SysBio 2018

Innsbruck, Austria

Elad Noor

March 2, 2018

Contents

1 Blackboard session: Metabolic Engineering

1.1 Introduction

1.1.1 Linear Programming (LP)

Every Linear Programming problem can be converted to a canonical form that is:

Primal

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & A\boldsymbol{x} \leq \boldsymbol{b} \\ \text{and} \quad & \boldsymbol{x} \geq \boldsymbol{0}\,. \end{aligned} \tag{1}$$

The dual is a symmetrical linear problem described by the following constraints

Dual

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{subject to} \quad & A^\top \boldsymbol{y} \geq \boldsymbol{c} \\ \text{and} \quad & \boldsymbol{y} \geq 0\,. \end{aligned} \tag{2}$$

If the primal has an optimal solution ($\boldsymbol{x}^*$), then the dual will also have an optimal solution ($\boldsymbol{y}^*$) and their values will be equal:

Strong duality theorem

$$\boldsymbol{c}^\top \boldsymbol{x}^* = \boldsymbol{b}^\top \boldsymbol{y}^* \tag{3}$$

Linear Programming problems can be solved very efficiently using established algorithms such as Simplex. Many good solvers exist for LP, including open source options such as glpk, CLP, and the internal R-project solver.

1.1.2 Mixed-Integer Linear Programming (MILP)

If a LP contains variables that can only have integer values (or in a more specific case, boolean values), the problem cannot be solved using the standard methods. In fate, MILP problems in general are NP-complete, i.e. computationally hard. The free solvers are typically not good enough for solving MILPs. Nevertheless, advanced commercial solvers such as CPLEX and Gurobi are quite efficient in solving even large MILPs in reasonable time, and provide free licenses for academics.

1.2 Convex analysis of a metabolic flux cone

The basic idea underlying most metabolic constraint-based models, is mass-balance combined with the steady-state assumption. Formally, given a stoichiometric matrix S, where rows represent metabolites and column represent reactions, and given a flux vector $\boldsymbol{v}$, the rate of change in metabolite levels ($\boldsymbol{c}$) is given by the equation.

Dynamic mass-balance

$$\frac{d\boldsymbol{c}}{dt} = S\boldsymbol{v} \tag{4}$$

For the subset of rows in S that represent internal metabolites (e.g. intracellular metabolites) which is denoted S_{int}, the pseudo steady-state assumption states that their concentrations do not change over time.

Pseudo steady-state

$$S_{\mathsf{int}}\boldsymbol{v} = 0 \tag{5}$$

Note that external metabolites are not constrained in this way. For example, the concentration of extracellular glucose in the medium decreases gradually while cells are growing at steady-state. Therefore, pseudo steady-state flux solutions are contained within the null-space (or kernel) of S_{int}, i.e. $\boldsymbol{v} \in \ker(S_{\mathsf{int}})$.

1.2.1 The flux cone

Typically, many of the reactions in our model will be considered irreversible, due to thermodynamic constraints (and we will elaborate on this topic in section 1.3). Therefore, a subset of the fluxes will be constrained to be positive:

$$\forall i \in I_{irr} \quad v_i \geq 0 \tag{6}$$

Therefore, the set of possible flux solutions would be all fluxes that satisfy both equation 5 and 6. Explicitly, it would be the intersection of all the hyperplanes that represent rows in S_{int} described by $S_{j*} \cdot \boldsymbol{v} = 0$ and all half-spaces corresponding to irreversible reactions $\boldsymbol{e_i} \cdot \boldsymbol{v} \geq 0$ (where $\boldsymbol{e_i}$ is the unit vector corresponding to reaction i). This intersection is an unbound convex polyhedron denoted the steady-state *flux cone* or $\mathcal{C}$ [9].

It is important to note, that the flux cone is unbounded in some directions, i.e. it can extend to infinity. Therefore, we typically ignore the absolute values of the fluxes (represented, for example, by $||\boldsymbol{v}||$) and consider only the relative values – i.e. the direction towards which the vector is pointing. It is often easier to only consider normalized vectors, such as ones whose biomass rate is set to 1.

Flux cone

$$\mathcal{C} = \{\boldsymbol{v} \in \mathbb{R}^n \mid S_{\mathsf{int}}\boldsymbol{v} = 0 \ \wedge \ v_i \geq 0 \ \ \forall i \in I_{irr}\} \tag{7}$$

1.2.2 Elementary Flux Modes

Elementary Flux Modes (EFMs) are defined as the set of all non-decomposable vectors in the flux cone. A non-decomposable vector is one whose *support* (the set of reactions with non-zero flux) is minimal – i.e. the flux cone does not contain any vectors whose support is a proper subset of it. EFMs also form a convex basis for the flux cone [17]:

$$\forall \boldsymbol{v} \in \mathcal{C} : \ \exists \lambda_j \geq 0 \ \text{ s.t. } \ \boldsymbol{v} = \sum_j \lambda_j \ \boldsymbol{e}^j \tag{8}$$

where $\{\boldsymbol{e}^j\}_j$ is the set of all EFMs. Figure 1 illustrates a model with 10 reactions and 9 EFMs.

