Computational tools for strain optimization by adding reactions

Sara Correia and Miguel Rocha

Abstract This paper introduces a new plug-in for the OptFlux Metabolic Engineering platform, aimed at finding suitable sets of reactions to add to the genomes of microbes (wild type strain), as well as finding complementary sets of deletions, so that the mutant becomes able to overproduce compounds with industrial interest, while preserving their viability. The optimization methods used are Evolutionary Algorithms and Simulated Annealing. The usefulness of this plug-in is demonstrated by a case study, regarding the production of vanillin by the bacterium *E. coli*.

1 Introduction

An important challenge in Metabolic Engineering (ME) consists in the identification of genetic manipulations to be applied to an organism, with the aim of constructing a mutant strain able to produce compounds of industrial interest. Based on the knowledge about the biological system and, more specifically, its metabolic network, we can manipulate the environment in which it develops, or alter it genetically, to maximize the production of a given compound [6].

Recently, advances have been achieved concerning the available knowledge of some biological organisms, for instance from the sequencing of their genomes and also from various types of high-throughput experimental data (e.g. gene expression, proteomics). However, the lack of tools to perform the analysis and interpretation of biological data still limits the use and interconnection of that knowledge [2].

In this context arises the OptFlux (http://www.optflux.org) [9], an open-source and modular platform for ME, incorporating strain optimization tasks, using Evolutionary Algorithms (EAs) [1] and Simulated Annealing (SA) [5]. OptFlux also allows the use of stoichiometric metabolic models for phenotype simulation of both

---

Sara Correia and Miguel Rocha
CCTC, University of Minho, Campus de Gualtar, Braga, Portugal
e-mail: `sarag.correia@gmail.com, mrocha@di.uminho.pt`

wild-type and mutant organisms, *Metabolic Flux Analysis* and pathway analysis using Elementary Flux Modes, among other features.

When performing strain optimization, some limitations arise from the metabolic models are incomplete [11] or the desired product can not be produced. In both cases, it will be necessary to find reactions to add to a metabolic model. In this paper, we present a new plug-in for OptFlux that allows to incorporate a set of reactions from an external database into an existing metabolic model performing phenotype simulation using those added reactions. Also, optimization methods will be put forward to allow the selection of the best set of reactions to add to the model. according to a given objective function (e.g. maximizing the production of a compound or filling gaps in the model).

## 2 Methods for phenotype simulation and strain optimization

The simulation process allows the prediction of the organism phenotype, using methods based on fundamental restrictions to the biological system. One of these methods is Flux Balance Analysis (FBA), that calculates the flux distribution making it possible to predict the growth rate of an organism or the rate of production of a metabolite, based on stoichiometric, reversibility and fluxes constraints [4]. FBA assumes that metabolic networks will reach a steady state constrained by the stoichiometry.

Predicting the metabolic state of an organism after a genetic manipulation (e.g. gene knockout) is a challenging task, because mutants are generally not subjected to the same evolutionary pressure that shaped the wild type. In these cases, other methods such *Minimization of Metabolic Adjustment* (MOMA) [12] and *Regulatory On/Off Minimization of metabolic fluxes* (ROOM) [13] are proposed to find a flux distribution for mutant strains.

Based on these methods, a question arises: how to find the ideal set of genes to be deleted to reach the desired phenotype? To try answer this question, the *OptGene* algorithm proposed by Patil et al [7] and its extensions made by Rocha et al [10] were proposed. In this last work, the authors' research group proposed a set-based representation that considered variable-sized solutions, allowing for solutions with different numbers of knockouts during the optimization process. Two optimization algorithms were developed: SA and Set-based EAs (SEAs). Both search for the optimum set size in parallel with the search for the optimum set of gene deletions.

