The Endogenics Simulation Engine

Overview of the mathematical modeling

Clément Surville, BioFi AG

July 2025

1 Introduction

Today's understanding of biology is based on decades of research in the lab, but also from different forms of modeling. The amount of data extracted from omics is exponentially increasing since the last decades and the advent of single-cell omics opens a new area of available data about biological systems and cell functions. These data allow to construct knowledge bases and knowledge graphs that are if not exhaustive, very large in terms of coverage of the biological pathways. The new challenge is to extract meaningful and tractable information from these bases to tackle problems in different fields: bioproduction, disease modeling, personalized medicine, etc...

Modelization is the future to overcome this challenge, as it provides reproducibility, predictive and dynamic results like phenotypes, disease trajectories, and is much cheaper and scalable than wet lab experiments. Despite such advantages, biological modeling faces different issues:

- *data integration*: to be biologically meaningful, models must be fed by real data. Omics data like metabolomics, proteomics, genomics, transcriptomics, etc... but also incorporate physical measurements of the living biological system (bioreactor conditions, body, cell/organoïds...)
- *networks interactions*: the different knowledge graphs like metabolic, gene regulatory, signaling networks must be integrated in a unified way and allow for multidimensional interactions, i.e. not considered independently.
- *nonlinear dynamics*: biology is highly dynamic and despite some form of homeostasis necessary for stability of life, these systems can evolve very rapidly to perturbations and in a non linear way,
- *stochasticity and quantum nature*: biological systems consist of microscopic elements that interact almost at the atomic scale. Randomness like mutations, collisional probabilities, stochasticity are ubiquitous and inherent of the biological processes.

At BioFi, we aim at providing a platform and a series of tools that bring modelling to the next level, where all these fundamental issues are taken into account.

In this white paper, we present some key concepts of our Endogenics Simulation Engine (ESE) that illustrate how we model from bottom up biological systems.

2 Master equation

In many systems governed by probabilistic and stochastic nature, like in biology, one can construct an equation representing the change (temporal) of the states as:

$$\frac{d}{dt}\vec{s} = \mathbf{A}(t)\vec{s}, \tag{1}$$

where $\vec{s}$ is the vector of states, and $\mathbf{A}(t)$ is a matrix of rates, or frequencies, and represents the probability of change. As an example, fluid dynamics and thermodynamics are based on such formalism (Boltzmann equation) and it represents accurately the collisional nature of fluids.

This system can be highly non linear if the matrix $\mathbf{A}(t)$ varies quickly or depends strongly on the states. It can also describe irreversible jumps between states, making such a formalism perfect for biological modelling. We will see that many mechanisms and processes fall into this master equation and we will provide in the end a brief overview of how to solve it numerically. For example, metabolic reactions, gene transcription, signal transduction, membrane transport, can be described mathematically in a way that allow to construct a master equation.

Our simulation platform and the ESE is based on constructing and evolving such master equation and we are sure that it highly contributes to the superior fidelity of our models.

3 Metabolic networks

Metabolic networks are central in understanding the biochemistry and the functions of the cell, and our knowledge has grown very quickly during the last decades. They couple metabolites (proteins, chemicals) through a set of reactions representing different processes: chemical interactions, enzymatic reactions, transport through membranes, diffusion, biomass production... Very large networks with more than 10^3 to 10^4 reactions and metabolites are available today, e.g. Recon, KEGG, HumanCyc, but they are not easily tractable.

Two principal approaches are used, usually in combination, to reduce and simplify these large networks in order to use them for predictions or for mechanistic understanding of diseases:

- *dimensionality reduction*: select the most relevant pathways (reactions and metabolites) for the problem at hand. In general, it involves several methods based on chemistry, thermodynamics, biology, domain specific knowledge; it can be curated manually or using data-driven methods like AI trained models.

- *temporality reduction*: reduce the computational cost of evolving the metabolic network. Solving the dynamical evolution of the amounts of metabolites can be expensive (it was a few decades ago, but it is less an issue today) and requires specific numerical methods. The majority of the analyses and study results of metabolic networks are obtained from quasi-steady assumptions, or network based analysis like mode decomposition.

At BioFi, we believe that to understand biology, and not only describe it, the problem must be tractable in size, but be really dynamic, like biology is. We thus use dimensionality reduction, but we search for accurate kinetics and *dynamical fidelity* of our models.