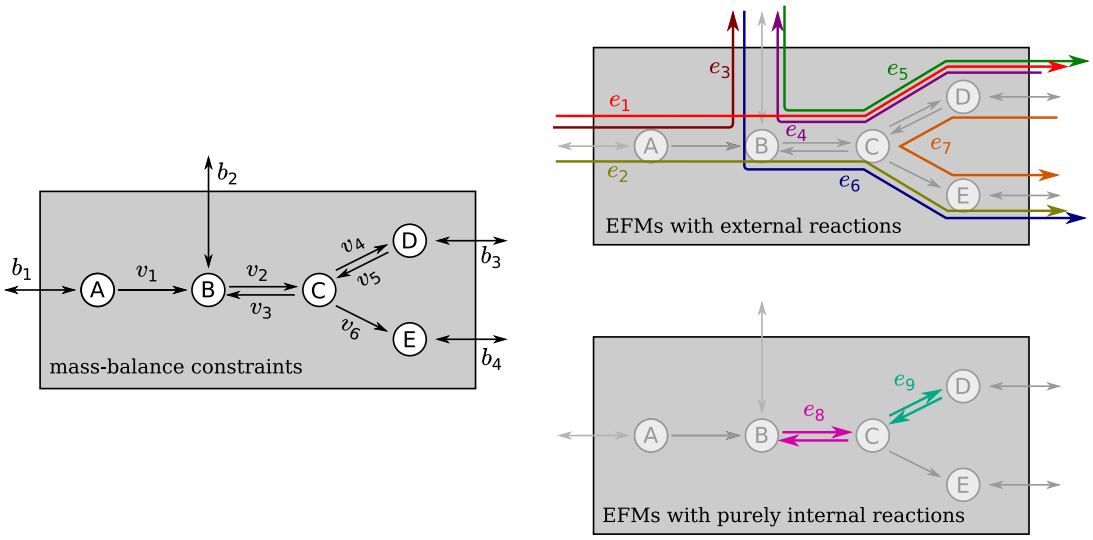

Figure 1: A toy example with 6 internal reactions and 4 exchange reactions. The flux cone is spanned by 9 EFMs.

EFMs are closely related to Extreme Pathways (EPs) and distinguishing them requires more subtle definitions. For further reading you can refer to [7].

1.2.3 Classification of EFMs

Even in our small toy example in Figure 1, we already encounter a common problem that arises when dealing with EFMs – i.e. cycles. In this example, it's quite obvious that the two EFMs that consist of combining two opposing reactions is futile and should be ignored. Larger networks, however, might have much larger cycles that are harder to identify and get rid of.

First, we need to distinguish between *primary exchange* and *currency exchange* reactions. The former are the standard reactions that exchange nutrients between the cell/compartment and its environment (sugar import, CO_2 export, etc.). *Currency exchange* reactions do not, in fact, exchange material between compartments, but are rather abstract representations of energy dissipation. A canonical example would be the ATP maintenance reaction that is present in most metabolic models: ATP + H_2O ⇀ ADP + P_i.

Now, we can define the three types of EFMs ([15], see Figure 2):

I – *Primary systemic EFMs* have at least one active primary exchange flux.

II – *Futile cycles* have no active primary exchange fluxes, and at least one active currency exchange flux.

III – *Internal cycles* have no active exchange fluxes (neither primary nor currency).

To illustrate the three types of pathways, we present another toy model with one currency exchange reaction and two primary exchange reactions (Figure 3). This model contains all the three types of EFMs.

In the next section, we will see how this three type classification is useful to separate thermodynamically feasible pathways from infeasible ones.

1.3 Thermodynamics

A fundamental principle in enzymatic catalysis is the notion that enzymes can accelerate the rate of a reaction, but not change its thermodynamic equilibrium, therefore they have no effect on the *direction* of flux.

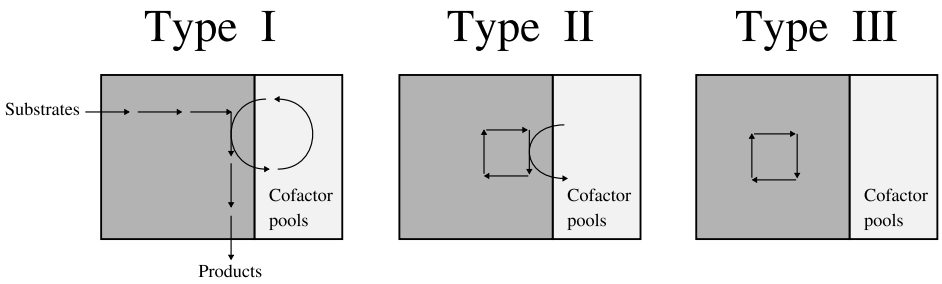

Figure 2: The three types of Elementary Flux Modes.

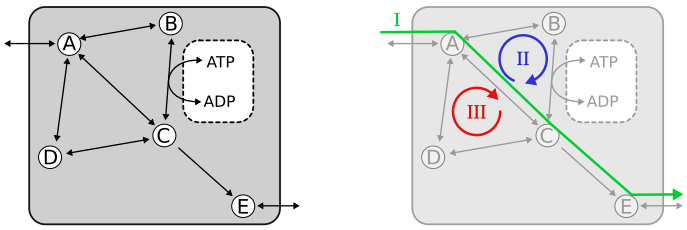

Figure 3: A small toy model with 8 internal reactions, 2 primary exchange reactions, and one currency exchange reaction. In this simple network, one can identify all three types of extreme pathways. The Type I pathway (green) is typically the type of solution that most constrain-based models are seeking. The Type II pathway ($A \rightarrow B \rightarrow C \rightarrow A$, blue) is a typical futile cycle, since it does not involve any primary exchange reactions, but does *waste* ATP. The Type III pathway ($A \rightarrow C \rightarrow D \rightarrow A$, red) is called internal since none of its reactions are exchange reactions. An internal cycle will never be thermodynamically feasible.

