Automated generation of resource allocation models at cellular scales

from prokaryotic to eukaryotic cells

Anne Goelzer

anne.goelzer@inrae.fr

Applied Mathematics and Computer Science,
from Genomes to the Environment
INRA Jouy-en-Josas, FRANCE
Resource allocation as a strong design principle in living organisms

Organisms

Models based on resource allocation

Biomass allocation by source-sink empirical relations

Nutrient partitioning model between life stages

High prediction capability despite their « simplicity »

Resource allocation between organs and life functions is a strong structural constraint

At cellular scale?
Growth Rate of *Escherichia coli*

ALLEN G. MARR

*Department of Microbiology, University of California, Davis, California 95616*

**Ribosomes**

**Non-ribosomal proteins**
A new perspective since 2009
How to predict resource allocation at genome scale?

**PERSPECTIVE (2009)**

**Shifts in growth strategies reflect tradeoffs in cellular economics**

Douwe Molenaar, Rogier van Berlo, Dick de Ridder, and Bas Teusink

**Interdependence of Cell Growth and Gene Expression: Origins and Consequences (2010)**

W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, Terence Hwa

**2009**

**Convex and genome-scale**

Cell design in bacteria as a convex optimization problem

Anne Goelzer, Vincent Fromion and Gérard Scorletti
Three (main) structural constraints

Resources (especially proteins) have to be shared by all biological processes (implicit feedback).
Resource sharing imposes constraints on cellular processes.
Detailed integration of production costs for protein synthesis

Transport
De novo synthesis by metabolic network

Amino acids

Ribosomes + tRNAs

ATP, GTP

Polypeptide

Chaperones

Ions, vitamins

Protein

(Active) molecular machine
→ capable of operate its function
e.g. capable of catalyzing the flux
Formalization into an optimization problem

Resource Balance Analysis (RBA)

For fixed $P_G \geq 0, \mu \geq 0$,

Find $R \geq 0, C \geq 0, \nu^x \in \mathcal{R}^m$, $|\nu_i| \leq k_{E_i} E_i$

subject to

$(C_{1a})$ For all $i \in I_p$,
$$- \sum_{j=1}^{m} S_{p_{ij}} \nu^x_j + \mu \left( \sum_{j=1}^{m} C_{M_{ij}}^M \nu^x_j + C_{R_i}^M R + C_{C_i}^M C + C_{G_i}^M P_{G_i}^x T \right) - \nu_Y = 0$$

$(C_{1b})$ For all $i \in I_c$,
$$- \sum_{j=1}^{m} S_{c_{ij}} \nu^x_j + \mu \bar{X}_{c_i} = 0$$

$(C_{1c})$ For all $i \in I_r$,
$$\sum_{j=1}^{m} S_{r_{ij}} \nu^x_j + \mu \left( \sum_{j=1}^{m} C_{M_{ij}}^R \nu^x_j + C_{R_i}^R R + C_{C_i}^R C + C_{G_i}^R P_{G_i}^x T \right) = 0$$

$(C_{1d})$ For all $i \in I_i$,
$$\sum_{j=1}^{m} S_{i_{ij}} \nu^x_j = 0$$

$(C_{2a})$ \( \mu(\sum_{j=1}^{m} C_{M_j}^R \nu^x_j + C_{R_j}^R R + C_{C_j}^R C + C_{G_j}^R P_{G_j}^x T) - k_T R = 0 \)

$(C_{2b})$ \( \alpha_c \mu(\sum_{j=1}^{m} C_{M_j}^C \nu^x_j + C_{R_j}^C R + C_{C_j}^C C + C_{G_j}^C P_{G_j}^x T) - k_C C = 0 \)

$(C_{3a})$ \( \sum_{j=1}^{m} C_{M_j}^D \nu^c_j + C_{R_j}^D R + C_{C_j}^D C + C_{G_j}^D P_{G_j}^c T - \bar{D}_c \leq 0 \)

$(C_{3b})$ \( \sum_{j=1}^{m} C_{M_j}^S \nu^s_j + C_{G_j}^S P_{G_j}^s T - \bar{D}_s \leq 0 \)


The RBA framework

- The feasibility problem is convex
- Equivalence with a Linear Programming (LP) optimization problem
  
  same complexity as FBA, efficient resolution at genome scale!