This work aims to enlarge the set of possible genetic modifications by addressing gene additions. In this case, using SEAs or SA approaches, the optimization process finds a set of new reactions to be added to the model. Optionally, a complementary set of reactions to remove can also be optimized. Optimization methods are the same that were used previously. The main difference lies in the representation of the solutions. Although still using a representation based on sets, it is necessary to integrate information regarding the reactions to be added. Thus, a new way of rep-

resenting solutions including two independent sets (knockouts and added reactions) was created. In Figure 1, the representation of one solution is depicted.
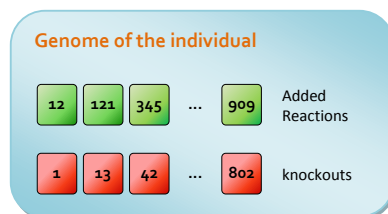

Fig. 1 Representation of the genome of an individual. Green squares represent reactions that will be added to the model (numbers are the reactions indexes in the external database). The knockouts are represented by red squares (numbers are indexes of reactions in the model).

3 OptFlux plug-in for adding reactions

A new plug-in was developed for OptFlux to allow the addition of external reactions to a metabolic model. The addition of new reactions can be made for phenotype simulation or conducting a strain optimization process. Methods to import, filter and visualize the external database of reactions are also available. The new functionalities can be accessed by the "Plugins/ Add Reactions" menu.

3.1 Import database of reactions

Importing an external database of reactions into OptFlux can be made using the same methods used for creating metabolic models (SBML [3] and flat text files). Also, a new format of text files is defined (details are in the site documentation) to allow a more flexible scheme. When using this format, the user can filter the input data files to select only reactions that satisfy some restrictions. This is useful for readability and to reduce the search space in the optimization tasks. In Figure 2, the application of two filters to a database is shown. After applying filters, the user obtains a set of reactions that will be imported to the OptFlux platform. The reaction database becomes available to use in simulation or optimization processes.
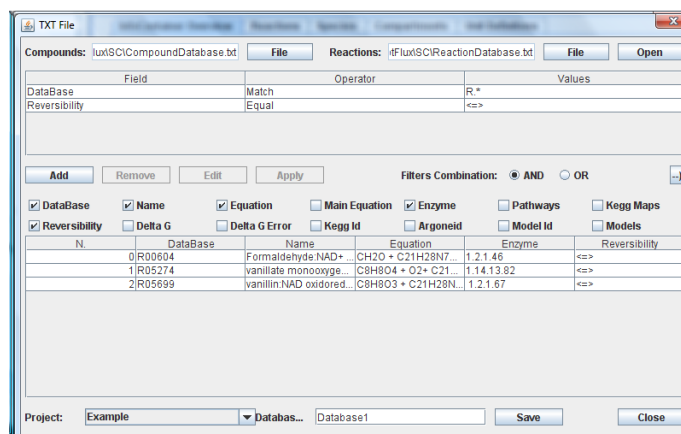
Fig. 2 Interface for selecting reactions and importing them to OptFlux. In this example, the user chooses only the reactions where ids start with "R" and that are reversible.

3.2 Mutant simulation by adding reactions

The phenotype simulation functionality allows mutant simulation by adding new reactions and optionally removing others from the model. After selecting the model to use, a previously loaded database is selected and the set of reactions to be added is chosen. Also, a set of knockouts can be selected. In Figure 3, the simulation interface is presented. During the configuration process, the user selects the simulation methods (FBA, MOMA or ROOM), the environmental conditions (the rates at which external metabolites can be consumed/ produced), and the objective function (e.g. the maximization/ minimization of a selected flux).

The result of mutant simulation can be observed in a specific interface (Figure 4), where the user can check the main results of the simulation: the list of added reactions, list of knockouts and values for all fluxes in the model.

3.3 Strain optimization by adding reactions

The strain optimization process tries to find a set of reactions to be added to the model to improve a given objective function (e.g. the production of specific product). The search can be for only a set of reactions to be added or the combination of added reactions and knockouts. In the interface (Figure 5), the user selects:

- **algorithm**: available optimization algorithms are EAs and SA;
- **simulation methods**: to be used in the simulation of each solution evaluated (FBA, MOMA or ROOM);
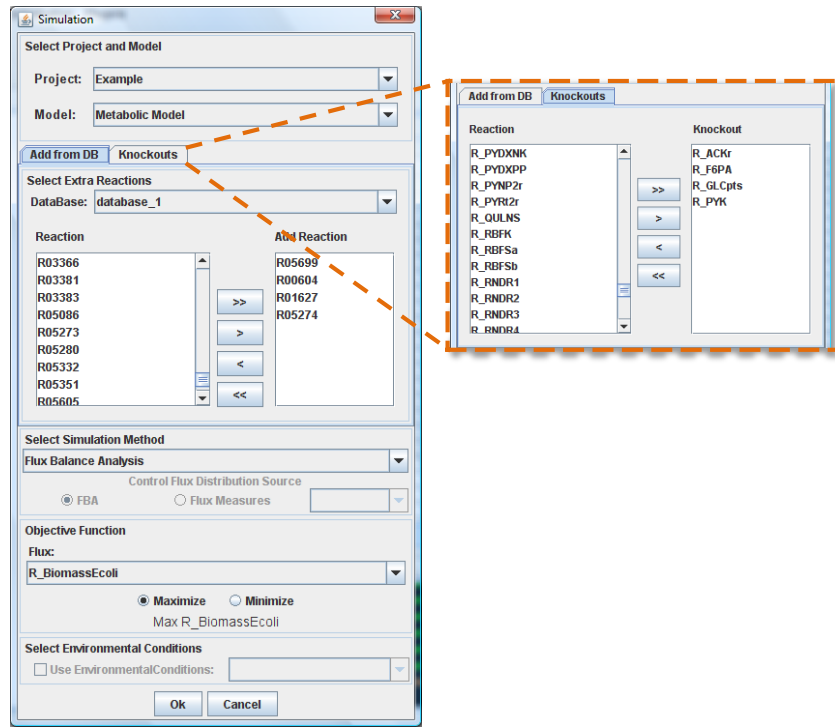
Fig. 3 Interface for mutant simulation. The case study of vanillin production is shown here (see below). In the example, 4 knockouts and 4 added reactions are selected.

- **objective function**: used to calculate the fitness value of each solution; options are the Biomass-Product Coupled Yield (BPCY) and Product Yield;
- **optimization basic setup**: configure the maximum number of solution evaluations, the maximum number of knockouts and added reactions and if the genome size should be fixed or have a variable size;
- **environmental conditions**: as defined for the simulation;
- **essential information**: define if it is possible to knockout some special type of reactions like drains, transport and critical reactions.

4 Results

4.1 Rebuilding gaps in the metabolic model

In this case study, used for validation purposes, OptFlux simplification methods were used to identify reactions constrained to a flux value of zero in the *E.coli*
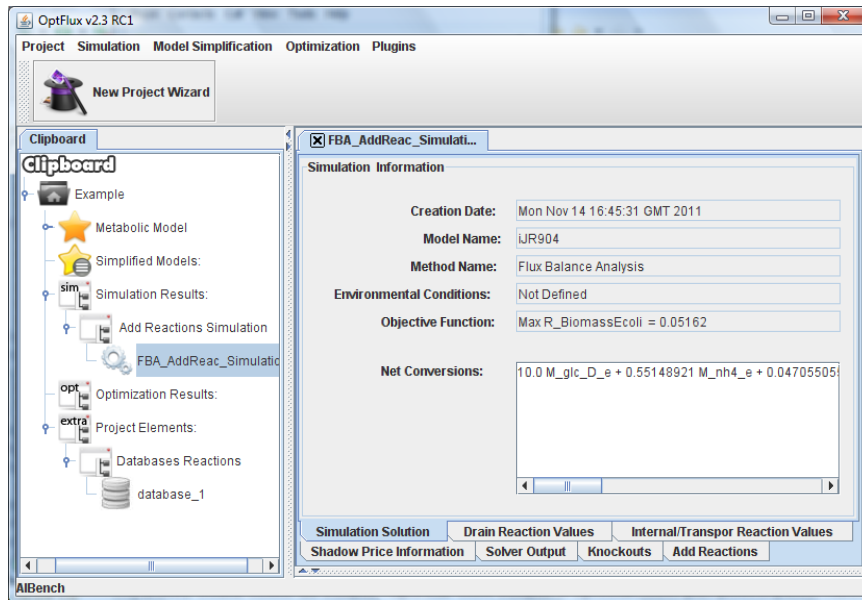
Fig. 4 Interface showing simulation results: in the left the clipboard shows the main objects and in the right side the visualization of the main results of mutant simulation are shown in distinct tabs.