1.3.1 The Gibbs free energy of reaction

The second law of thermodynamics states that entropy can never decrease over time in an isolated system. In living cells, which are open systems in constant pressure, temperature, and pH, entropy can be replaced by the concept of transformed Gibbs free energy of reaction [1], denoted $\Delta_r G'$. Then, for a chemical reaction to be feasible, it must impose a negative change in $\Delta_r G'$:

$$\Delta_r G' < 0 \tag{9}$$

For the purpose of constraint-based modeling, it is enough to understand how each $\Delta_r G'$ is affected by its reactant concentration. We typically use the assumption that biochemical reactions occur in dilute aqueous solutions (i.e. the activity coefficient of uncharged species is unity). In this case, the transformed Gibbs free energy of reaction j is given by:

$$\Delta_r G'_j = \Delta_r G'^{\circ}_j + RT \cdot \ln(Q'_j) \tag{10}$$

$$Q'_j = \prod_i c_i^{S_{ij}} \tag{11}$$

where R is the gas constant, T is the temperature (in Kalvin), Q' is the biochemical reaction quotient, c_i is the molar concentration of a reactant (substrate of product) and S_{ij} are the corresponding stoichiometric coefficients (i.e. a row in S_{int}). $\Delta_r G'^{\circ}_j$ represents that $\Delta_r G'_j$ in standard conditions, which is typically defined as 1 M for each one of the reactants. One way to measure $\Delta_r G'^{\circ}_j$, is to let the reaction run until it reaches equilibrium, and measure the concentrations of all reactants. The value of Q' in equilibrium is called the *equilibrium constant* and is denoted K'. Also, since know that at equilibrium $\Delta_r G'_j = 0$, we can solve for $\Delta_r G'^{\circ}_j$ and get:

$$\Delta_r G'^{\circ}_j = -RT \cdot \ln(K'_j) \tag{12}$$

For further reading on thermodynamics of biochemical reactions, see [2, 12].

Finally, we would like to rewrite equations 10 in a more convenient way that corresponds well with our linear algebra notation. First, one can notice that taking the log from Q' makes it a linear function of the log-concentration

values: $\ln(Q'_j) = \sum_i S_{ij} \cdot \ln(c_i)$. In addition, from now on we will use a matrix notation for calculating the vector of all reaction Gibbs energies ($\boldsymbol{\Delta_r G'}$):

$$\boldsymbol{\Delta_r G'} = \boldsymbol{\Delta_r G'^{\circ}} + RT \cdot S^{\top} \boldsymbol{x} \tag{13}$$

where we define $\boldsymbol{x}$ as the vector of log-concentrations, i.e. $x_i = \ln(c_i)$.

1.3.2 Irreversibility is an approximation

In pure physics terms, every chemical reaction should be reversible. There are cases, where the equilibrium is so far from unity, which makes it impossible for the reaction to reach equilibrium in practice. In these cases, it is sometimes convenient to assume that the reaction is irreversible and complete ignore the reverse direction. In defining the flux cone, this assumption can significantly reduce the solution space and number of EFMs.

Most thermodynamics-aware models, however, cannot make this assumption. Every reaction must have a defined equilibrium constant, no matter how extreme it might be. Therefore, from now on, we assume $\Delta_r G'^{\circ}_j$ has a finite value, even for reactions that are defined as irreversible in our model. There is no universal threshold for $\Delta_r G'^{\circ}_j$ which make a reaction irreversible. It is context dependent and a topic of much debate [10, 6, 13].

Beyond $\Delta_r G'^{\circ}_j$, probably the most important parameters for determining the reversibility are the range of concentrations allowed for each of the reactants. Basically, a reversible reaction should have both positive and negative values to its $\Delta_r G'_j$, depending on the chosen reactant concentrations within their predefined ranges. If we denote the vectors of lower and upper bounds as $\boldsymbol{b}^L$ and $\boldsymbol{b}^U$, respectively, the constraint on $\boldsymbol{x}$ would be:

$$\ln(\boldsymbol{b}^L) \leq \boldsymbol{x} \leq \ln(\boldsymbol{b}^U) \tag{14}$$

It is easy to see, that the lowest value for $\Delta_r G'_j$ is achieved when all substrate concentrations are set to their value in $\boldsymbol{b}^U$, and all product concentrations to $\boldsymbol{b}^L$. The highest value of $\Delta_r G'_j$ occurs in the opposite extreme. Then, it become straightforward to check if a reaction is, by itself, reversible or not.

1.3.3 Thermodynamic Feasibility of EFMs

Given a specific elementary flux mode $\boldsymbol{e}^j$ (or any flux vector $\boldsymbol{v}$, for that matter), we would like to know whether it is thermodynamically feasible according to the second law. First, every one of the support reactions must be feasible in the direction defined by $\boldsymbol{e}^j$. However, the same metabolite can be a substrate for one reaction, and a product for another and therefore there is inter-dependence between the $\Delta_r G'_j$ values of these reactions. Sometimes, even though a series of reactions can be individually feasible, their combination is infeasible. This has been termed a *distributed bottleneck* by Mavrovouniotis [10] in 1993.

