Large resource allocation models of cells

Hugo Dourado[1], Anne Goelzer[2], Pranas Grigaitis[3], Wolfram Liebermeister[2], and Elad Noor[4]

[1] Institute for Computer Science and Department of Biology, Heinrich-Heine Universität, Düsseldorf, Germany; [2] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France; [3] Systems Biology Lab, Amsterdam Institute of Molecular and Life Sciences (AIMMS), Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; [4] Department of Plant and Environmental Sciences, Weizmann Institute of Science, 76100 Rehovot, Israel

Abstract

Resource balance analysis models are cell models based on three basic constraints formulated at genome-scale: stationary fluxes (balancing production and consumption fluxes, uptake and excretion fluxes, as well as compound dilution by cell growth); flux coupling constraints relating fluxes to the amounts of catalyzing enzymes (or other machines); and density constraints, limiting molecule amounts in cell compartments, or molecule concentrations. These constraints narrow down the solution space predicted by FBA towards more physiological solutions. Large resource allocation models build on the same principles, and have been implemented as different variations (RBA models, ME-models, and pc-models).

Keywords: Cell model, Resource allocation, Constraint, Biomass

Contributions: This chapter was planned and outlined by the authors after discussions with O. Bodeit, M. Lercher, and R. Steuer, and has been reviewed by S. Waldherr and F. Moro.

To cite this chapter: H. Dourado, A. Goelzer, P. Grigaitis, W. Liebermeister, and E. Noor. Large resource allocation models of cells. Version April 2024. The Economic Cell Collective. Economic Principles in Cell Biology (2024). No commercial publisher | Authors open access book.

The authors are listed in alphabetical order.

Chapter overview

- Resource balance analysis models are cell models based on three basic constraints formulated at genome-scale: stationary fluxes (balancing production and consumption fluxes, uptake and excretion fluxes, as well as compound dilution by cell growth); flux coupling constraints relating fluxes to the amounts of catalyzing enzymes (or other machines); and density constraints, limiting molecule amounts in cell compartments, or molecule concentrations.
- These constraints narrow down the solution space predicted by FBA towards more physiological solutions
- Large resource allocation models build on the same principles, and have been implemented as different variations (RBA models, ME-models, and pc-models).

9.1. Overcoming the limitations of FBA in predicting phenotypes

In the previous chapters, we have discussed two principal approaches to modeling biochemical systems. To keep the number of variables low, but with intention of well-parametrizing the model, one can construct small, coarse-grained models of growing cells (Chapter 8 in [1]). On the contrary, Flux Balance Analysis (FBA) models can accommodate a very large number of variables (Chapter 5 in [1]), making them an excellent choice to model metabolic networks at genome-scale.

The small, coarse-grained models are a very suitable platform for investigation of base principles of life. Likely the best example to illustrate this is the work of Douwe Molenaar and co. [2], where a self-replicator model was used to proposed that the low-yield, or substrate-inefficient ("wasteful") metabolic strategies are adopted as a consequence of these pathways being more efficient in terms of protein use, compared to the high-yield pathways. In other terms, the growth output of the "wasteful" strategy per unit protein is higher than the "efficient" one. Thus we now believe that fermentation of glucose, often called under an umbrella term "overflow metabolism", will take place in many organisms if the substrate in their environments is abundant enough.

However, we know that the chemistry of life is extremely diverse, and even such a familiar concept as fermentation can become complicated. Take three representatives of the tree of life: a bacterium *Escherichia coli*, budding yeast *Saccharomyces cerevisiae*, and mammalian, say, human cells. All three exhibit overflow metabolism - even when enough oxygen is available in the environment - yet the underlying biochemistry tells us that *E. coli* ferments glucose into acetic acid, *S. cerevisiae* - into ethanol, and human cells - into lactate. Bringing more contrasts on the table, there might be extreme differences in a single taxon already: some yeasts, for instance, will never produce ethanol when oxygen is present; some of them have lost the ability to do respiration at all over the course of evolution. This might sound like playing a trivia game, but in many cases, meaningful modeling of complex biological systems requires both taking and making biochemical insight. Therefore, when we aim not only to uncover the underlying principles, but also to learn biochemistry, FBA models have an upper hand.

Yet we already know that the predictions of canonical FBA models are limited to substrate-efficient metabolic states. Continuing with the example of the overflow metabolism, FBA models would predict *E. coli* or *S. cerevisiae* to respire on minimal medium with glucose as the main carbon source – irregardless of the maximal flux of glucose into the cell. Thus the prediction of substrate-inefficient metabolism using FBA over the years used to rely on introducing additional, mainly empirical (e.g. maximal oxygen uptake), constraints onto the system [3]. Moreover, we can impose only linear constraints in FBA models, and this greatly reduces our

options.

