Reg4OptFlux: an OptFlux plug-in that comprises meta-heuristics approaches for Metabolic Engineering using Integrated Models

Orlando Rocha*, Paulo Vilaça*†, Miguel Rocha* and Rui Mendes*
*CEB/CCTC, University of Minho, Portugal
Email: orocha@deb.uminho.pt, mrocha@di.uminho.pt, rcm@di.uminho.pt
†SilicoLife, Portugal
Email: pvilaca@silicolife.com

***Abstract*—Metabolic engineering (ME) strategies have been implemented over the last few years, in order to improve microbial strains of interest in industrial biotechnology. With the advent of experimental data concerning to regulatory aspects, several efforts have been conducted to incorporate this information in genome-scale metabolic models, aiming at the improvement of phenotype simulation methods. However, most of these methods can be used only by computer science experts, since they are not available in user-friendly software ME frameworks. This work presents Reg4OptFlux, a computational framework for ME, that integrates methods for phenotype simulation and optimization strain design, relying on integrated metabolic and regulatory models. Meta-heuristic approaches such as Evolutionary Algorithms and Simulated Annealing were appropriately modified to accommodate the optimization tasks, and were applied to study the optimization of ethanol and succinic acid production using an integrated model of the *E.coli* host. The framework was implemented as a plug-in for OptFlux, an open-source software for ME, and it is available in the OptFlux web site (www.optflux.org).**

***Keywords*-Metabolic engineering; Integrated models; Regulatory models; Meta-heuristics approaches; Open-source;**

I. Introduction

The inherent complexity of cellular systems has led to the development of a variety of *in silico* modeling approaches for the simulation, optimization and analysis of biological processes. These efforts have been driven by the development of engineered microbial strains capable of accomplishing desired biotransformations, or overproduction of valuable biochemicals. In recent years, constraint-based modeling approaches have been widely applied to analyze, interpret and predict cellular phenotype under defined environmental conditions. These approaches impose governing physicochemical constraints, such as flux capacities, thermodynamics and mass conservation, in order to reduce the solution space of the feasible flux distributions [1].

Flux Balance Analysis (FBA) has been the most successfully used constraint-based method, relying on the linear optimization of an objective function, commonly the maximization of biomass, to reach an optimal flux distribution [1], [2]. Notwithstanding, alternative approaches, such as the method of minimization of metabolic adjustment (MOMA) [2] and the regulatory on/off minimization (ROOM) [3] methods, were developed to address aspects related with the phenotype simulation of mutant strains affected by genetic perturbations.

Bi-level optimization strategies have been applied to pinpoint genetic modifications that can lead to improvement of biochemical production [4]. OptKnock [5] was the first optimization method designed to obtain reaction deletion strategies for the overproduction of a metabolite, where a bi-level optimization problem is reformulated into a mixed integer linear programming (MILP) problem, based on the strong duality theorem. Since then, several variations of OptKnock have been implemented. OptReg [6] extended OptKnock, introducing up/down regulation of various reactions in addition to knockouts. OptGene and its variants [7], [8] presented an alternative meta-heuristic optimization approach using Evolutionary Algorithms (EAs) and Simulated Annealing (SA) to find the best set of gene knockouts. Despite of providing near-optimal solutions when compared to OptKnock, it demands less computational time to solve larger size problems, being also more flexible regarding to the objective function that can optimized.

With the advent of experimental data concerning regulatory aspects, several efforts have been performed to incorporate regulatory information (e.g. association of a transcriptional/translational layer) in constraint-based models, to improve their accuracy in phenotype predictions. Regulatory constraints were firstly introduced in constraint-based models in the regulatory FBA (rFBA) method presented by Covert *et al.* [9]. A Boolean logic representation was used to characterize the transcriptional regulatory structure. The capabilities of this approach were enhanced in posterior work of the authors [10], [11]. Shlomi *et al.* [12] presented the Steady-state Regulatory Flux Balance Analysis (SR-FBA), an extension of rFBA, using a MILP approach to simulate a pair of consistent metabolic and regulatory steady states, by satisfying both regulatory and metabolic constraints. Later, Kim and Reed presented OptORF [4], an optimization method for strain design, using simultaneously regulatory and metabolic information. A bilevel optimization approach based on the OptKnock approach is used to identify the optimal metabolic and regulatory gene deletions, as well as

978-1-4799-5937-2/14/$31.00 ©2014 IEEE

genes to over-express, that maximize the production of a target metabolite being the Boolean gene-reactions rules and transcriptional regulatory rules imposed as constraints.