Using Linear Programming, it is relatively simple to test thermodynamic feasibility:

$$\begin{aligned}
\mathsf{Find} \quad & \boldsymbol{x} \\
\mathsf{such\ that} \quad & \\
\boldsymbol{\Delta_r G'} &\equiv \boldsymbol{\Delta_r G'^{\circ}} + RT \cdot S^{\top} \boldsymbol{x} \\
\boldsymbol{\Delta_r G'} &< 0 \\
\ln(\boldsymbol{b}^L) &\leq \boldsymbol{x} \leq \ln(\boldsymbol{b}^U)
\end{aligned} \tag{15}$$

1.3.4 Max-min Driving Force

In order to provide a more quantitative measure for the thermodynamic feasibility of a given pathway, the Max-min Driving Force (MDF) method provides a relatively simple measure. We adjust the Linear Program in Equation 15, by adding a margin variable (B) and maximizing its value:

$$\begin{aligned}
\mathsf{MDF} &\equiv \max_{B,\boldsymbol{x}} B \\
\mathsf{such\ that} \quad & \\
\boldsymbol{\Delta_r G'} &\equiv \boldsymbol{\Delta_r G'^{\circ}} + RT \cdot S^{\top} \boldsymbol{x} \\
\boldsymbol{\Delta_r G'} &< -B \\
\ln(\boldsymbol{b}^L) &\leq \boldsymbol{x} \leq \ln(\boldsymbol{b}^U)
\end{aligned} \tag{16}$$

If the MDF is positive, then there exists a set of concentrations $\boldsymbol{x}$ (inside the allowed range) such that $\boldsymbol{\Delta_r G'} \leq -B < 0$, i.e. this pathway is thermodynamically feasible. The larger B is, the "more feasible" it is, since we can keep all reactions

farther away from equilibrium. From non-equilibrium thermodynamic theory, we know that reactions that are far from equilibrium tend to be more efficient (as described by the flux-force relationship [3]). An example for MDF calculation is given in Figure 4.

An online interface for running MDF analysis can be found here: `http://equilibrator.weizmann.ac.il/pathway/`.

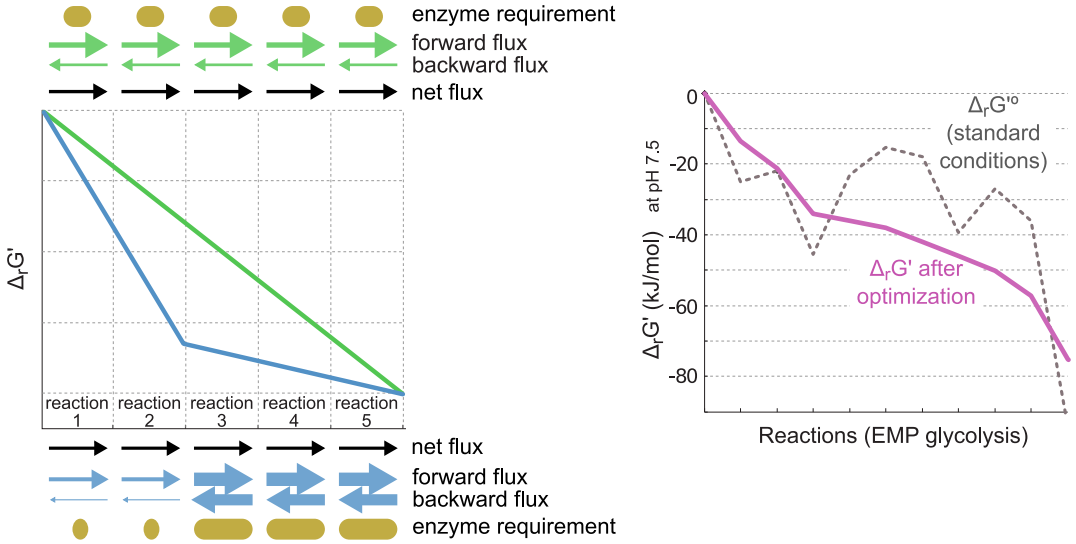

Figure 4: (left) Schematic comparison between two pathways. Each pathway starts and ends with the same compounds, employs five enzymes and carries the same net flux. The kinetic parameters of all enzymes in both pathways, as well as enzyme and metabolite concentrations, are assumed to be identical. (right) Energetic profile of Embden-Meyerhof-Parnas glycolysis. Dashed black line corresponds to $\Delta_r G'^{\circ}$ values (metabolite concentrations of 1M) of pathway reactions at pH 7.5. Red line corresponds to $\Delta_r G'$ values of pathway reactions after an optimization procedure that maximizes the driving force of the thermodynamic bottleneck reactions. Figure is from Noor et al. [14].

1.3.5 Thermodynamic Flux Balance Analysis (TFBA)

Thermodynamic FBA (also known as Thermodynamic-based Metabolic Flux Analysis [6]) was designed to deal with thermodynamically infeasible flux solutions within the framework of FBA:

$$
\begin{aligned}
&\textbf{FBA} \\
\boldsymbol{v}^* &= \arg\max_{\mathrm{v}} \boldsymbol{c}^\top \boldsymbol{v} \\
&\text{such that:} \\
S_{\text{int}}\boldsymbol{v} &= \mathbf{0} \qquad (17) \\
\boldsymbol{v}_{LB} &\leq \boldsymbol{v} \leq \boldsymbol{v}_{UB} \qquad (18)
\end{aligned}
$$

where $\boldsymbol{c} \in \mathbb{R}^r$ is the objective function, and the constants are the internal stoichiometric matrix $S_{\text{int}} \in \mathbb{R}^{m \times r}$. $\boldsymbol{v}_{LB}$ and $\boldsymbol{v}_{UB}$ are the lower and upper bounds on the fluxes, typically relevant only for exchange fluxes.