Overall, we frequently seek to take the advantageous points of both "schools of modeling", however, this is where we need to start doing compromises. In the ideal world, the self-replicator models from the Chapter 8 in [1] would have to be extended with explicit kinetics and thermodynamic constraints to obtain a detailed cell model. However, the number of variables would increase tremendously, and non-linear optimization is very inefficient already past even small systems. On the contrary, we could try to advance on the FBA-type models by introducing the concepts of protein economy (Chapter 7 in [1]) at genome-scale. Following our best understanding, these, again, would constitute non-linear relationships (e.g. enzyme kinetics), yet large-scale non-linear programming is not a viable option either. Thus simplifications are currently necessary to keep linearity (and convexity) to solve optimization problems for large-scale models.

So can we make large-scale models tractable? If we 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 of metabolites in the growing cell volume, or simply - the growth rate μ of the cell.

9.1.1. Why growth rate?

Under the assumption of the balanced growth, the number of copies of every metabolite in the cell is doubled between two consecutive cell divisions. If metabolites are described by their concentration, dilution by growth can be effectively modeled of every metabolite by a "consuming reaction", 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 critical value $\mu_{\max}$, the maximal possible growth rate for our model. Finding this critical value as well as the corresponding optimal fluxes $\mathbf{v}$ and compound concentrations $\mathbf{c}$ is relatively easy, and can be done by bisection: solving a series of Linear Programming problems (checking for potential solutions $(\mathbf{v}, \mathbf{c})$ for different values of μ).

9.1.2. Replacing complex kinetics by catalytic constraints

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 in [1], 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 in [1].

To linearize the expression for v, while keeping the dependence on e, we need to replace the relationship $k = v/e$ by a fixed number, and so k becomes a model parameter. If the metabolite concentrations were known (experimentally, or from kinetic models under optimality assumptions, see Chapter 6 in [1]), the value of k could be computed. Otherwise, it can also be determined experimentally, by measuring v and e and setting $k = v/e$ [4], which is feasible for a limited number of enzymes, however. 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 linear constraints on the enzyme. If we consider coefficients k and $k^{'}$ to approximate enzyme kinetics in the forward and backward direction, respectively, the flux the enzyme e catalyzes should satisfy $-e\ k^{'} \leq v \leq e\ k$. We set $k^{'} = 0$ for irreversible reactions, and, for simplicity reasons, we usually assume $k = k^{'}$ for reversible reactions, unless kinetic measurements are available that suggest otherwise. This relationship can be formulated as enzyme capacity constraints in order to replace the kinetic rate laws in the FBA model. By writing down such constraints for each enzyme in the model, we can couple the metabolic fluxes with the demand for enzymes, needed to operate these fluxes.

9.1.3. Overview of existing FBA extensions

The linearization approach described above can be successfully used for very large models, making the genome-scale models of resource allocation possible. What we commonly refer to as "resource allocation models" therefore formalize the mathematical relationships defining the interactions and allocation of resources between the cellular processes to describe optimal resource allocation using constraint-based models. All these relationships take the form of linear, growth-rate dependent equalities and inequalities, and, when linearized, form a convex feasibility problem [5, 6, 7].

By itself, the idea of constraining metabolic models to represent limited metabolic capacity of cells is not a new one. There are two ways to approach this budgeting problem: "protein budgeting", where a fixed amount of protein needs to be partitioned in the optimal manner (maximizing growth), and "resource budgeting", where models include both the protein budgeting and the descriptions of demands for protein synthesis. However, "protein budgeting" problems assume that investments in protein production follow the budget, and not vice versa.

Some extensions of FBA account for extra empirical constraints on the total concentration of metabolic enzymes (FBA with molecular crowding, or FBAwMC [8]), or on proteome sectors (Constrained-Allocation FBA, or CAFBA [9]). While these can predict metabolic states more reliably, the empirical constraints come as model assumptions and thus cannot be understood by the models themselves. In these models, the primary assumption is that the cell phenotype is obtained by genetic regulations, and the main goal and utility of genetic regulation can be interpreted as ways of saving resources. Thus in many cases when we predict cell phenotype maximizing growth, we find predictions in good agreement with the experimental observations. Therefore, resource allocation models extend and embed the ideas of proteome partitioning beyond frameworks like CAFBA and GECKO [10], or representing metabolic capacity limitations beyond FBAwMC.

Currently, there are three main implementations of large-scale resource allocation models: Resource Balance Analysis (RBA) [11], Models of Metabolism and Macromolecular Expression (ME-models) [12] and proteome-constrained models (pc-models) [13]. All these implementations are formalized as LP feasibility problems at fixed growth rate. 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 that, for instance, RBA does. Therefore, limitations on the capacity of exchange fluxes (as in FBA) are necessary to obtain a solution.

