, such as the transcription machinery, the protein secretion apparatus (see [6, 15]), or molecule turnover [16], as well as other types of constraints. ■

How to incorporate the medium composition. We represent the medium composition by extracellular concentrations, which determine the efficiencies of metabolic transporters via Michaelis-Menten like rate laws (as nonlinear $k(c)$ functions; see section 8.1.2). For an extracellular nutrient $\mathrm{S_{ext,i}}$ with concentration $s_{\mathrm{ext,i}} \geq 0$, the efficiency of the corresponding metabolic transporter(s) is given by $\mathrm{k_E(s_{ext,i})} = \frac{\mathrm{k_{cat} s_{ext,i}}}{\mathrm{K_m + s_{ext,i}}}$, with parameters $\mathrm{k_{cat}}$ and $\mathrm{K_m}$ for the turnover number and the affinity of the transporter, respectively.

Convexity. For given growth rate and medium composition, all equalities and inequalities in our RBA problem $\mathcal{P}_{\mathrm{rba}}(\mu)$ is linear in the decision variables $(\mathbf{y}, \nu)$ and is proven to be convex [17, 18]. By construction, the feasibility domain of $\mathcal{P}_{\mathrm{rba}}(\mu)$ corresponds to the set of all possible phenotypes of the cell at a growth rate $\mu \geq 0$.

Parsimonious resource allocation. The principle of parsimonious resource allocation between cellular processes coincides with principles of optimal performance, for example, that a cell phenotype should maximize growth rate. Importantly, for an RBA problem with given parameters, there exists a maximal growth rate $\mu^* \geq 0$, such that for any μ, $\mathcal{P}_{\mathrm{rba}}(\mu)$ is feasible if and only if $\mu \leq \mu^*$ [17, 18]. For a given medium composition, the maximal growth rate μ^* can computed by using a bisection algorithm, in which a series of LP problems are solved to narrow down the exact growth rate at which the problem becomes infeasible. Together with the maximal feasible growth rate one obtains the optimal cell configuration maximizing growth $(\mu^*, \mathbf{y}^*, \nu^*)$.

Exploration of the feasibility domain. It was proven that the feasibility domain contracts with increasing growth rate [17, 18]. In the same vein as Flux Variability Analysis ([19] and), the feasibility domain can be explored at optimal (μ^*) or sub-optimal $(\mu \leq \mu^*)$ growth rates. For one decision variable y_i (resp. ν_i), two LP problems are solved, where (i) constraints C_1, C_2 and C_3 remain unchanged; (ii) the decision variable y_i (resp. ν_i) is maximized (LP 1) and minimized (LP 2). This operation is repeated for each decision variable to obtain *in fine* the feasibility domain of all decision variables. In practice, at the optimum, the cell configuration $(\mu^*, \mathbf{y}^*, \nu^*)$ is often unique. Indeed, non-unique solutions will exist if two alternative metabolic pathways have exactly the same cost in resources. Since all enzymes have different amino acid sequences, use different cofactors, are differently localization, etc, this is highly unlikely. A caricatural example of a model with non-unique solutions would be one in which an enzyme pool is arbitrarily split into two, and the two new “enzyme species” are given different names, although they are physically exactly the same.

RBA and other frameworks for cell modelling with protein constraints RBA extends and embeds the ideas of proteome partitioning beyond CAFBA [20] and GECKO [21], of cytosolic limitations beyond FBAwMC [22]. The other large-scale resource allocation models are ME-models [9] . Originally, ME-models were considered as an extension of M-models, by including predictions for mRNA, protein, and ribosome levels. Importantly, they do not consider density constraints (C_3). Therefore, limitations on the capacity of exchange fluxes (as in FBA) are necessary to obtain a solution. ME-models are also formalized as LP feasibility problems at fixed growth rate. As in RBA, the parsimonious use of resources is obtained by maximizing growth, i.e. by solving a series of LP problems. . Actually, these methods can be reformalized using RBA constraints.

8.3.2 What do RBA constraints mean?

RBA describes optimal resource allocation in constraint-based, genome-scale whole-cell models and formalizes the mathematical relationships defining the interactions and allocation of resources between the cellular processes. All these relationships take the form of linear, growth-rate dependent equalities and inequalities, and form the convex feasibility problem $\mathcal{P}_{\mathrm{rba}}$ [17, 14, 18]. For cells growing in exponential phase at a growth rate μ, our constraints C_1, C_2 and C_3 formalize different structural constraints:

(I) the metabolic network has to produce all metabolic precursors necessary for biomass production and mass conservation must hold for all intracellular molecule species - i.e. intracellular metabolites and molecular machines - (equalities C_1).