We use available metabolic models, provided by the user or in public data bases (e.g. Reactome), and apply standard reduction methods if necessary to minimize the network size, e.g. COBRA methods. We also construct models from the ground up to small or medium sized models, i.e. of the order of 10 to 10^2 reactions and metabolites.

3.1 Metabolic networks theory

The set of metabolites involved in the metabolic network model and the set of reactions (or equations) that defines their relations and evolution can be written as a system of equations. Suppose one has n metabolites indexed as M_i, with $i \in [1, n]$, and one has m reactions in total, indexed with $j \in [1, m]$. A basic example of such a reaction is in the form:

$$\alpha_1 M_1 + \alpha_2 M_2 \longrightarrow \alpha_3 M_3 + \alpha_4 M_4. \tag{2}$$

The α_i are the stoichiometric coefficients of the metabolites in this particular reaction. The evolution of the concentrations of the metabolites $[M_i]$ is provided by the differential form of this reaction:

$$\begin{aligned}
\frac{d}{dt}[M_1] &= -\alpha_1 v_j, &(3)\\
\frac{d}{dt}[M_2] &= -\alpha_2 v_j, &(4)\\
\frac{d}{dt}[M_3] &= \alpha_3 v_j, &(5)\\
\frac{d}{dt}[M_4] &= \alpha_4 v_j, &(6)
\end{aligned}$$

where v_j is the rate of this j-th reaction, the so called *flux of the reaction*. For a directed reaction like this, $v_j \geq 0$. To generalize, if we give $\beta_{i,j} = -\alpha_{i,j}$ for the reactants of the j-th reaction, $\beta_{i,j} = \alpha_{i,j}$ for its products, and $\beta_{i,j} = 0$ otherwise, then the differential form becomes a system of n equations:

$$\begin{aligned}
\frac{d}{dt}[M_1] &= \beta_{1,j} v_j,\\
&\vdots\\
\frac{d}{dt}[M_i] &= \beta_{i,j} v_j,\\
&\vdots\\
\frac{d}{dt}[M_n] &= \beta_{n,j} v_j.
\end{aligned}$$

As a result, the set of the m reactions provides m similar systems, of which the contributions to the variation of the concentrations sum up, giving the final system:

$$\begin{aligned}
\frac{d}{dt}[M_1] &= \beta_{1,1} v_1 + ... + \beta_{1,j} v_j + ... + \beta_{1,m} v_m,\\
&\vdots\\
\frac{d}{dt}[M_i] &= \beta_{i,1} v_1 + ... + \beta_{i,j} v_j + ... + \beta_{i,m} v_m,\\
&\vdots\\
\frac{d}{dt}[M_n] &= \beta_{n,1} v_1 + ... + \beta_{n,j} v_j + ... + \beta_{n,m} v_m.
\end{aligned}$$

If $[\vec{M}]$ is the n-vector of concentrations, $\vec{v}$ the m-vector of the reaction fluxes, and $S_{n,m}$ the $n \times m$ matrix of the generalized stoichiometric coefficients (the βs), we obtain the compact matrix form:

$$\frac{d}{dt}[\vec{M}] = \mathbf{S_{n,m}}\ \vec{v}, \tag{7}$$

This is the base of the classical analysis of metabolic networks. One can obtain interesting information about the system by exploring quasi-steady solutions, i.e. *when the concentrations of metabolites stabilize*,

$$\frac{d}{dt}[\vec{M}] \sim 0 = \mathbf{S_{n,m}}\ \vec{v}. \tag{8}$$

The family of methods that follow this assumption is the *flux balance analysis*, or FBA, which try to calculate the optimum fluxes $\vec{v}$ solution of the above system that maximizes objective functions under other constrains. Typical objective functions are biomass production, or ATP consumption. This approach falls under the general context of linear constrain based optimization.

At BioFi, we go much further by searching the dynamics of these networks, and to do so, we need to answer: What are the value of the reaction rates? Can we parametrize them? What is the uncertainty?

3.2 Law of mass action and Michaelis-Menten

The collisional and quantum nature of biochemistry are respected by parametrizing the reaction rates as function of the number of reacting molecules, and in particular in the continuous form

$$v \propto [M] \tag{9}$$

This fundamental property leads to a different reaction regimes: *first order* when $v = k_1[M_i]$, *second order* when $v = k_2[M_i][M_j]$. In general, a reaction with two reactants is second order.