9.2. Types of constraints in resource allocation models

As indicated above, fine-grained models of resource allocation build on the genome-scale metabolic models (GEMs) to encompass all the reactions that potentially could happen in a metabolic network. The technical advance, when constructing such models, is to impose sets of additional constraints onto GEMs to couple the metabolic fluxes with investment into metabolic pathways (production of enzymes). To the date, different implementations of this concept were proposed to predict optimal resource allocation in different microorganisms [14].

The general description of these constraints in fact is the same as for small, coarse-grained self-replicator models, only the number of individual constraints increases. Moreover, every of the constraints described can be split into a number of constraints, considering only a subset of fluxes in the model (e.g., fluxes taking place in a certain cell compartment). Although the precise formulations vary, resource allocation models build on three principal types of constraints (Figure 9.1):

(1) Mass-conservation constraints
(2) Flux coupling constraints
(3) Compartment capacity, or protein density, 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. They are implemented by setting target values for amounts and/or fluxes defining a viable cell in a given (or several) environmental conditions, but they are not structural constraints. These constraints usually are added *ad hoc* and do not need to bear any functional meaning *per se*. We will now expand on the three types of constraints used in resource allocation models; note that the description is not exhaustive and peculiarities may vary among different formulations.

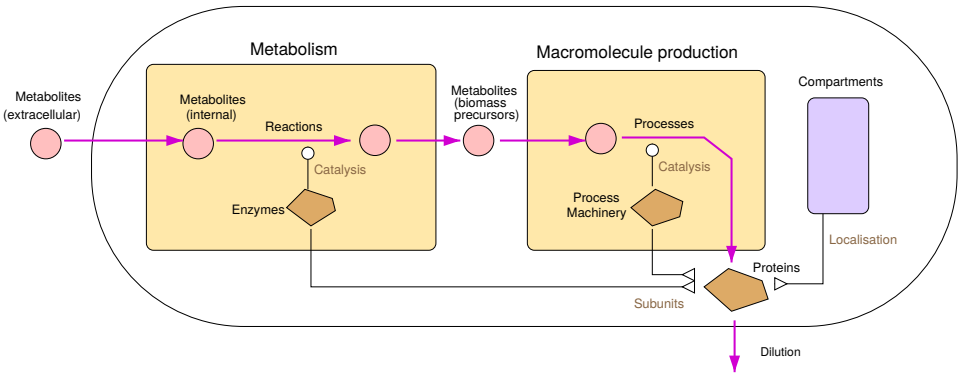
9.2.1. Mass-conservation constraints

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. The reactions which represent either protein folding modeled as the conversion of the "unfolded" protein species into the "folded" ones, and the dilution-by-growth is modeled as a sink for the "folded" protein species ("folded" $\rightarrow \emptyset$).

9.2.2. Flux coupling constraints

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 (in most cases, k_{cat} values), which are used as model parameters. We establish the coupling between fluxes and protein synthesis by setting $v = k_{\mathrm{cat}}\ e\ \eta$, where e is the enzyme concentration and $0 < \eta \leq 1$ is an efficiency term summarizing the effects of reaction thermodynamics, enzyme saturation, and possibly small-molecule regulation. The value for η can be either assumed or fitted from experimental data, and when $\eta = 1$, the enzyme is considered to operate at its maximal rate. Coupling constraints are introduced to couple both (i) the metabolic reactions with enzyme usage (as described above) and (ii) protein turnover

(A)



(B)

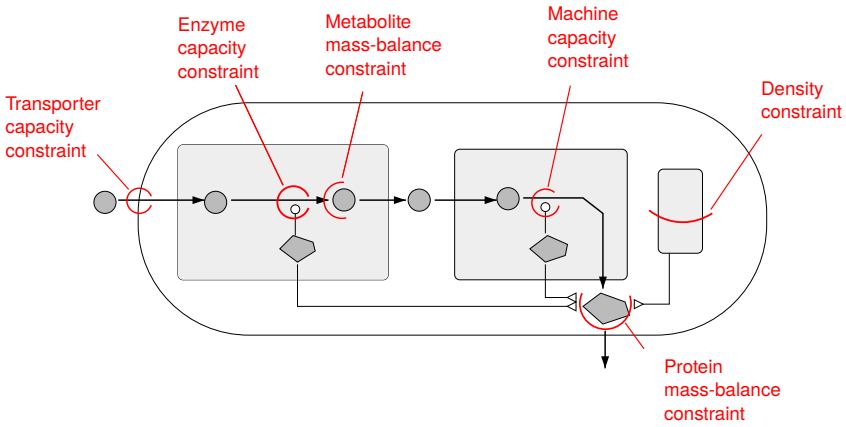


Figure 9.1: Overview of biological components and mathematical constraints in large-scale resource allocation models. Here, Resource Balance Analysis (RBA) model is taken as an example. (A) Typically, an RBA model describes metabolisms and macromolecule production in a growing cell (yellow blocks). Precursors 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) Sets of 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.