Most of these methods have been integrated in software frameworks, such as the COBRA Toolbox [13], the Tiger Toolbox [14], OptFlux [15] and CellNetAnalyser [16]. All these suites allow users to perform the most common tasks used in this area, such as FBA and flux variability analysis (FVA). However, only OptFlux and COBRA are able to perform strain optimization tasks. Moreover, some of these frameworks are capable of loading regulatory information, but none of them can perform strain optimization tasks using simultaneously regulatory and metabolic models. Also, most of the referred frameworks are implemented in MATLAB, a comercial platform. Moreover, COBRA and Tiger do not provide a user-friendly interface, an important feature for researchers. In contrast, OptFlux is a Java open-source framework that provides user-friendly interfaces to their features. Here, we present a plug-in to provide features for strain design in OptFlux, which encompass strain optimization algorithms and phenotype simulation methods, able to work with integrated metabolic and regulatory models.

II. Core features of OptFlux

OptFlux is an extensive platform fully implemented in the Java language, that provides an extensive set of functionalities for Metabolic Engineering through user-friendly interfaces. Stoichiometric metabolic models can be loaded from different model formats (i.e. flat files, Systems Biology Markup Language (SBML) standard and Metatool format). Several simulation algorithms are implemented in OptFlux to perform *in silico* phenotype simulations of wild-type and mutant strains (e.g FBA, parsimonious FBA (pFBA) [17], MOMA, ROOM). Specific environmental conditions can be established to impose constraints over the fluxes. All the information concerning models, such as reactions, metabolites, stoichiometric matrix and gene-protein-reaction associations (GPRs) can be either visualized or exported. Strain optimization using Single and multi-objective Evolutionary Algorithms (EAs), Simulated Annealing (SA) or OptKnock can be performed to identify sets of reaction deletions that optimize a given set of objective functions related with desired industrial goals.

III. Proposed Framework

A. Integrated models

Integrated metabolic and transcriptional regulatory models consist of the aggregation of two layers representing the metabolism and regulation. The integration process is performed through the mapping of GPR associations present in the metabolic model with genes present in the regulatory model. Regulatory models are merely qualitative, where gene relationships and environmental cues are described in Boolean logic. The aim is to provide a Boolean rule for each regulated gene, describing how they are affected by regulatory events (transcription factors or other regulatory genes) or even by environmental conditions. Logical operators, AND, OR, NOT are used in the characterization of these perturbations, to achieve a binary state ”on” or ”off” of each gene. These genes can be either metabolic, those present in GPR associations, or regulatory, those that involved in regulatory processes. Environmental conditions are related to external stimuli, such as compounds in the media or stimuli that can cause perturbations in the system (e.g stress, conditions).

The integrated framework supports the simultaneous loading of both metabolic and regulatory models. It is also possible to create an integrated model, from a metabolic model (containing GPR associations) previously loaded, by joining the regulatory network. The integration process of the two models relies on the following operations:

- Gene connections: through the mapping of genes that are defined in the regulatory model to metabolic genes present in the GPR associations of the metabolic model.
- Mapping of the environmental conditions: by verifying which conditions are equal in the two models. In cases where the conditions correspond, these are considered as ”true” in the regulatory model, otherwise they are considered as ”false”. Additionally, the user may set the conditions that are only present in the regulatory model (to be ”true” or ”false”).
- Mapping of genes and their products: the association between transcriptional factors and the genes that encode them is performed. Then, an abstract syntax tree for each regulatory rule is assembled, which will be used to calculate the binary state of the corresponding gene in the simulation. The state ”true/false” of any transcription factor that is present in a regulatory rule, but is not associated with a gene responsible for its encoding, can also be changed by the user.

B. Methods for phenotype simulation

The regulatory framework comprises two distinct methods to perform the phenotype simulation. The Boolean Regulatory Steady State Constraint-Based Approach (BRSS-CBA) method developed by the authors [18], and the aforementioned SR-FBA. The BRSS-CBA is based on a steady state approximation like rFBA, using a two step approach to reach a steady-state flux distribution consistent with the Boolean regulatory network state. Briefly, in the first step, the regulatory network is simulated to calculate which genes are inactive. The process continues with a constraint-based simulation method using the information of the previous step as constraints. The Boolean simulation follows a synchronous and deterministic Boolean network simulation method, assuming that all variables are updated simultaneously in every step. The assumption is that if the systems have the same initialization, they will reach always the same state. Thus, in this process, the regulatory network is iterated until an attractor is found, representing the steady-state of the network. In case the system reaches a "cyclic attractor" (i.e. the system oscillates between two or more states), only genes that have an "off" state in all states are set to "off". As mentioned before, the aim is to gather a hypothetical set of genes that exhibits an inactive state, resulting from the Boolean operations carried out with the regulatory model. Subsequently, this information is transferred to the metabolic network through the mapping to GPR associations. Finally, the metabolic simulation is performed using one of the methods present in OptFlux (i.e., pFBA, FBA, MOMA, ROOM).