As an example, we can express the kinetic law of enzymatic reactions using this law of mass action. It is the basics of the Michaelis-Menten model, where S is a substrate, E the enzyme, C an enzymatic complex formed during the the reaction, and P the product of the overall reaction:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_{cat}} P + E. \tag{10}$$

The constants k represent the kinetic constants of proportionality of the mass action law. One can show after some calculations that the rate of the overall reaction is approximately:

$$v_{MM} = \frac{k_{cat}[E][S]}{K_m + [S]}, \tag{11}$$

with $K_M = (k_{cat}+k_{-1})/k_1$, the Michaelis-Menten constant. This has the dimension of a concentration; it can be called a cutoff concentration, as it is when $[S]$ is well above it that the reaction rate saturates.

3.3 Integration to master equation

If we come back to the above example Equation 2, and now the rate of the reaction is expressed as function of the reactant's concentration (e.g. second order law), we can convert the reaction system

to:

$$\frac{d}{dt}[M_1] = -\alpha_1 k_2 [M_2][M_1], \tag{12}$$

$$\frac{d}{dt}[M_2] = -\alpha_2 k_2 [M_1][M_2], \tag{13}$$

$$\frac{d}{dt}[M_3] = \alpha_3 k_2 [M_2][M_1], \tag{14}$$

$$\frac{d}{dt}[M_4] = \alpha_4 k_2 [M_2][M_1]. \tag{15}$$

It turn out to be in the form

$$\frac{d}{dt}[M_1] = -\omega_1 [M_1], \tag{16}$$

$$\frac{d}{dt}[M_2] = -\omega_2 [M_2], \tag{17}$$

$$\frac{d}{dt}[M_3] = \alpha_3(\omega_1/\alpha_1[M_1] + \omega_2/\alpha_2[M_2])/2, \tag{18}$$

$$\frac{d}{dt}[M_4] = \alpha_4(\omega_1/\alpha_1[M_1] + \omega_2/\alpha_2[M_2])/2. \tag{19}$$

which is the base to construct a system of reactions in the form of

$$\frac{d}{dt}[\vec{M}] = \mathbf{A}(t)[\vec{M}]. \tag{20}$$

We can easily implement complex kinetic laws in the same manner, and have a framework for metabolic networks that respect the first principles of biology. This method also provides very good non linear coupling between reactions, because the instant contributions of each reaction in the network are added in the matrix-vector product. But our approach is not restricted to metabolic networks.

4 Gene regulation networks

Gene regulatory networks relate the expression of genes to the activation of promotion or inhibition transcription factors (TF). These proteins bind to the genetic code and facilitate the transcription or reduce the transcription. Several aspects are responsible for the action of transcription factors: number of active sites, turnover rate, clumping, DNA conformation, RNA transcription, ... In the end, if we map the concentration of TFs and the one of the gene product (the produced protein), some trends can be drawn: (i) at low level of TF, low production (ii) at high level of TF, saturated production, (iii) an intermediate cutoff concentration make the switch between the two regimes.

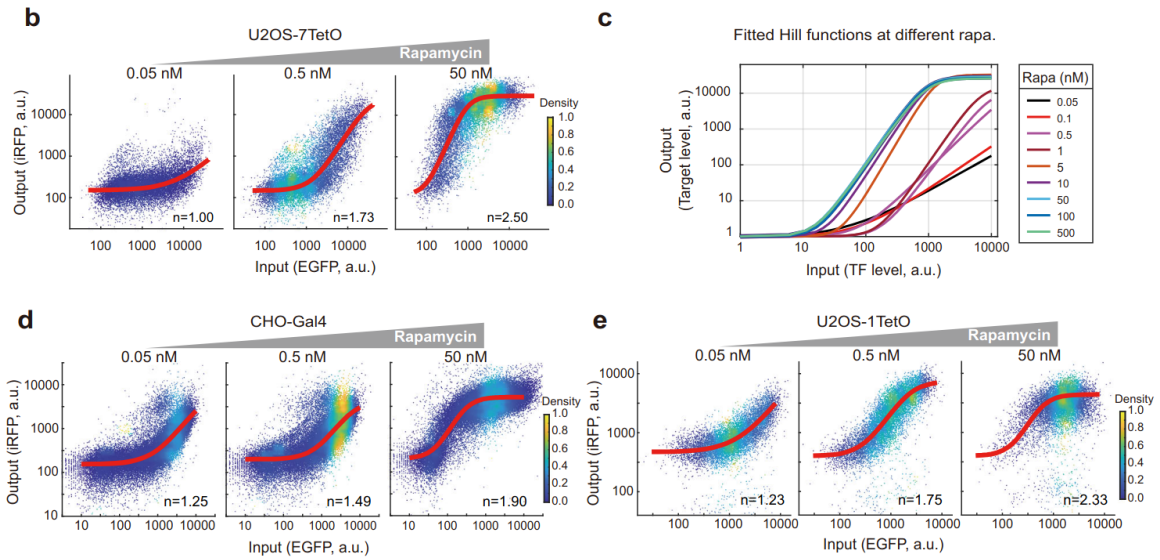
These results are well known since the 1900's and the work of Hill on O_2 binding on hemoglobin and are still under research. For example, the effect of clustering is investigated in Wu, et al. 2022. "Modulating Gene Regulation Function by Chemically Controlled Transcription Factor Clustering." Nature Communications 13 (1): 2663. In their figure 4, they show measurements of levels of TFs in experiments.