TFBA adds another two sets of variables – the boolean flux indicators ($\boldsymbol{y} \in \{0,1\}^r$), and the log-concentrations ($\boldsymbol{x} \in \mathbb{R}^r$). Then, the following constraints are added to the Linear Problem (which is now actually a Mixed-Integer

Linear Problem – MILP):

$$
\begin{aligned}
&\textbf{TFBA}\\
\boldsymbol{v}^* &= \arg\max_{\mathrm{v}} \boldsymbol{c}^\top \boldsymbol{v}\\
&\text{such that:}\\
S_{\mathsf{int}}\boldsymbol{v} &= \mathbf{0}\\
\boldsymbol{v}_{LB} &\le \boldsymbol{v} \le \boldsymbol{v}_{UB}\\
0 &\le M\boldsymbol{y} - \boldsymbol{v} \le M && (19)\\
0 &< M\boldsymbol{y} + \boldsymbol{\Delta_r G'} < M && (20)\\
\boldsymbol{\Delta_r G'} &= \boldsymbol{\Delta_r G'^{\circ}} + RT \cdot S^\top \boldsymbol{x} && (21)\\
\ln(\boldsymbol{b}^L) &\le \boldsymbol{x} \le \ln(\boldsymbol{b}^U) && (22)
\end{aligned}
$$

The new constants are the vector of standard Gibbs energies of reaction $\boldsymbol{\Delta_r G'^{\circ}}$ (in units of kJ/mol), the gas constant $R = 8.31$ J/mol/K and temperature $T = 300$ K. M which is a very large number (larger than any of the possible flux and Gibbs free energy). Note also the difference between the internal stoichiometric matrix (S_{int}) used for the pseudo steady-state assumption, versus the full stoichiometric matrix (S) used in the Gibbs free energy calculation.

Equations 21-22 should be familiar, and are exactly the same as in previous sections. To understand how the other two constraints (19-20) enforce thermodynamic feasibility, we consider three possible cases for each reaction j separately:

1. $v_j > 0$: the only possible value that y_j can have is 1, otherwise, $My_j - v_j$ would be negative and violate constraint 19. Therefore, constraint 20 becomes $0 < M + \Delta_r G'_j < M$, which means that $\Delta_r G'_j < 0$.
2. $v_j < 0$: the only possible value that y_j can have is 0, otherwise, $My_j - v_j$ would be larger than M and violate constraint 19. Therefore, constraint 20 becomes $0 < \Delta_r G'_j < M$, which means that $\Delta_r G'_j > 0$.
3. $v_j = 0$: both 0 and 1 are possible solutions for y_j. Therefore, there are no constraints on $\Delta_r G'_j$.

Summarizing these 3 cases, one can concisely write:

$$\forall j : v_j = 0 \ \vee \ \mathrm{sign}(v_j) = -\mathrm{sign}(\Delta_r G'_j)\,, \tag{23}$$

which is exactly the second law of thermodynamics.

1.3.6 Loopless Flux Balance Analysis (ll-FBA)

The loopless algorithm [16] is very similar to TFBA, except that there are no actual thermodynamic values. This way, thermodynamically infeasible internal (Type III) cycles are eliminated, while all other pathways are kept [11]. The set of equations describing ll-FBA are:

$$
\begin{aligned}
&\textbf{ll-FBA}\\
\boldsymbol{v}^* &= \arg\max_{\mathrm{v}} \boldsymbol{c}^\top \boldsymbol{v}\\
&\text{such that:}\\
S_{\mathsf{int}}\boldsymbol{v} &= \mathbf{0} && (24)\\
\boldsymbol{v}_{LB} &\le \boldsymbol{v} \le \boldsymbol{v}_{UB} && (25)\\
0 &\le M\boldsymbol{y} - \boldsymbol{v} \le M && (26)\\
0 &< M\boldsymbol{y} + \boldsymbol{\Delta_r G'} < M && (27)\\
\boldsymbol{\Delta_r G'} &\in \ker(S)^\perp && (28)
\end{aligned}
$$

here, $\boldsymbol{\Delta_r G'}$ is not constrained by the $\boldsymbol{\Delta_r G'^{\circ}}$ and the metabolite concentrations, but is only required to be orthogonal to the null-space of S (or, equivalently, to be in image($S^\top$)).

Why does ll-FBA not eliminate type II cycles from the set of flux solutions? These futile cycles "waste" resources such as ATP, but are thermodynamically feasible. They are only considered "cycles" because we chose to give co-factors a special status (namely, they are external metabolites that do not need to be kept at steady-state). In other words, type II EFMs are in the null-space of S_{int}, but not in the null-space of S (just like type I EFMs). Therefore, it is possible to assign negative $\Delta_r G'$ values to all reactions in a type II EFMs.

Interestingly, this argument can be applied also to the reverse of such futile cycle. Running a futile cycle in reverse is sometimes called an Energy Generating Cycle (EGC), and is obviously unrealistic. Furthermore, unlike type III cycles that do not affect the biomass rate in standard FBA, EGCs have the potential to increase the maximal biomass yield and pose a more serious problem in FBA models [5]. This is one of the cases where TFBA differs from ll-FBA, as it prevents the use of EGCs completely, while ll-FBA doesn't.

1.4 Metabolic engineering algorithms

1.4.1 Phenotypic Phase Plane (PPP)

This useful concept derived from FBA, is a term for 2D projections on the flux polytope. Typically, one chooses specific reaction (e.g. exchange of succinate) for the y-axis and the FBA objective (i.e. biomass rate) for the x-axis. The projection of the flux polytope onto this plane creates a convex polygon shape and helps to visualize how important a certain flux is for creating biomass. Often for cases where the selected flux is a fermentation product, the PPP is referred to as a *production envelope*. It is standard practice to present strain-design results by overlaying two or more PPPs, e.g. a knockout strain versus the wild-type (Figure 5). Using such comparisons, it is easy to see if the knockout is useful for forcing a cell to produce the desired chemical byproduct, by coupling its production to the biomass rate.