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 parameterizing 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 [12]. 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 $v = k_{\mathrm{cat}}\ e\ \eta$), and simplify such factors as enzyme saturation and thermodynamic driving force into a single value of factor η. 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 $k_{\mathrm{eff}} = k_{\mathrm{cat}}\ \alpha$, 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 $\eta = 1$. 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 [15].

9.2.3. Protein density constraints

The final layer of information in the fine-grained resource allocation models is a set of *protein density* constraints. These constraints describe the [upper] limit of cellular process(es), e.g. maximal protein capacity of a compartment. These constraints are formulated as weighted sums of protein abundance, and usual weighing multipliers are proportional to the molecular weight of the protein. Usually, the density constraints are expressed in terms of (usually 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.

9.2.4. Interpreting the consequences of the additional constraints

We have briefly discussed what types of additional constraints need to be implemented to extend FBA models to account for cellular resource allocation, and now let us recap on what these sets of rules mean in biological terms. The constraints described above shall couple the metabolic fluxes with the production of enzymes that operate these functions, so the model has to produce amino acids and generate ATP in order to use them for protein translation. Moreover, the enzyme demand will be coupled with the production of the macromolecular machines required to produce, fold, and degrade these enzymes (ribosomes, chaperones, and proteases, respectively), requiring the same building blocks (see Chapter 2 in [1]). These constraints therefore formalize a self-replicating molecular system in balanced growth subject to different structural constraints:

1. 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.
2. 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;
3. the intracellular density of compartments and the occupancy of membranes must not exceed the defined limits.

As highlighted before, the biological interpretation of the additional constraints discussed above is rather

Box 9.A : 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: $v_{\mathrm{synthesis,e}} = v_{\mathrm{degradation,e}} + v_{\mathrm{dilution,e}}$, or, rewritten with the respective parameters, $v_{\mathrm{synthesis,e}} = (k_{\mathrm{deg,e}} + \mu)\ e$. Here, $k_{\mathrm{deg,e}}$ 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*, [16]). 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). *E. coli* maintains rather constant protein content in dry weight across growth rates (ca. 0.55 $(g$ protein$)\ gDW^{-1}$) [17, 18]. On the contrary, the protein content is known to vary in *S. cerevisiae* as a function of growth rate [16].

universal for different implementations of resource allocation models, with minor deviations in terminology and/or formulation. To illustrate how resource allocation models are built from conventional GEMs, and how the respective models are formalized in mathematical terms, in the following we will consider one of the popular formulations of resource allocation models in more depth.

9.3. Resource Balance Analysis (RBA) models

Resource Balance Analysis (RBA) has been developed as (and is considered to be) a flexible and generic modeling framework which describes the functioning of an organism using the most relevant set of linear equality and inequality constraints, described in general terms in Section 9.2. As a consequence, an RBA model includes all known metabolic reactions coupled to relevant cell processes with major protein investments (production of biomass precursors; including, but not limited to protein translation, protein folding, protein transmembrane transport, and protein degradation). Where applicable, circumstantial information can be included into the model to establish the dependency of enzyme activity on metal ions, vitamins, and/or cofactors. Which metabolic reactions and cell processes are regarded as relevant may vary between organisms and is a modeler's choice.

9.3.1. Building a draft RBA model

The software package `RBApy` [19] contains all the routines needed to build and simulate RBA models. In order to build a new RBA model, it takes a genome-scale metabolic network in SBML format [20] as an input, together with additional information to formulate the additional constraints described in the previous section. Different types of biological data, are needed to build an RBA model for an organism:

- Amino acid sequences for metabolic enzymes and macromolecular machines (e.g. ribosomes and chaperones),
- If applicable, stoichiometry of known cofactors (e.g. metal ions),

- Efficiencies of metabolic enzymes,
- Molecular weights and localization of proteins (for density constraints),
- Any empirical constraints on concentrations or fluxes ("targets", see previous section).

The software routine first extends the input GEM to include the description of protein turnover in the cell. The software extracts information from the input files on (i) protein sequences and cofactors, (ii) the subunit stoichiometry of protein complexes, and (iii) protein localization (using information from public databases such as UniProt). Using this information, reactions corresponding to protein synthesis, folding, degradation, and dilution by growth are added. Finally, the software maps enzymes to the reactions they catalyze and to the proteins they consist of, and the output of the routine is a draft (uncalibrated) RBA model.