The alternative is the SR-FBA method that applies a MILP formulation to maximize the flux through the biomass reaction, assuming the following constrains: (1) metabolic constraints, (2) regulatory constraints, (3) gene-to-reactions mapping and (4) reaction enzyme state constraints. The regulatory and GPR Boolean rules are transformed into linear equations, by applying the following transformations:

- a = b AND c is formulated as $-1 \leq 2b + 2c - 4a \leq 3$
- a = NOT b is formulated as a+b=1

Other Boolean operators (including OR) are defined using the operators presented above.

C. Algorithms for strain optimization

Similarly to Optflux, single and multi-objective Evolutionary Algorithms or Simulating Annealing methods can be applied to perform the optimization tasks. These optimization methods were originally developed within the authors research group and their implementation can be found in detail in [8]. These algorithms had to be adapted to find the best gene knockout strategies using integrated models. Instead of selecting the reactions to be deleted, the algorithm selects a hypothetical set of genes (metabolic or regulatory), that when eliminated lead to an increase in the production of a desired metabolite, as encoded by the objective function.

1) Solution encoding and evaluation: Only gene deletions are encoded in the solution, using a fixed or variable size set-based representation. The solution consists of a set of integer values, representing the indexes of genes to be deleted. Therefore, this information is used in the aforementioned simulation methods to define which genes are deactivated ("off" state) in the initial Boolean state representation of the genes. Then, the simulation is executed and the output flux distribution will be used in an objective function to evaluate the fitness value. The available objective functions are the Biomass-Product Coupled Yield (BPCY) [7] and the Product Yield with Minimum Biomass (PYMB). In BPCY, the fitness value is calculated by the following formula:

$$BPCY = \frac{P \times B}{S} \qquad (1)$$

where P is the flux of the desired metabolite, B is the value of biomass flux and S is the substrate intake flux. On the other hand, in PYMB, the fitness value is calculated through the ratio between the flux of the desired product and the substrate flux. However, a threshold value is applied that represents the minimal acceptable biomass flux (a percentage value regarding to the wild type strain). Figure 1 shows a general scheme of the optimization algorithms.

2) Pre-processing and post-processing: Due to the high number of variables (i.e genes and reactions) present in the models, these problems are computationally intensive. Therefore, in most cases it is suitable to reduce the number of decision variables to improve the convergence of the algorithms. In this context, a feature is provided to verify the essential genes (e.g their removal leads to non growth phenotypes). Thus, in the optimization tasks, these genes are not considered as targets for deletion, reducing the search space. In a similar fashion, at the end of the optimization process, a simplification of the solutions is performed by removing all unnecessary genes that do not affect the objective function value, keeping only the relevant knockouts.

IV. Case studies

Two case studies were performed, both considering an integrated model of metabolism and regulation for *E. coli*, presented by Covert *et al* [11]. The goal is to produce ethanol and succinic acid with glucose as the limiting substrate. The simulation is conducted in anaerobic growth conditions, establishing the following parameters: $glc(e) = 18.5$, $O_2 = 0$. A series of experiments were set to analyze the behavior of the aforementioned optimization algorithms, using both phenotype simulation methods BRSS-CBA and SR-FBA. Moreover, each experiment was conducted using both objective functions: BPCY and PYMB. The variable size variant was used, enabling different alternatives for the cardinality of the maximum knockout set (k = 2, 4, 6, 8). The termination criteria was set to 50,000 objective function evaluations in all cases. Each experiment was repeated five times due to the stochastic nature of the optimization algorithms. This is a small number of runs, due to the time needed to run simulations using SR-FBA (a computationally demanding MILP formulation). The same gene knockout optimization experiments were also conducted for succinate, using the same conditions, yet without the integration of the regulatory network, for comparison purposes.