- For a set of **given extracellular nutrient concentrations**, we can prove that there exists a **maximal growth rate value**
  
  • **without setting an objective function** (contrary to FBA);
  • **defined by a trade-off on the resource allocation** (especially on proteins);
  • for which a resource distribution (enzyme/ribosomes) exists;
  • and can be efficiently computed through the iterative resolution of LP optimization problems;

- Every mechanism saving resources increases the growth rate

- Theoretical prediction of induced/repressed sub-systems in the metabolic network (towards the prediction of genetic regulations)

  RBA computes (a) the maximal **growth rate**, (b) the **metabolic fluxes** including the substrate uptake and by-product secretion rates, and (c) the genome-scale resource allocation including the **absolute abundances of enzymes, transporters, ribosomes, and chaperones**, i.e. the phenotype of the organism
Rewriting the RBA problem in a more compact way

For fixed $P_G \geq 0$, $\mu \geq 0$,

\begin{align*}
\text{find} & \quad Y \in \mathbb{R}^{m+p}_{\geq 0}, \nu \in \mathbb{R}^m, \\
\text{subject to} & \quad -\Omega \nu + \mu(C^S_Y Y + C^S_B \bar{B} + C^S_G P_G) = 0 \\
(C_1) & \quad \mu(C^M_Y Y + C^M_G P_G) - K_T Y \leq 0 \\
(C_{2a}) & \quad -K'_E Y \leq \nu \leq K_E Y \\
(C_{2b}) & \quad C^D_Y Y + C^D_G P_G - \bar{D} \leq 0
\end{align*}
For fixed $P_G \geq 0$, $\mu \geq 0$, find

$$Y \in \mathbb{R}_{\geq 0}^{m+p}, \nu \in \mathbb{R}^m,$$

subject to

$$(C_1) \quad -\Omega \nu + \mu (C^S_Y Y + C^S_B \bar{B} + C^S_G P_G) = 0$$
$$\mu (C^M_Y Y + C^M_G P_G) - K_T Y \leq 0$$
$$-K^r_E Y \leq \nu \leq K_E Y$$
$$C^D_Y Y + C^D_G P_G - \bar{D} \leq 0$$

From the stoichiometry of chemical reactions

From annotation & bioinformatics
752 parameters to be estimated

For fixed $P_G \geq 0, \mu \geq 0$, 

\[
\begin{align*}
\text{find} & \quad Y \in \mathbb{R}_\geq 0^{m+p+m}, \nu \in \mathbb{R}^m, \\
\text{subject to} & \\
(C_1) & \quad -\Omega\nu + \mu(C^S_Y Y + C^F_B \bar{B} + C^S_G P_G) = 0 \\
(C_{2a}) & \quad \mu(C^M_Y Y + C^M_G P_G) - K_T Y \leq 0 \\
(C_{2b}) & \quad -K_E Y \leq \nu \leq K_E Y \\
(C_3) & \quad C^D_Y Y + C^D_G P_G - \bar{D} \leq 0 
\end{align*}
\]

Q.: quantitative  
RA.: relative/absolute

Proteomics (Q., RA.)  
Fluxomics  
From literature or biomass comp or data.

Total protein content + proteomics + protein localization
Identification of apparent catalytic rate of $\approx 600$ enzymes (Consistency with the expected distribution)

Quantitative prediction of the resource allocation between 72 cellular processes
Quantitative prediction of the resource allocation between 72 cellular processes.

Metabolic Engineering

Quantitative prediction of genome-wide resource allocation in bacteria

Anne Goelzer, Jan Muntele, Victor Chubukov, Matthieu Jules, Eric Prestet, Rolf Nöiker, Mahendra Mariadassou, Stéphane Aymerich, Michael Hecker, Philippe Noirot, Dörte Becher, Vincent Fromion

Institutes: INRA, UREAP, Malzeve, F-28350 Jouy-en-Josas, France; Institute for Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland; Institute for Microbiology, F-92350 Jouy-en-Josas, France; Aachen University, Aachen, Germany.

R² = 0.823

Graph showing the relationship between measured growth rate (1/h) and predicted protein cost (MDa).
Generation of RBA models for prokaryotes
What do we need to create a RBA model?

A metabolic network including the description of **enzymatic complexes**

- Literature or automatic reconstruction

- **Uniprot**
  - Information on protein sequences
  - Amino acid sequences, Localization ions, cofactors, vitamins, structure, …

Description of non-metabolic molecular machines

- Ribosomes, chaperones, etc.