(II) the capacity of each type of molecular machine must be sufficient to ensure its function, i.e. to catalyze chemical conversions at a sufficient rate (inequalities C_{2b} for the metabolic enzymes and transporters, C_{2a} for the molecular machines of macromolecular processes);

(III) the intracellular density of compartments and the occupancy of membranes must not exceed the defined limits (inequalities C_3).

Other physiological constraints can be added, for instance a defined amount of DNA or membrane lipids per cell. They are implemented by setting target values for concentrations and/or fluxes defining a viable cell in a given (or several) environmental conditions, but they are not structural constraints. Intrinsically, the RBA framework captures local objectives (often considered as alternative objective functions in standard CBM methods) as the efficient use of ATP production, or saving enzyme (i.e. the criterion beyond parsimonious FBA [23]) within structural constraints.

Towards more autonomous CBM models. In standard constraint-based modeling methods such as FBA [24], the biomass composition and the cell configuration are fixed *a priori* by the chemical reaction of biomass formation and by the limitation of capacities of cell exchange fluxes (see chapter 5). In RBA, metabolic fluxes (including cell exchange fluxes) and protein levels result from a global compromise in terms of the parsimonious use of resources. When the predicted configuration (Y, ν) is changing, so are the amino acid and nucleotide requirements, and thus so is the biomass composition (see section 8.4 for further details). The final use of extracellular nutrients also results from a compromise in terms of cost (induction of the catabolic pathway) and benefit (the nutrient yields resources) for the cell.

8.3.3 How to build and calibrate RBA models?

For a given organism, an RBA model can include all known metabolic reactions and all (potentially relevant) cell processes with major protein investments (including in particular metabolism, transport, ribosomes, chaperones). Which metabolic reactions and cell processes are regarded as relevant may vary between organisms and is a modeling choice.
RBA integrates the description of the stoichiometry of the metabolic network, the composition of some macromolecular components as DNA, mRNA, lipids, cell wall components, but also a detailed description of molecular machines including the precise protein and nucleic sequences, their dependency on metallic ions, vitamins, cofactors for activity, and their cellular localization.

Requirement for RBA model creation. Several pieces of information, that is, several types of biological data, are needed to build an RBA model for an organism:

- Metabolic network and reaction stoichiometry
- Mapping from reactions to enzymes and further to proteins
- Protein amino acid sequences and cofactors (e.g. metal ions)

- Efficiencies of metabolic enzymes,
- Types of cellular machines (e.g. ribosomes and chaperones), their efficiencies and cofactor demands
- Molecule sizes and localisations (for density constraints)
- Other constraints on concentrations or fluxes ("targets") defining a viable cell

In practice, RBA models can be built and simulated using a software called RBApy [15]. It takes as input (i) a genome-scale metabolic network in SBML format [25] and having a high-quality of mapping between genes and reactions; (ii) the description of molecular machines involved in macromolecular processes (e.g ribosomes); (iii) the NCBI taxon of the organism of interest. Then it automatically extracts information on protein sequences and cofactors, subunit structures, localization from public repositories such as Uniprot, maps enzymes to the reactions they catalyze and to the proteins they consist of, and builds a draft uncalibrated RBA model.

Calibration of RBA model parameters. An RBA model may contain a high number of model parameters. How can we obtain them? Remember that the main parameters to be estimated are the abundance of housekeeping proteins $(\mathbf{p})$, densities $(\bar{\mathbf{d}})$ and efficiencies of molecular machines $(\mathbf{K}_{\mathrm{E}}, \mathbf{K}_{\mathrm{E}}^{'}, \mathbf{K}_{\mathrm{T}})$[1]. As we learned in chapter 3, the rate of an enzymatic reaction, $v = e\, k_{\mathrm{app}}$, depends on the enzyme's efficiency or "apparent catalytic rate", given by $k_{\mathrm{app}} = f(\mathbf{c}) = k_{cat}^{+} \cdot \eta^{\mathrm{rev}}(\mathbf{c}) \cdot \eta^{\mathrm{sat}}(\mathbf{c}) < k_{cat}^{+}$. The k_{app} values are always below the k_{cat} value, but may vary from state to state depending on metabolite concentrations. Since internal metabolite concentrations $\mathbf{c}$ are unknown and difficult to measure at genome-scale, we cannot estimate k_{app} from the explicit kinetic law $f(\mathbf{c})$. We need to obtain these k_{app} parameters empirically, for example by measuring the flux v and the protein abundance e in one condition and taking their ratio. Hence, for a given environmental condition, RBA parameters can be estimated using quantitative proteomics [26] for $\mathbf{p}$ and $\bar{\mathbf{d}}$, and a combination of quantitative proteomics and fluxomics (if available) as in [6] or FBA to estimate the flux distribution as in [15] for efficiency parameters. To account for variable enzyme efficiencies, one may make the simplifying assumption that enzyme efficiencies depend mostly on growth rate. By estimating the enzyme efficiencies at different growth rates and interpolating between them, one obtains empirical relationships between them and the growth rate [6] that can be further used in $\mathcal{P}_{\mathrm{rba}}(\mu)$. For instance, several estimates of enzymatic efficiencies obtained in several contrasted growth conditions will provide a relationship such as $\mathbf{K}_{\mathrm{E}}(\mu)$ instead of a constant $\mathbf{K}_{\mathrm{E}}$ value.