V. Results and discussion

Figure 2 summarizes the results obtained in both case studies, showing the mean of the best solutions obtained in the runs for each maximum of allowed knockouts. Figures 2.a) and 2.b) suggest that all used approaches reached similar optimization results in ethanol production. Nevertheless, the PYMB objective function leads to solutions that can achieve better ethanol production rates. From the results

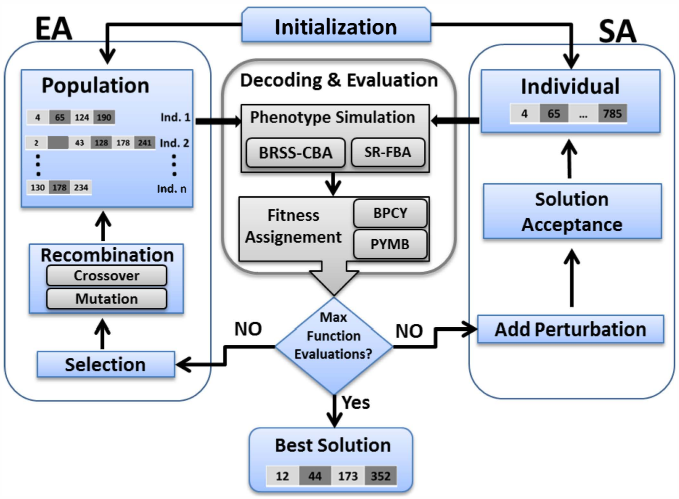

Figure 1. Structure of the strain optimization algorithms used in this work.

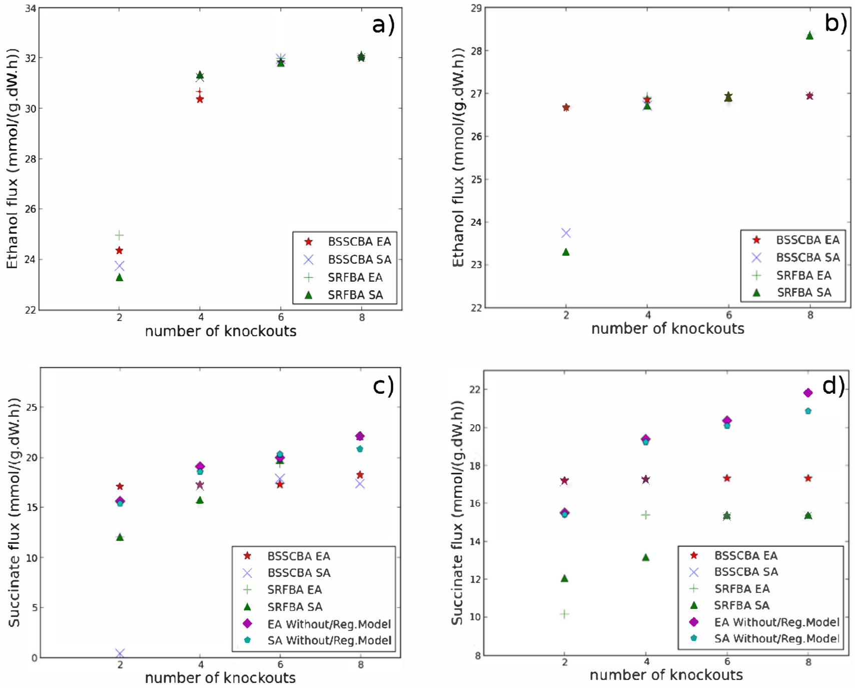

Figure 2. Results obtained by EA and SA using both simulation methods. Figure (a) and (c) are the results using PYMB; Figure (b) and (d) are the results using BPCY.

obtained in the optimization of succinate, a dependence of the approaches to the employed objective function seems to exist with the better results being found by SR-FBA(EA/SA) using the PYMB objective function (Figure 2.c) and BSS-CBA(EA) using the BPCY objective function (Figure 2.d). There were no improvements when incorporating information about the regulatory network in the case of succinate production, being even observed a decrease in succinate production comparing to solutions from optimizations without the regulatory network. This can be explained by the fact that the regulatory model of *E.coli* contains only a small part of the regulatory aspects, and most of the existing regulatory rules are still incomplete. Thus, these results indicate that it is necessary to perform a detailed analysis of the regulatory model to verify the accuracy of the gene rules related to the succinate production, which is not the purpose of this work. The results obtained for ethanol were also compared to the ones reported in [4], and it was verified that the