- **Literature, Databases**

Optimal resource allocation

- Growth rate
  - Metabolic fluxes, protein content

- **Parameters**
  - Efficiencies of molecular machines, \( P_G \) amount, etc.
RBApyp: a software system for bacterial resource allocation models

Inputs (mandatory):
- A genome-scale metabolic model (GSMM) including gene association (i.e. boolean AND/OR)
- The NCBI Taxon ID

Additional inputs (for model refinement)
- Definition of molecular machines,
- Composition of rRNAs, tRNAs

Additional inputs (for model calibration)
- Quantitative (or relative-absolute) proteomics
- Fluxomics (or prediction of metabolic fluxes)

Outputs:
- A RBA model in XML files
- Simulation results in text files
Validation of RBAPy on *Bacillus subtilis*

(A) Model created from SBML and Uniprot files with default parameters and default processes, where automatic merging for some identifiers failed (e.g. tRNAs IDs)

(B) After matching SBML metabolite identifiers with Uniprot cofactor identifiers, and RBAPy identifiers for metabolites involved in processes

(C) After calibration of molecular machine efficiencies, adjustment of subunit stoichiometries of enzyme complexes and molecular machines from the hand-curated model, and adding metabolic demands for flagella movement and membrane biosynthesis.
A RBA model for *Escherichia coli* created from scratch

Source of information:
- The iJO1366 metabolic model [1]
- Literature, Uniprot

Estimation of parameters:
- Apparent catalytic rates $k_E$, efficiencies of molecular machines in glucose minimal medium (data from [2]+[3])
- Total protein abundance per compartment wrt growth rate, etc. based on [2]

A RBA model for *Escherichia coli*

Source of information:

- The iJO1366 metabolic model [1]

Flux visualization using Escher maps
Protein visualization using Proteomaps

Need additional information for the use of Proteomaps and Escher maps

- BIGG identifiers (Escher [1])
- Functional annotation (Proteomaps [2])

A versatile modeling framework

A new cellular process can be included straightforward by adding new capability constraints and new decision variables in RBA

- Necessitate to detail the production cost
- Introduce new parameters to identify

A new compartment can be included straightforward by adding new density constraints

- Introduce new parameters to identify

Possible currently by editing the XML files manually of RBApy

Future RBApy releases will contain tools to facilitate the integration of other processes and compartments
Towards RBA models for eukaryotic cells
How to integrate

\[
\sum_{k} C_k = C_T \\
\sum_{k} N_k = N_T
\]

Mesophyll cells

Guard cells

Transport De novo synthesis by metabolic network

Amino acids

Ribosomes

Polypeptide

ATP, GTP

Chaperones

Active protein

Transport

Amino acids

Peptoglycan

Cofactors

Others

Membrane

Cell wall synthesis

Nucleotides

Lipids

Replication

Transcription

DNA

RNA

Molecular processes

Cell components

Energy and Reducing power

Anabolism

Precursors

NH3
Formalization into a LP optimization problem

\[ P^{e}_{\text{rba}}(\mu): \text{For a fixed vector of concentrations } P_G \in \mathbb{R}^{N_g}_{\geq 0}, \text{ and the growth rate } \mu \geq 0 \text{ of the cell, find} \]

\[
\begin{align*}
    Y \in \mathbb{R}^{N_y}_{\geq 0}, & \quad \nu \in \mathbb{R}^{N_m}, & \quad f \in \mathbb{R}^{N_c}_{\geq 0}, \\
    -\Omega \nu + \mu(C^S_Y Y + C^S_G P_G + C^S_B \bar{B} + C^S_F f \hat{B}) = 0 & \quad \mu(C^M_Y Y + C^M_G P_G) - K_T Y \leq 0 & \quad -K'_E Y \leq \nu \leq K_E Y \\
    C^D,^{iq}_Y Y + C^D,^{iq}_G P_G - C^D,^{iq}_F f \leq 0 & \quad C^D,^{eq}_Y Y + C^D,^{eq}_G P_G - C^D,^{eq}_F f = 0 & \quad C^F f - \bar{C} = 0 \\
    f_V \leq I_V f \leq \bar{f}_V 
\end{align*}
\]

- For fixed growth rate the optimization problem is a LP problem
- There exists a maximal growth rate \( \mu^* \) such as \( Prba(\mu) \) is feasible for lower \( \mu \) values, and unfeasible for upper values
- The optimal \( \mu^* \) can be computed by a bisection algorithm by solving a series of LP.
Formalization into a LP optimization problem

\( P_{rba}^e(\mu) \): For a fixed vector of the cell, the problem is a LP optimization problem.

For fixed growth rate the optimization problem is a LP problem.

There exists a maximal growth rate \( \mu^* \) such as \( P_{rba}^e(\mu) \) is feasible for lower \( \mu \) values, and unfeasible for upper \( \mu \) values.

The optimal \( \mu^* \) can be computed by a bisection algorithm by solving a series of LP.

RBA for eukaryotic cells: foundations and theoretical developments

Anne Goelzer, Vincent Fromion

INRA, UR1404, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France.

firstname.name@inra.fr

Abstract

Resource allocation models were recently identified as new ways to investigate cell design principles. In particular, the Resource Balance Analysis (RBA) framework is the first constraint-based modeling method capable of accurate quantitative predictions of the genome-wide resource allocation. Initially developed and validated on bacteria, the objective of this paper is to provide the mathematical foundations of the extension of the RBA framework to eukaryotic cells. We especially investigate the way to handle the cellular compartments in order to formalize eventually the functioning of organelles. It turns out that the final RBA problem for eukaryotic cells is close to the one of prokaryotic cells from a theoretical point of view. The mathematical properties that were already identified on the prokaryotic RBA framework can be easily transposed to eukaryotic cells. In particular, the eukaryotic RBA problem can be solved easily at the cell scale by Linear Programming.