8.4 Prediction of variable biomass composition

Cell models describe, among other things, what a cell is composed of (see chapter 2). In such models, cell composition may either serve as input information (e.g. putting a certain amount of lipids as a constraint) or model results (to be validated by data).Cell composition may be described broadly by molecule classes (e.g. the overall amounts of lipids, protein, DNA, RNA, and so on), or more precisely for individual molecule types (indiviual lipids, proteins, mRNA species, and so on), or even in terms of atomic composition (which in turn gives clues about the amounts of molecule classes). Another measure of biomass composition is the amounts of precursors that are made in metabolism in order to produce macromolecules. In FBA, specifically, "biomass" refers to the proportions of these precursor amounts (including cofactors), assuming that the total production may change, but the proportions stay fixed.

Cell composition may greatly vary between cell types, and also for a given cell type between conditions (for instance, because of the different ribosome content of a cell at different growth rates). This variable composition often poses a challenge for models. Just like the uptake rates, the varying biomass composition reflects complex global rearrangements of resources, and choices between metabolic strategies. In classical FBA, uptake rates and biomass composition need to be predefined; if they are not known from experiments, predictions may not be reliable (see chapter 5).

A main achievement of RBA models is that the biomass composition is not fixed and given as input information as in FBA, but is predicted by the model itself. The model predicts the optimal macromolecule composition of the cell and –

[1] $\bar{\mathbf{b}}$ is deduced from the biomass reaction of the genome-scale metabolic network of the organism

unlike in small self-replicator models (see the Molenaar-type models in chapter **??**) - this prediction is very detailed: it follows from the predicted proteome and may comprise details like trace elements (e.g. iron, which appears in some of the proteins) and whose contribution to the overall biomass will vary with the expression level of those proteins. How does this work? In RBA models, in contrast to FBA, uptake rates and biomass composition are usually not predefined (although this is possible, if required, and may be used to fit models to data), but obtained as model predictions from resource allocation ideas. For example, in RBA, the proportions of biomass precursors would reflect the types and amounts of macromolecules (to be made from these precursors) in the optimal state, which in turn (among other things) reflect the amounts of enzymes needed to make the precursors.

Recommended readings

- Website rba.inrae.fr for further details on RBA

Exercises

Exercise 1 The role of metabolite concentrations

- What if the metabolite content of the cell has been underestimated? Assume that the amount of small metabolites in cells is currently underestimated. What problems in model predictions would arise from the fact? In what way would predictions (by FBA or other methods) be distorted?
- In what way would a cell, in reality, profit from a lower small metabolite content? Can we assume that the ratio between small metabolites and proteins is optimized? Describe possible aspects of this compromise!

Bibliography

[1] M.T. Wortel, H. Peters, J. Hulshof, B. Teusink, and F.J. Bruggeman. Metabolic states with maximal specific rate carry flux through an elementary flux mode. *FEBS Journal*, 281(6):1547–1555, 2014. doi: 10.1111/febs.12722.

[2] Stefan Müller, Georg Regensburger, and Ralf Steuer. Enzyme allocation problems in kinetic metabolic networks: Optimal solutions are elementary flux modes. *Journal of theoretical biology*, 347:182–190, 2014. doi: 10.1016/j.jtbi.2013.11.015.

[3] M Dinh and V Fromion. Rba like problem with thermo-kinetics is non convex. *arXiv preprint arXiv:1706.01312*, 2017. doi: 10.48550/arXiv.1706.01312.

[4] Vicente Acuna, Flavio Chierichetti, Vincent Lacroix, Alberto Marchetti-Spaccamela, Marie-France Sagot, and Leen Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60, 2009. doi: 10.1016/j.biosystems.2008.06.015.