9.3.2. Mathematical description of a RBA problem

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 (see Chapter **??** in [1]), 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:

1. 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 $\mathbf{m} \triangleq (m_1,\ldots,m_{n_m})^T$;
2. 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 $\mathbb{P}$;
3. 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 $\mathcal{P}_{\mathrm{rba}}(\mu)$) 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 $\mu \geq 0$,

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

where all the inequalities are defined component-wise and:

- $\mathbf{\Omega}$ 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.

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 [5, 7]. 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. 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 $\mu \geq 0$.

We conclude this with some 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 5 in [1]).
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 [6, 21], 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 cellular processes and molecular machines, such as the transcription machinery, the protein secretion apparatus (see [21, 19]), or molecule turnover [22], as well as other types of constraints.

9.3.3. Simulation and analysis of RBA models

How to incorporate the medium composition. We represent the medium composition in two aspects, namely (i) qualitatively, by allowing exchange of the medium metabolites in the model ($UB_{Exchange,n} > 0$). Note that some metabolites, although not explicitly represented by the growth media, should also adhere to this rule (e.g. oxygen, water, and protons). The (ii) quantitative composition of the growth medium is determined by extracellular concentrations, which, in turn, dictate the efficiencies of metabolic transporters via Michaelis-Menten-like rate laws (as nonlinear $k(c)$ functions; see section 9.1.1). 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.

Calibration of model parameters. An RBA model may contain a high number of model parameters. First, the global parameters to be estimated are related to cell composition: (i) the concentrations of bulk biomass components $\bar{\mathbf{b}}$, which is usually deduced from the biomass reaction of the genome-scale metabolic network of the organism. Using quantitative proteomics data [23], one can infer (ii) the protein densities in different compartments ($\bar{\mathbf{d}}$), and (iii) the abundance of housekeeping (unspecified) proteins ($\mathbf{p}$).

The next set of parameters we need to collect concerns the efficiencies of molecular machines ($\mathbf{K_E}$, $\mathbf{K_E'}$,$\mathbf{K_T}$). As we learned in Chapters 2 in [1] and 3 in [1], the rate of an enzymatic reaction v depends on the enzyme's efficiency or "apparent catalytic rate", given by $v = e\, k_{\mathrm{app}}$, with $k_{\mathrm{app}} = f(\mathbf{c}) = k_{\mathrm{cat}}^{+} \cdot \eta^{\mathrm{rev}}(\mathbf{c}) \cdot \eta^{\mathrm{sat}}(\mathbf{c}) < k_{\mathrm{cat}}^{+}$. 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, efficiency parameters can be estimated using quantitative proteomics in combination with fluxomics [21] or FBA to estimate the flux distribution [19]. 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 efficiency and the growth rate [21] to be used in $\mathcal{P}_{\mathrm{rba}}(\mu)$. For instance, several estimates of enzymatic efficiencies obtained in contrasting growth conditions will provide a relationship $\mathbf{K_E}(\mu)$ instead of a constant $\mathbf{K_E}$ value.

Obtaining the RBA solution for a given parameter set. 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 $\mu \leq \mu^*$ [5, 7]. 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. A real-life example would be simulating growth in glucose-limited chemostat cultures under different dilution rates D. With increasing D, the glucose availability increases, and a set of n different glucose uptake rates q_{Glc} ($q_{Glc,1}$, $q_{Glc,2}$, …, ${}_qGlc, n$) can be subjected to an RBA model to obtain a set of optimal metabolic states (μ_1^*, μ_2^*, …, μ_n^*).

Together with the maximal feasible growth rate one obtains the optimal cell configuration maximizing growth $(\mu^*, \mathbf{y}^*, \nu^*)$. The principle of optimal performance, in this case, that a cell phenotype should maximize growth rate, in fact, coincides with the principle of parsimonious resource allocation between cellular processes.

Exploration of the feasibility domain. Although RBA models inherently reduce the solution space due to principle of parsimonious resource allocation, the solutions obtained might still contain considerable flux

variability. In the same vein as Flux Variability Analysis ([24], see Chapter 5 in [1]), 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.

It was proven that the feasibility domain becomes smaller with increasing growth rate [5, 7], so it might be worthwhile to probe the solution space at slow-growth regimes. 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.

9.3.4. Use of -omics data-informed k_{app} *vs.* naïve k_{cat} values

The three most popular formalisms of fine-grained resource allocation models, RBA [11], ME-models [12], and pc-models [13], are variations on the same theme, as shown in the general discussion of the underlying constraints in Section 9.2. Thus most of the ideas, concepts, and constraints are equivalent (or at least highly similar) in their biological interpretation. Most of the differences arise from the approach taken towards parametrization of these models, and consequently, interpretation of model output. Here we will discuss an example where implementations differ significantly.