This paves the way to future developments of RBA models for eukaryotic cells.
What are the parameters?

- $K_E$, $K'_E$: the apparent catalytic rates of enzymes and transporters.
  
  → Literature/database or Total protein content, proteomics, fluxomics

- $K_T$: the efficiency of macromolecular machines as ribosomes.
  
  → Literature or Total protein content, proteomics, Ribosome abundance

- $P_G$: Abundance of proteins (per compartment) for which the activity is not explicitly described in the model.
  
  → Total protein content, proteomics, protein localization

- $\bar{B}$: Abundance of macrocomponents of biomass as total DNA, mRNA, cell wall, Lipids, Starch, free AA, etc.
  
  → Biomass composition

- $\bar{C}$: Link between compartments and membranes of organelles as the surface/volume ratio.
  
  → Literature

RBA models for prokaryotic and eukaryotic cells are highly similar!
Only a few changes to RBApyp are necessary

- Improve the protein localization management

- Include additional cellular processes in compartments (e.g. translation process in mitochondrion and in cytoplasm) by default

- Implement the additional constraints related to compartment management

- Tools to facilitate the manual curation

A proof-of-concept on the leaf of *Arabidopsis thaliana (under progress)*
RBA for other eukaryotic cells (1/2)

« Straightforward » in theory, but ....

Problem in information description within GSMM ...
- Standards such as SBML need to evolve to account for molecular machine descriptions and template-based chemical reactions
- Need to unify identifiers of molecules, reactions, cellular functions across regna (plant, mammals, microorganisms)

... and beyond
- Molecular machines (including enzymatic complexes) are poorly described in databases (but recent progress like Reactome)
- Use of ontologies to formally describe the organism as a « system » (i.e. composed of molecular machines dedicated to a cellular function)
- Transfer of the knowledge from a model organism to another one (genericity vs specificity)

→ Strong link with the knowledge representation field in biology
RBA for other eukaryotic cells (2/2)

Heterogeneity in organism description

Mammalian cells
- Active community to build a curated Human GSMM including gene association
- High degree of homology between mammals
- Automatic reconstruction based on orthologues search for common metabolic pathways
- Use of transcriptomics or proteomics to specialize GSMM per organ

→ Mouse GSMM generated from this procedure

Availability of data for RBA parameters identification?

Plant cells
- GSMMs usually do not include the description of enzymatic complexes (AND/OR rules)
- Difficulty to link genes IDs in GSMM and Uniprot
  → Lack of a unified standard between databases
- GSMMs need to be specialized by organs and by developmental stages
- Localization of isoforms could be improved by transcriptomics or proteomics

→ Strong link with the bioinformatics community to infer sequence-based information such as protein localization, structure of the molecular machine complex, etc.

→ Strong link with the biostatistician community to build (for instance) specialized GSMM from omics data
Conclusion and perspectives
the RBA framework

- Extension of the RBA theoretical framework (under progress and/or validation)
  - To dynamical conditions (dRBA)
  - To stochastic fluctuations in gene expression
  - To include thermodynamics and kinetics constraints to predict the metabolite abundances
  - To predicts the emergence of regulatory networks
  - To the chemostat

- RBA for bioengineering to aid strain design

- Automatic generation of RBA models (RBAPy)
  - Extension to eukaryotic cells
  - Integration of facilities for model manipulation, adaptation and visualization
  - Integration of additional methods of simulation (such as dRBA)

- Resource allocation for other prokaryotes and multi-cellular organisms (under progress and/or validation)
  - Microorganisms: Escherichia coli, Streptomyces coelicolor, Ralstonia solanacearum, Synechocystis sp PCC6803, yeast
  - Plant: Arabidopsis thaliana, Zea mays
Acknowledgments

INRA-MaIAGE

V. Fromion
L. Tournier, S. Fischer, M. Dinh, W. Liebermeister
S. Fischer
M. Mariadassou, P. Nicolas

Ecole Centrale de Lyon
G. Scorletti

INRA-Micalis

M. Jules, S. Aymerich (Grignon)
P. Noirot, E. Prestel, E. Dervyn (Jouy)

The BaSynthec partners

J. Muntel R. Nölker, D. Becher M. Hecker and U. Mader (Greifswald),
V. Chubukov and U. Sauer (ETHZ)