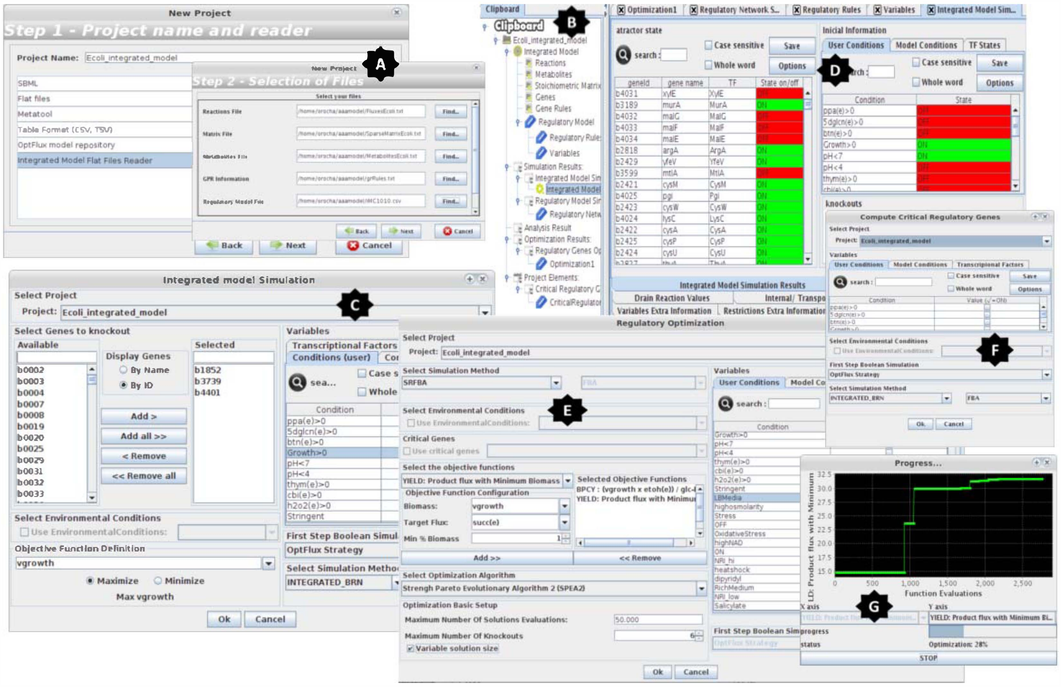

Figure 3. Screenshots of Optflux with the proposed plug-in. (A) loading models; (B) clipboard containing the OptFlux datatypes; (C) mutant simulation; (D) view of the simulation results; (E) strain optimization interface; (F) critical genes interface; (G) optimization runtime interface

proposed optimization methods could achieve similar results. Therefore, these results show that the implemented methodologies are able not only to find similar strategies to the ones previously published, but also to suggest good alternative solutions concerning to gene deletions, in the overproduction of ethanol. Moreover, the implemented features are more user friendly than the existing strain design methods that rely on integrated models.

VI. Features of the Implemented Plug-in

All the features provided by this framework can be accessed by a user-friendly Graphical User Interface (GUI), in the form of an OptFlux plug-in. The instructions concerning to the installation procedure and the functionalities present in the plug-in are available on http://darwin.di.uminho.pt/optfluxwiki/index.php/OptFlux3:RT. Figure 3 shows some screenshots to illustrate the framework layout. The main features are the following:

- **Integration of metabolic and regulatory models**: users can load both models simultaneously (Figure 3.A) or create an integrated model joining the regulatory network with an existing model in Optflux. After loading and integrating all information, they can be accessed in the clipboard (Figure 3.B).
- **Simulation**: users can perform wild type and mutant strain simulations by applying different configurations such as: gene knockouts, environmental conditions, transcriptional factors, external stimuli (defined in the regulatory model) and the simulation method (Figure 3.C). Moreover, users can perform simulations of the regulatory network using similar configurations, with a simulation method based on the first stage step of BSS-CBA (just the Boolean simulation method).
- **Analysis of critical genes**: essential genes can be computed by both simulation methods, in which the user can use different inputs of external stimuli that are defined in the regulatory model (Figure 3.F). These can be saved to a text file and loaded for future use in optimizations tasks.
- **Optimization**: users can perform strain optimization tasks using single and multi-objective EA or SA. Users may configure different parameters in the optimization procedures, such as: objective function, desired flux, critical genes, simulation algorithm, external stimuli, environmental conditions, maximum number of knockouts and number of evaluations. In addition, in multi-objective optimization, multiple objective functions can be established (Figure 3.E). The algorithms progress can be monitored (or stopped) while running (Figure 3.G).
- **Results visualization**: All the results obtained in the operations can be visualized in appropriate graphical interfaces (Figure 3.D).