In resource allocation models, two types of constraints define the proteome capacity at given growth rate μ, the protein density vector $\bar{\mathbf{b}}$, and the fraction of housekeeping proteins $\mathbf{p}$ in the proteome. The remaining proteome space is to be distributed among the proteins that are explicitly defined in the model. The RBA formalism requires to formulate the function $k_{\text{app}}(\mu)$ (or $\mathbf{K}_{\text{E}}(\mu)$ in the RBA problem, Section 9.3.2) for every protein in the model using -omics data (see Section 9.3.3), and the fraction of the "housekeeping" proteins in the proteome is determined from data for each simulation.

Conversely, the formulation of pc-models [13] allows more flexibility to the "unspecified" protein UP, represented by a single artificial protein of average size and amino acid composition. Instead of setting a fixed amount allocated to $\mathbf{p}$ which changes across conditions, one can determine the *minimal* fraction of this protein in proteome UP_{min}, and formulate the demand to produce UP as an inequality constraint $UP \geq UP_{min}$. Interestingly, in *Saccharomyces cerevisiae*, the proteome mass fraction occupied by non-metabolic proteins is relatively constant under different glucose-limited conditions, as determined by quantitative proteomics data (see [13], Fig. S1 for a plot).

This inequality constraint can be interpreted as the upper limit of available protein space, i.e., under fixed protein density $\mathbf{y} + \mathbf{p} = \text{const.}$, the proteome not occupied by $\mathbf{y} \stackrel{\Delta}{=} \mathbf{e} + \mathbf{m}$ is allocated to $\mathbf{p}$. Since now the model can distribute the proteome among explicitly-defined $vs.$ unspecified protein freely, the procedure of fitting k_{app} values is no longer a prerequisite. Using k_{cat} values, collected from literature/databases/own experimental measurements, rather than apparent k_{app} values, has consequences both for predictions and the data use: first, the model prediction on the protein use is the "demand" of the enzyme and is strictly coupled to the flux through the enzyme (equivalent to the ECM1 layer of enzyme costs in the *Enzyme Cost Minimization* method, Chapter 6 in [1]). Second, the condition-dependent quantitative proteomics data can be used as validation dataset for model predictions instead [25], as the predicted protein abundance is not dependent on these datasets.

Using less data for parameter fitting and redirecting these data-rich datasets towards validation of model prediction strengthens the argument for using resource allocation models for learning new biology, and already has real-life examples. For instance, the discrepancies in predicted $vs.$ observed levels of glycolytic enzymes at glucose-scarce conditions in [13] inspired the same team to revisit the question whether the high levels of glycolytic enzymes represent the optimal expression given very low thermodynamic driving force and undersaturation of glycolytic enzymes. Comparing predictions of *Enzyme Cost Minimization* models with the results of the pc-model and experimental data, [15] proposed that *S. cerevisiae* expresses genuine excess of glycolytic enzymes in glucose-limited conditions, meant to amply consume any glucose as soon as it appears in the environment.

9.4. Biomass composition: both a constraint and a prediction

Cell models describe, among other things, what a cell is composed of (see Chapter 2 in [1]). In FBA, specifically, biomass refers to the proportions of different molecule classes (e.g. lipids, protein, DNA, RNA, cofactors) in 1 gram dry weight of cells, and biomass composition needs to be defined prior to optimization. Since, at least for FBA models of microbes, biomass production usually is the optimization objective, the literature frequently refers to the mathematical description of cell composition as "biomass objective function" (BOF). In most cases, it is assumed that the proportions of biomass constituents are fixed, only the total production (flux through BOF) changes.

For the predictions of FBA models to be reliable, a high-quality BOF is a must (see Chapter 5 in [1]). Therefore, there is a sustained effort to experimental determination biomass composition, even for *E. coli* [26]; for more details on the usual experimental measurement methods, see the box in Chapter 2 in [1]. In case supporting data are available, the cell composition in the BOF may be described in a more fine-grained manner for individual molecule types (e.g., individual lipids, proteins, mRNA species, etc), or even in terms of atomic composition (which in turn gives clues about the amounts of molecule classes). So, overall, the biomass composition acts as a global, and one of the most stringent, constraint on the predicted solution space in FBA-based models.

However, cell composition may greatly vary not only between (micro-)organisms, or different cell types within the same organism, but also for a the same organism/cell type across different conditions. Budding yeast *S. cerevisiae*, for instance, exhibits rather linear relationships between the proportions of bulk biomass constituents as a function of growth rate in glucose-limited cultures [27]. This variable composition often poses a challenge for models: just like the uptake rates, the varying biomass composition reflects complex global rearrangements of resources (for instance, different ribosome content at different growth rates [28] leads to changes in RNA-to-protein ratio in the cells), and choices between metabolic strategies (e.g. depletion of storage carbohydrates in glucose-fermenting *S. cerevisiae* [16]).