[5] Maxime Mahout, Ross P Carlson, and Sabine Peres. Answer set programming for computing constraints-based elementary flux modes: Application to escherichia coli core metabolism. *Processes*, 8(12):1649, 2020. doi: 10.3390/pr8121649.

[6] Anne Goelzer, Jan Muntel, Victor Chubukov, Matthieu Jules, Eric Prestel, Rolf Nölker, Mahendra Mariadassou, Stéphane Aymerich, Michael Hecker, Philippe Noirot, et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic engineering*, 32:232–243, 2015. doi: 10.1016/j.ymben.2015.10.003.

[7] Laurent Tournier, Anne Goelzer, and Vincent Fromion. Optimal resource allocation enables mathematical exploration of microbial metabolic configurations. *Journal of mathematical biology*, 75(6):1349–1380, 2017. doi: 10.1007/s00285-017-1118-5.

[8] Kobe De Becker, Niccolò Totis, Kristel Bernaerts, and Steffen Waldherr. Using resource constraints derived from genomic and proteomic data in metabolic network models. *Current Opinion in Systems Biology*, 29:100400, 2022. doi: 10.1016/j.coisb.2021.100400.

[9] Edward J O'Brien, Joshua A Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9(1):693, 2013. doi: 10.1038/msb.2013.52.

[10] Pranas Grigaitis and Bas Teusink. An excess of glycolytic enzymes under glucose-limited conditions may enable saccharomyces cerevisiae to adapt to nutrient availability. *FEBS letters*, 2022.

[11] André B. Canelas, Cor Ras, Angela ten Pierick, Walter M. van Gulik, and Joseph J. Heijnen. An in vivo data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. *Metabolic Engineering*, 13(3):294–306, 2011. doi: 10.1016/j.ymben.2011.02.005.

[12] E Martinez-Salas, JA Martin, and M Vicente. Relationship of escherichia coli density to growth rate and cell age. *Journal of bacteriology*, 147(1):97–100, 1981. doi: 10.1128/jb.147.1.97\-100.1981.

[13] Steven B Zimmerman and Stefan O Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli. *Journal of molecular biology*, 222(3):599–620, 1991. doi: 10.1016/0022-2836(91)90499-v.

[14] Anne Goelzer and Vincent Fromion. Bacterial growth rate reflects a bottleneck in resource allocation. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1810(10):978–988, 2011. doi: 10.1016/j.bbagen.2011.05.014.

[15] Ana Bulović, Stephan Fischer, Marc Dinh, Felipe Golib, Wolfram Liebermeister, Christian Poirier, Laurent Tournier, Edda Klipp, Vincent Fromion, and Anne Goelzer. Automated generation of bacterial resource allocation models. *Metabolic engineering*, 55:12–22, 2019. doi: 10.1016/j.ymben.2019.06.001.

[16] Anne Goelzer and Vincent Fromion. Rba for eukaryotic cells: foundations and theoretical developments. *bioRxiv*, page 750182, 2019. doi: 10.1101/750182.

[17] Anne Goelzer, Vincent Fromion, and Gérard Scorletti. Cell design in bacteria as a convex optimization problem controller. In *CDC*, 2009. doi: 10.1016/j.automatica.2011.02.038.

[18] Anne Goelzer, Vincent Fromion, and Gérard Scorletti. Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6):1210–1218, 2011. doi: 10.1016/j.automatica.2011.02.038.

[19] Radhakrishnan Mahadevan and Chrisophe H Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276, 2003. doi: 10.1016/j.ymben.2003.09.002.

[20] Matteo Mori, Terence Hwa, Olivier C Martin, Andrea De Martino, and Enzo Marinari. Constrained allocation flux balance analysis. *PLoS computational biology*, 12(6):e1004913, 2016. doi: 10.1371/journal.pcbi.1004913.

[21] Benjamín J Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*, 13(8):935, 2017. doi: 10.15252/msb.20167411.

[22] Alexei Vazquez, Qasim K Beg, Marcio A DeMenezes, Jason Ernst, Ziv Bar-Joseph, Albert-László Barabási, László G Boros, and Zoltán N Oltvai. Impact of the solvent capacity constraint on e. coli metabolism. *BMC systems biology*, 2(1):1–10, 2008. doi: 10.1186/1752-0509-2-7.

[23] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390, 2010.

[24] Amit Varma and Bernhard O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and environmental microbiology*, 60 (10):3724–3731, 1994. doi: 10.1128/aem.60.10.3724-3731.1994.

[25] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. doi: 10.1093/bioinformatics/btg015.

[26] Mélisande Blein-Nicolas and Michel Zivy. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1864(8):883–895, 2016. doi: 10.1016/j.bbapap.2016.02.019.