The mathematical function that provides such a variation curve is given by the Hill's function.

$$H(x) = \frac{x^n}{X_c^n + x^n}, \text{ with } n \geq 1. \tag{21}$$

The model of gene transcription is thus given by the following rate law:

$$\frac{d}{dt}[P] = \pm k \frac{[TF]^n}{X_c^n + [TF]^n} \tag{22}$$



g. 4 TF clustering propensity modulates the gene regulation functions of synthetic TFs. a Assay design. **b-c**, TF clustering propensity modulates the ene regulation function of the U2OS-7TetO system. Steady-state TF (EGFP) and reporter (iRFP) signals (i.e., inputs and outputs) were quantified by flow

Figure 1: Gene transcription curves obtained experimentally in Wu, et al. 2022. (Fig 4)

where the gene product's concentration $[P]$ increases $(+)$ if TF is a promoter, or decreases $(-)$ if TF is an inhibitor. This transcription rate is governed by collisional and binding effects, that are highly related to concentrations and that are well described by our master equation. As a result, we express the gene transcription regulation by:

$$\frac{d}{dt}[P] = \pm k \frac{[TF]^{n-1}}{X_c^n + [TF]^n}[TF] = \pm\omega[TF]. \tag{23}$$

A network of gene regulation will help to construct a system of regulations that can be expressed as a matrix-vector product in the same form as the master equation. In this context, the master equation has the additional advantage to make easy setting additional constrains. One important is that the total effect of the regulations on the production of a protein (or a gene product) must be positive. i.e. if the inhibition is larger than promotion, the transcription is not active. Such a condition is equivalent to:

$$\frac{d}{dt}[P_i] = \sum_j A_{i,j}[P_j] \geq 0, \tag{24}$$

for each component $[P_i]$ of the gene regulatory network, and $A_{i,j}$ being the elements of the matrix $\mathbf{A}$ than contains the aggregated rates of transcription.

This framework makes our ESE ready to handle complex gene networks and resolve them dynamically with a high degree of non linearity. Moreover, we can couple them directly with metabolic networks as we use the same description as a master equation. This is crucial as many gene products cascade down to other transcription factors via metabolic reactions, as thus feedback loops between the two networks exist. Moreover, many other biological pathway fall into this framework, in particular signaling pathways, which are perfectly suited as ligand binding is at the source of gene transcription models. Thus signaling networks can be accurately incorporated and coupled in our models.

5 Conclusions

We have constructed a simulation framework which consist of: (i) principles based on data, biology, physics and statistics, (ii) a mathematical modeling of networks that respects these principles, (iii) a numerical implementation in our Endogenics Simulation Engine (ESE).

While the details of this implementation are our technological property and will not be disclosed, we can generally say that the master equation framework is adequate to construct different solvers for the time integration. We have designed a fast and robust solver, which can handle large integration time steps. It is perfectly suited for optimization strategies where a lot of models must be run. We also provide a more accurate high order solver that is slightly more restrictive, but provides the best precision once the model is adapted to the phenotype or the target state.

By combining data integration from knowledge, biological fidelity in the ESE modelling, numerical optimization and convergence, we at BioFi can provide a robust platform that assists biologists and researchers in better understanding biology, planning experiments, design and validate future medicine.