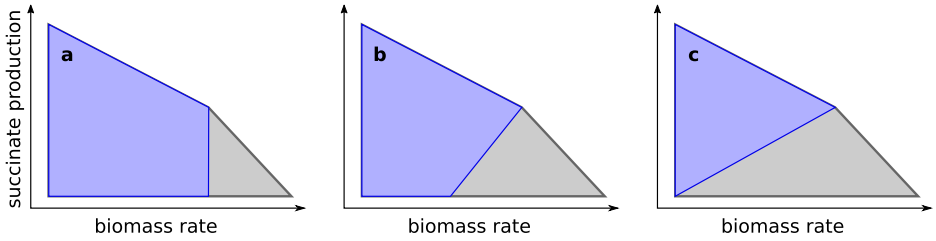

Figure 5: Phenotypic phase planes for anaerobic succinate production from glucose based on the *E. coli* core model. (a) wild-type strain (light gray) vs. triple-deletion mutant (ACKr, ATPS4r, FUM) resulting in a design without growth-coupling (purple). (b) wild-type strain vs. triple-deletion mutant (ACALD, PYK, ME2) resulting in a partially growth-coupled design. (c) wild-type strain vs. double-deletion mutant (ACALD, LDH_D) resulting in a fully growth-coupled design. Figure is from Machado and Herrgård [8].

1.4.2 OptKnock and its derivatives

Burgard, Pharkya, and Maranas [4] phrased the OptKnock optimization problem first as a bi-level mixed-integer linear optimization problem. The inner problem is standard FBA, maximizing an objective function $\boldsymbol{c}_{\mathsf{in}}^{\top}\boldsymbol{v}$. This encodes the assumption that each knockout strain we consider for the outer problem, evolves to maximize the objective (which is typically set to be the biomass function). Then, we choose the one knockout that maximizes the outer objective $\boldsymbol{c}_{\mathsf{out}}^{\top}\boldsymbol{v}$ – typically the secretion rate of the desired product (e.g. succinate).

$$
\begin{aligned}
&\textbf{OptKnock}\\
&\max_{\boldsymbol{y}} \quad \boldsymbol{c}_{\mathsf{out}}^{\top}\boldsymbol{v}\\
&\text{subject to}\\
&\qquad\qquad \max_{\boldsymbol{v}} \ \boldsymbol{c}_{\mathsf{in}}^{\top}\boldsymbol{v}\\
&\qquad\qquad \text{subject to}\\
&\qquad\qquad\qquad\qquad S_{\mathsf{int}}\boldsymbol{v} = \boldsymbol{0}\\
&\qquad\qquad\qquad\qquad v_{\mathsf{LB},\,j}\cdot y_j \le v_j \le y_j \cdot v_{\mathsf{UB},\,j}\\
&\qquad\qquad \boldsymbol{y}\in\{0,1\}^r\\
&\qquad\qquad \textstyle\sum_j (1-y_j) \le K
\end{aligned}
\tag{29}
$$

where, K is the maximum number of allowed knockouts. The inner LP is a standard FBA optimization problem, except that a few of the reactions are "knockout out" – when a reaction indicator is set to 0 by the external LP it forces the corresponding flux to zero, i.e. $y_j = 0 \rightarrow v_j = 0$. In addition, we set $v_{\mathsf{LB},\,biomass} > 0$ and $y_{biomass} = 1$, since we only want to consider mutants that show at least some minimal growth. Finally, the external LP places an upper bound on the total number of knockouts, which is useful for two reasons: (1) making sure we don't have too

many knockouts that would make the genetic manipulation impractical, and (2) to reduce the computational load required to solve the MILP. Nevertheless, even for small K, solving the bi-level MILP directly is not tractable for large networks (such as the genome-scale *E. coli* model). Fortunately, Burgard, Pharkya, and Maranas [4] provided a solution, by converting into a single-level MILP using duality theory.

The dual of the inner optimization problem in 29 is:

$$\begin{aligned} \min_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\omega}} \quad & \sum_j y_j(v_{\mathsf{UB},\,j}\cdot\mu_j - v_{\mathsf{LB},\,j}\cdot\omega_j) \\ \text{subject to} \quad & \\ & S_{\mathsf{int}}^\top\boldsymbol{\lambda} + \boldsymbol{\mu} - \boldsymbol{\omega} \geq \boldsymbol{c}_{\mathsf{in}} \\ & \boldsymbol{\mu},\boldsymbol{\omega} \geq 0\,. \end{aligned} \tag{30}$$

Since duality theory states if the optimums solutions of the primal and dual problems are bounded, their objective function values must be equal to one another, so ensuring optimality can be achieved by equating the two objectives: $\boldsymbol{c}_{\mathsf{in}}^\top\boldsymbol{v} = \sum_j y_j(v_{\mathsf{UB},\,j}\cdot\mu_j - v_{\mathsf{LB},\,j}\cdot\omega_j)$. However, using this equation as a constraint in the outer optimization problem would make it non-linear, since it contains products of two variables ($y_j\cdot\mu_j$ and $y_j\cdot\omega_j$). We can, however, use the fact that y_j are binary in order to cast this expression into a linear form.

First, we introduce a new set of auxiliary variables $\boldsymbol{m}\in\mathbb{R}^r$, and constrain them to be $m_j = y_j\cdot\mu_j$ by adding the following:

$$\begin{aligned} 0 \leq \quad & m_j \quad \leq y_j\cdot M \\ \mu_j - (1-y_j)\cdot M \leq \quad & m_j \quad \leq \mu_j\,. \end{aligned} \tag{31}$$

As before, M is a constant larger than any possible value of μ_j. Considering the two cases for y_j, one can easily see that $y_j = 0 \;\rightarrow\; m_j = 0$ and $y_j = 1 \;\rightarrow\; m_j = \mu_j$, which is what we wanted to achieve. We do exactly the same for $\boldsymbol{\omega}$ and another set of variables $\boldsymbol{u}\in\mathbb{R}^r$. Then we can rewrite the dual objective as: $\sum_j y_j(v_{\mathsf{UB},\,j}\mu_j - v_{\mathsf{LB},\,j}\omega_j) = \boldsymbol{v}_{\mathsf{UB}}^\top\boldsymbol{m} - \boldsymbol{v}_{\mathsf{LB}}^\top\boldsymbol{u}$.