VII. Conclusions

Computational optimization tools are essential in Metabolic engineering, since they can contribute significantly to the improvement of microbial strains, reducing the production cost of valuable compounds with interest to the industry. In this work, a plug-in that makes the OptFlux platform the first available software to integrate regulatory and metabolic models is proposed, allowing both phenotype simulation and strain optimization operations. The software is available enlarging the tool-set at the disposal of the ME community. The main driving idea was to create tools able to use information concerning to regulatory aspects, that could help in finding gene knockout strategies, leading to the overproduction of desired compounds. Thus, the proposed plug-in combines optimization algorithms and phenotype simulation methods supporting the use of integrated models to help users in pinpointing genetic modifications. The results show that the implemented methods can provide insights of hypothetical gene knockout strategies, that lead to an improvement of desired products. In future work, the authors intend to improve the capabilities of the software, by implementing OptORF and new methods for simulation of Boolean networks.

Acknowledgment

The work is funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project ref. COMPETE FCOMP-01-0124-FEDER-015079 and by Escola de Engenharia da Universidade do Minho - Agenda 2020.

References

[1] K. J. Kauffman, P. Prakash, and J. S. Edwards, "Advances in flux balance analysis," *Current Opinion in Biotechnology*, vol. 14, no. 5, pp. 491–496, Oct. 2003.

[2] D. Segrè, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 23, pp. 15 112–7, Nov. 2002.

[3] T. Shlomi, O. Berkman, and E. Ruppin, "Regulatory on/off minimization of metabolic flux changes after genetic perturbations." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7695–700, May 2005.

[4] J. Kim and J. L. Reed, "OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains." *BMC systems biology*, vol. 4, p. 53, Jan. 2010.

[5] A. P. Burgard, P. Pharkya, and C. D. Maranas, "Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization," *Biotechnology and Bioengineering*, vol. 84, no. 6, pp. 647–657, 2003.

[6] P. Pharkya and C. D. Maranas, "An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems." *Metabolic engineering*, vol. 8, no. 1, pp. 1–13, Jan. 2006.

[7] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen, "Evolutionary programming as a platform for in silico metabolic engineering." *BMC bioinformatics*, vol. 6, no. 1, p. 308, Jan. 2005.

[8] M. Rocha, P. Maia, R. Mendes, J. P. Pinto, E. C. Ferreira, J. Nielsen, K. R. Patil, and I. Rocha, "Natural computation meta-heuristics for the in silico optimization of microbial strains." *BMC bioinformatics*, vol. 9, p. 499, Jan. 2008.

[9] M. W. Covert, C. H. Schilling, and B. Palsson, "Regulation of gene expression in flux balance models of metabolism." *Journal of theoretical biology*, vol. 213, no. 1, pp. 73–88, Nov. 2001.

[10] M. W. Covert and B. O. Palsson, "Constraints-based models: regulation of gene expression reduces the steady-state solution space." *Journal of theoretical biology*, vol. 221, no. 3, pp. 309–25, Apr. 2003.

[11] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks." *Nature*, vol. 429, no. 6987, pp. 92–6, May 2004.

[12] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, "A genome-scale computational study of the interplay between transcriptional regulation and metabolism," *Molecular Systems Biology*, vol. 3, no. 1, 2007.

[13] J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. O. Palsson, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0." *Nature protocols*, vol. 6, no. 9, pp. 1290–307, Sep. 2011.

[14] P. A. Jensen, K. A. Lutz, and J. A. Papin, "TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks," *BMC Systems Biology*, vol. 5, no. 1, p. 147, 2011.

[15] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. a. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha, "OptFlux: an open-source software platform for in silico metabolic engineering." *BMC systems biology*, vol. 4, p. 45, Jan. 2010.

[16] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles, "Structural and functional analysis of cellular networks with CellNetAnalyzer." *BMC systems biology*, vol. 1, no. 1, p. 2, Jan. 2007.

[17] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. O. Palsson, "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models." *Molecular systems biology*, vol. 6, p. 390, Jul. 2010.

[18] P. Vilaça, I. Rocha, and M. Rocha, "A computational tool for the simulation and optimization of microbial strains accounting integrated metabolic/regulatory information." *Bio Systems*, vol. 103, no. 3, pp. 435–41, Mar. 2011.