A main advance of resource allocation models, compared to conventional FBA models, is that only a part of the biomass composition is given as input information just like in FBA ($\bar{\mathbf{b}}$ in RBA). The proteome composition, on the contrary, becomes a genuine prediction of the optimization procedure. Unlike small self-replicator models (see the models in Chapter 8 in [1]), this prediction is very detailed, as the the predicted proteome composition is represented by the sum of individual protein abundances. Moreover, if proteins require trace elements or cofactors (e.g. iron in iron-containing proteins) for function, the demand and contribution to the overall biomass of these metabolites will also be predicted by the model (as it will vary with the expression level of those proteins).

In theory, the abundance of biomass constituents other than proteome could be formulated in the way they

become predictions of the resource allocation models, rather than hardcoded inputs. Following the idea implemented in the small, coarse-grained models of [2], one could set relationships between, e.g., protein density in the cells and production of lipids (in [2], the biological interpretation was to maintain the surface area-to-volume ratio constant). Currently this is not widely accepted as a standard practice, and, as we can see from the example above, requires comprehensive experimental evidence, which, by itself, could be interpreted still as "input to the model".

9.5. Concluding remarks

In this chapter, we have considered the resource allocation models, the extensions of FBA models which couple metabolic networks with the macromolecular machinery that is required to operate them. These models append existing FBA models with a large number of additional reactions, metabolites, constraints, and model parameters, and, overall, offer a fine-grained representation of cellular economy. Many of kinetic parameters cannot be accurately measured for individual enzymes, and/or are condition-dependent. The *quantitative* nature of the predictions of resource allocation (and the most cellular decisions/phenotype shifts), however, are largely governed by global constraints: for instance, when the protein density $g\ gDW^{-1}$ in a compartment reaches its upper limit (=that compartment is fully packed with protein), the cells switch from fully-respiratory to respiro-fermentative growth (see [12] for *E. coli*, or [13] for *S. cerevisiae*). Unlike the kinetic parameters, which are rather uncertain, these "global" constraints are based on more trustworthy evidence.

Thus these models still retain a reasonable compromise concerning numerical tractability and model complexity, and can accurately predict complex adaptations, which cannot be captured by GEMs in an autonomous way, i.e. without the addition of empirical constraints on fluxes. A successful use case of using resource allocation models is dissecting iron economy, using RBA models: some proteins require iron for their function, and the cell growth can become iron-limited in some conditions. The RBA model was used to predict cell behavior under iron starvation, and the predictions suggested couple of scenarios, (i) the cell may *increase* the import of iron, but also (ii) *avoid* using proteins that contain iron (and the pathways in which they operate) [21, 29].

As with the biomass composition, another aspect of resource allocation models (and FBA-based models in general) with some duality in its interpretation is the objective function. Although its validity has been always debated since conception, maximization of instantaneous growth rate as the optimization objective has shown incredible success in predicting microbial physiology. The current approach we apply for resource allocation models still remains the FBA-based assumption that the desired cell phenotypes are the ones maximizing instantaneous growth rate μ. This time, however, the μ is also a model variable, so we have to apply bisection to obtain *the* optimal solution for each parameter set we use in resource allocation models.

It is becoming more and more evident that a lot of phenotypes (and microbial species!) we try to predict divert from the principle maximization of instantaneous growth rate. For instance, the most experimental microbial physiology research has been focused on carbon-limited (C-limited) cultures, especially the yeast work in Delft, the Netherlands (see [27, 16] for examples). It seems that the maximization of growth rate work very well in C-limited case, and the success of resource allocation models to quantitatively capture these phenotypes [12, 21, 13] affirms this assumption. But is C-limitation descriptive of natural environments? Let us continue the argument with yeasts as an example.

Yeasts in the wild, for instance, very frequently are subjected to feast-famine cycles in terms of carbon availability, and one could argue that these yeasts should act as glucose-limited in the famine phase of the

cycle. Yet the current opinion in the yeast ecology seems to see feast-famine cycles as a continuous, although reduced, supply of carbon, and steer towards embracing a higher role of nitrogen (N) limitation in natural environments instead. Currently, our understanding of N-limited growth is not very comprehensive, and N-limitation is also a case where the instantaneous growth rate maximization breaks down: the pc-models of *S. cerevisiae* cannot quantitatively capture the cell behavior under N-limited conditions (Pranas Grigaitis, *unpublished*).