Using this notion, we get a single-level MILP problem for OptKnock:

OptKnock

$$\begin{aligned} \max_{\boldsymbol{y},\boldsymbol{v},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\omega},\boldsymbol{m},\boldsymbol{u}} \quad & \boldsymbol{c}_{\mathsf{out}}^\top\boldsymbol{v} \\ \text{subject to} \quad & \\ & S_{\mathsf{int}}\boldsymbol{v} = \boldsymbol{0} \\ & v_{\mathsf{LB},\,j}\cdot y_j \leq v_j \leq y_j\cdot v_{\mathsf{UB},\,j} \\ & \boldsymbol{c}_{\mathsf{in}}^\top\boldsymbol{v} = \boldsymbol{v}_{\mathsf{UB}}^\top\boldsymbol{m} - \boldsymbol{v}_{\mathsf{LB}}^\top\boldsymbol{u} \\ & \boldsymbol{c}_{\mathsf{in}} \leq S_{\mathsf{int}}^\top\boldsymbol{\lambda} + \boldsymbol{\mu} - \boldsymbol{\omega} \\ & 0 \leq \boldsymbol{\mu},\boldsymbol{\omega} \\ & 0 \leq \boldsymbol{m} \leq \boldsymbol{y}\cdot M \\ & \boldsymbol{\mu} - (1-\boldsymbol{y})\cdot M \leq \boldsymbol{m} \leq \boldsymbol{\mu} \\ & 0 \leq \boldsymbol{u} \leq \boldsymbol{y}\cdot M \\ & \boldsymbol{\omega} - (1-\boldsymbol{y})\cdot M \leq \boldsymbol{u} \leq \boldsymbol{\omega} \\ & \textstyle\sum_j (1-y_j) \leq K \end{aligned} \tag{32}$$

1.4.3 Derivatives of OptKnock

A comprehensive review by Machado and Herrgård [8] lists the many published algorithms for *in silico* strain design that followed after OptKnock (see Figure 1.4.3).

RobustKnock is an improvement for OptKnock, introduced by Tepper and Shlomi [18] in 2010, which accounts for competing pathways that might reduce the chemical production rates. One common issue with standard OptKnock, is that it only looks at the maximal rate of the chemical production rate: $\max \boldsymbol{c}_{\mathsf{out}}^\top\boldsymbol{v}$. However, as visualized by the production envelope in Figure 5a, there might be redundancy in the solution space, where the same biomass rate can be achieved without any production of the desired chemical. RobustKnock addresses this problem by performing a max-min optimization, i.e. maximizing the lowest value of $\boldsymbol{c}_{\mathsf{out}}^\top\boldsymbol{v}$ over all $\boldsymbol{v}$ that satisfy the inner LP. This method yields solutions like in Figure 5b-c.

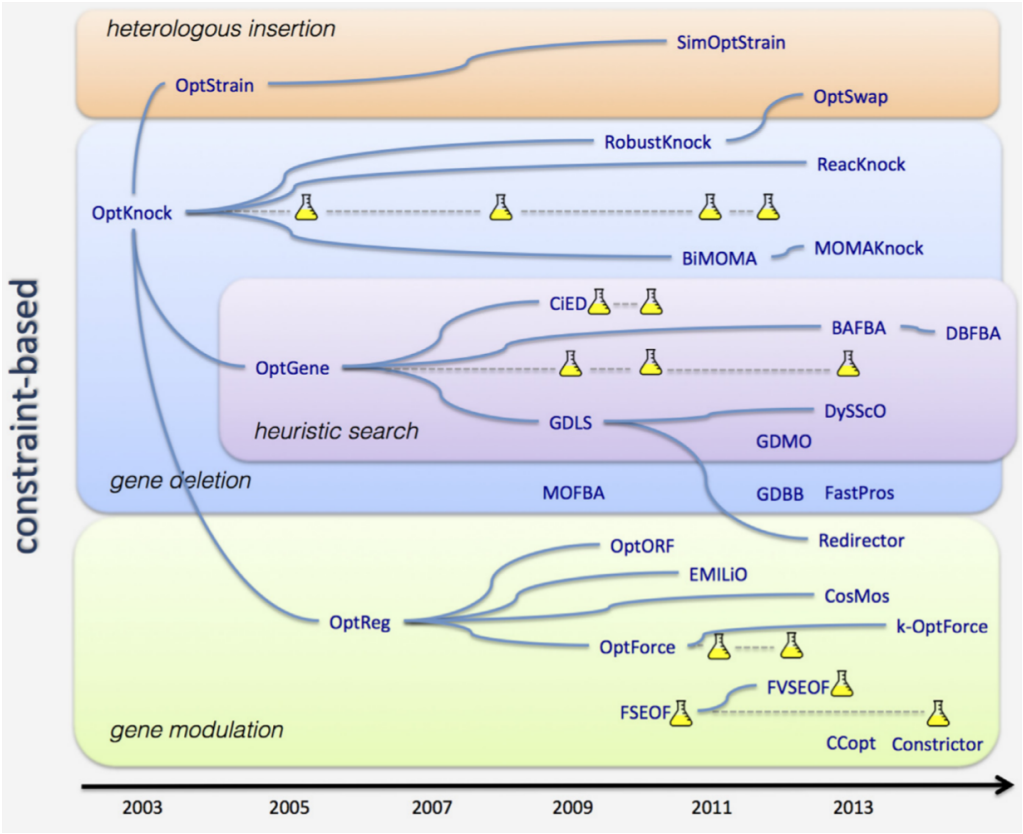

Figure 6: A timeline of all constraint-based strain design methods that followed OptKnock (figure is from [8]).

References

[1] Robert A. Alberty. *Biochemical Thermodynamics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Mar. 2006. ISBN: 0-471-75798-5.

[2] Robert A. Alberty, Athel Cornish-Bowden, Robert N. Goldberg, Gordon G. Hammes, Keith Tipton, and Hans V. Westerhoff. "Recommendations for Terminology and Databases for Biochemical Thermodynamics". In: *Biophysical Chemistry* 155.2 (May 1, 2011), pp. 89–103. ISSN: 0301-4622.

[3] Daniel A. Beard and Hong Qian. "Relationship between Thermodynamic Driving Force and One-Way Fluxes in Reversible Processes". In: *PLoS ONE* 2.1 (Jan. 3, 2007), e144.

[4] Anthony P. Burgard, Priti Pharkya, and Costas D. Maranas. "Optknock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization". In: *Biotechnology and Bioengineering* 84.6 (Dec. 20, 2003), pp. 647–657. ISSN: 1097-0290.

[5] Claus Jonathan Fritzemeier, Daniel Hartleb, Balázs Szappanos, Balázs Papp, and Martin J. Lercher. "Erroneous Energy-Generating Cycles in Published Genome Scale Metabolic Networks: Identification and Removal". In: *PLOS Computational Biology* 13.4 (Apr. 18, 2017), e1005494. ISSN: 1553-7358.

[6] Christopher S. Henry, Linda J. Broadbelt, and Vassily Hatzimanikatis. "Thermodynamics-Based Metabolic Flux Analysis". In: *Biophysical Journal* 92.5 (Mar. 1, 2007), pp. 1792–1805. ISSN: 0006-3495.

[7] Steffen Klamt and Jörg Stelling. "Two Approaches for Metabolic Pathway Analysis?" In: *Trends in Biotechnology* 21.2 (Feb. 1, 2003), pp. 64–69. ISSN: 0167-7799, 1879-3096. pmid: 12573854.

[8] Daniel Machado and Markus J. Herrgård. "Co-Evolution of Strain Design Methods Based on Flux Balance and Elementary Mode Analysis". In: *Metabolic Engineering Communications* 2 (Dec. 1, 2015), pp. 85–92. ISSN: 2214-0301.

[9] Sayed-Amir Marashi, Laszlo David, and Alexander Bockmayr. "Analysis of Metabolic Subnetworks by Flux Cone Projection". In: *Algorithms for Molecular Biology* 7 (May 29, 2012), p. 17. ISSN: 1748-7188.

[10] Michael L Mavrovouniotis. "Identification of Localized and Distributed Bottlenecks in Metabolic Pathways". In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 1 (1993), pp. 275–283. ISSN: 1553-0833. pmid: 7584346.

[11] Elad Noor, E Nathan Lewis, and Ron Milo. "A Proof for Loop-Law Constraints in Stoichiometric Metabolic Networks". In: *BMC Syst. Biol.* 6 (Jan. 2012).

[12] Elad Noor, Avi Flamholz, Wolfram Liebermeister, Arren Bar-Even, and Ron Milo. "A Note on the Kinetics of Enzyme Action: A Decomposition That Highlights Thermodynamic Effects". In: *FEBS Letters.* A century of Michaelis - Menten kinetics 587.17 (Sept. 2, 2013), pp. 2772–2777. ISSN: 0014-5793.

[13] Elad Noor, Arren Bar-Even, Avi Flamholz, Yaniv Lubling, Dan Davidi, and Ron Milo. "An Integrated Open Framework for Thermodynamics of Reactions That Combines Accuracy and Coverage". In: *Bioinformatics* 28.15 (Aug. 1, 2012), pp. 2037–2044. ISSN: 1367-4803.

[14] Elad Noor, Arren Bar-Even, Avi Flamholz, Ed Reznik, Wolfram Liebermeister, and Ron Milo. "Pathway Thermodynamics Highlights Kinetic Obstacles in Central Metabolism". In: *PLOS Computational Biology* 10.2 (Feb. 20, 2014), e1003483. ISSN: 1553-7358.

[15] Nathan D. Price, Iman Famili, Daniel A. Beard, and Bernhard Ø Palsson. "Extreme Pathways and Kirchhoff's Second Law". In: *Biophysical Journal* 83.5 (Nov. 1, 2002), pp. 2879–2882. ISSN: 0006-3495. pmid: 12425318.

[16] Jan Schellenberger, Nathan E. Lewis, and Bernhard Ø. Palsson. "Elimination of Thermodynamically Infeasible Loops in Steady-State Metabolic Models". In: *Biophysical Journal* 100.3 (Feb. 2, 2011), pp. 544–553. ISSN: 0006-3495.

[17] Christophe H. Schilling, David Letscher, and Bernhard Ø. Palsson. "Theory for the Systemic Definition of Metabolic Pathways and Their Use in Interpreting Metabolic Function from a Pathway-Oriented Perspective". In: *Journal of Theoretical Biology* 203.3 (Apr. 7, 2000), pp. 229–248. ISSN: 0022-5193.

[18] Naama Tepper and Tomer Shlomi. "Predicting Metabolic Engineering Knockout Strategies for Chemical Production: Accounting for Competing Pathways". In: *Bioinformatics* 26.4 (Feb. 15, 2010), pp. 536–543. ISSN: 1367-4803.