So the selection of a suitable optimization objective can be a choice followed by huge success, but also, the optimal solution might end up contradicting the existing knowledge. How can we try to mitigate that? One huge advance of resource allocation models is that at any condition, the available solution space is greatly reduced, compared to conventional FBA. We can argue that we have introduced a whole new set, a whole new type of constraints into the model by accepting assumptions stemming from the metabolism-molecular machinery coupling. In theory, we should be able to reason further regarding any additional (even empirical/*ad hoc*) constraints and/or additional optimization objectives which would bring our model predictions closer to observed biology. Just remember: fitting models is not a sin; but nontransparent/reckless fitting is! After all, modeling is an art, and there is no one cookbook that represents the ground truth: we should be free to explore the secrets of biology, as unrealistic as our assumptions are at times.

A final remark on modeling being an art. In this book, we have explored several types of cell models of different size, detail, and assumptions behind. This whole hierarchy and diversity of different implementations and formalisms might seem overcomplicated and unnecessary, although it is a mere reflection that "one size does *not* fit all". In the following chapters we shall continue discussing further model types, and we invite (future) modelers to be creative, mix, match, and tailor different models (and modeling types) to advance biology. The compromise between fine-grained but liner modeling $vs.$ complex kinetics that materialized into resource allocation models is an inspiring example of how one can push bounds of different methods.

Recommended readings

RBA website Website `rba.inrae.fr` for further details on RBA. Under *Tools*, there are example models and Jupyter notebooks for running them.

Review article on large-scale resource allocation models K. de Becker *et al.* "Using resource constraints derived from genomic and proteomic data in metabolic network models" *Curr Opin Syst Biol* 2022, 29:100400

Problems

Problem 9.1 The role of metabolite concentrations The available cell space for proteins depends on the assumed space occupied by small metabolites.

1. 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?
2. 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! For inspiration, see [30].

Bibliography

[1] The Economic Cell Collective, editor. *Economic Principles in Cell Biology*. Free online book, 2023. doi: 10.5281/zenodo.8156386.

[2] Douwe Molenaar, Rogier van Berlo, Dick de Ridder, and Bas Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular Systems Biology*, 5(1):323, 2009. doi: https://doi.org/10.1038/msb.2009.82.

[3] Daan H De Groot, Julia Lischke, Riccardo Muolo, Robert Planqué, Frank J Bruggeman, and Bas Teusink. The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints. *Cellular and Molecular Life Sciences*, 77:441–453, 2020. doi: 10.1007/s00018-019-03380-2.

[4] Dan Davidi, Elad Noor, Wolfram Liebermeister, Arren Bar-Even, Avi Flamholz, Katja Tummler, Uri Barenholz, Miki Goldenfeld, Tomer Shlomi, and Ron Milo. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k_{cat} measurements. *Proceedings of the National Academy of Sciences*, 113(12):3401–3406, 2016. doi: 10.1073/pnas.1514240113.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Ibrahim E Elsemman, Angelica Rodriguez Prado, Pranas Grigaitis, Manuel Garcia Albornoz, Victoria Harman, Stephen W Holman, Johan van Heerden, Frank J Bruggeman, Mark MM Bisschops, Nikolaus Sonnenschein, et al. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies. *Nature Communications*, 13(1):1–12, 2022. doi: 10.1038/s41467-022-28467-6.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] Laurence Yang, James T Yurkovich, Colton J Lloyd, Ali Ebrahim, Michael A Saunders, and Bernhard O Palsson. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Scientific reports*, 6(1):36734, 2016. doi: 10.1038/srep36734.

[26] Vetle Simensen, Christian Schulz, Emil Karlsen, Signe Bråtelund, Idun Burgos, Lilja Brekke Thorfinnsdottir, Laura García-Calvo, Per Bruheim, and Eivind Almaas. Experimental determination of Escherichia coli biomass composition for constraint-based metabolic modeling. *Plos one*, 17(1):e0262450, 2022. doi: 10.1371/journal.pone.0262450.

[27] H. C. Lange and J. J. Heijnen. Statistical reconciliation of the elemental and molecular biomass composition of Saccharomyces cerevisiae. *Biotechnology and Bioengineering*, 75(3):334–344, 2001. doi: 10.1002/bit.10054.

[28] Eyal Metzl-Raz, Moshe Kafri, Gilad Yaakov, Ilya Soifer, Yonat Gurvich, and Naama Barkai. Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *eLife*, 6, aug 2017. doi: 10.7554/eLife.28034.

[29] 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.

[30] Naama Tepper, Elad Noor, Daniel Amador-Noguez, Hulda S Haraldsdóttir, Ron Milo, Josh Rabinowitz, Wolfram Liebermeister, and Tomer Shlomi. Steady-State metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. *PLoS One*, 8(9):e75370, September 2013. doi: 10.1371/journal.pone.0075370.