model. The model is reduced eliminating those reactions and a database is created with the removed reactions (407). In each run, three randomly selected reactions are further removed from the new reduced model and inserted into the database. The optimization methods must find these reactions and re-integrate them in the model to maximize biomass production. This process was repeated 10 times for SA and EA. The number of evaluations needed to find the solution in each run are given in Table 1.

Table 1 Number of function evaluations to find the optimal solution using SA and EA.

Test reactions	EA	SA
TPI,TKT1 e TKT2	500	300
IGPS, IDOND e ENO	2060	1120
MDH, ICDHyr e CBMK	2700	2930
IPPS, HSST e GSNK	9550	1735
PANTS, P5CR e ORPT	8240	4270
ADCL, IMPD e PSERT	6750	11103
RPI, TALA e ACLS	1215	1680
ACOTA, DDPA e PFL	2020	998
PRPPS, SPMS e TRDR	1035	5302
A5PISO, RPI e TYRTA	9065	7650
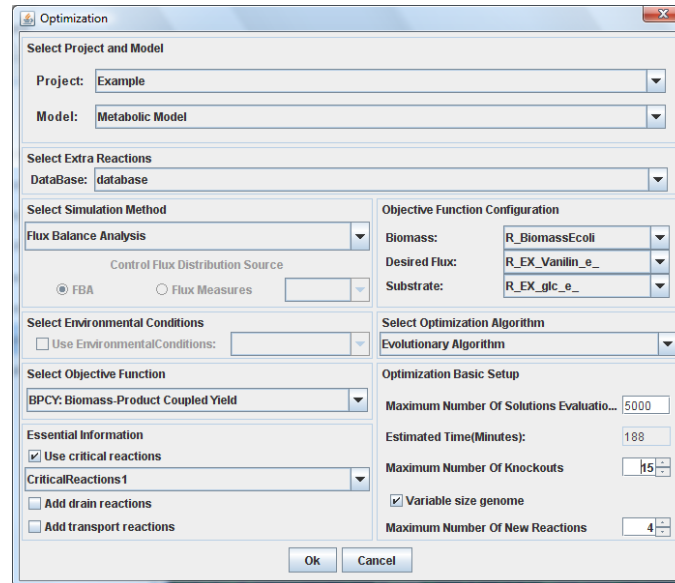
Fig. 5 Interface for strain optimization processes. In this example, an EA is configures, the simulation method is FBA, the objective function is BPCY, essential information uses the critical reactions, a maximum of 15 knockouts and 4 new reactions are permitted in variable sized sets.

4.2 Vanillin case study

This case study aims to identify new pathways for the production of vanillin from glucose in *E. coli* and validate the implemented simulation method. To demonstrate the validity of the simulation process, we used the previous study with the OptStrain framework [8]. To proceed with the test it was required to build a database of reactions to add to the metabolic model. The added reactions to the metabolic model can be observed in Figure 6.

The simulation was performed for each of the three sets of knockouts in the paper, considering the substrate flux of 10 $mmol/gDWh^{-1}$ and the objective function the maximization of biomass. FBA was used in simulation process. The obtained results agree with the one from the previous work [8], thus validating our implementation.

The next step was to run the strain optimization process to find a set of added reactions and knockouts, that maximizes vanillin production coupled with the organism growth. The process was run 30 times for each EA and SA using as objective function the Biomass-Product Coupled Yield (BPCY). Previously, it was necessary change the metabolite ids of metabolic model for those used in database.

Table 2 shows the 95% confidence interval of results obtained in the optimization process, considering the best solution from each run.

Comparing these results with the ones obtained in the previous study [8], we see that the BPCY value of their solution was 0.035 ($BPCY = (6.787 \times 0.052)/10 =$
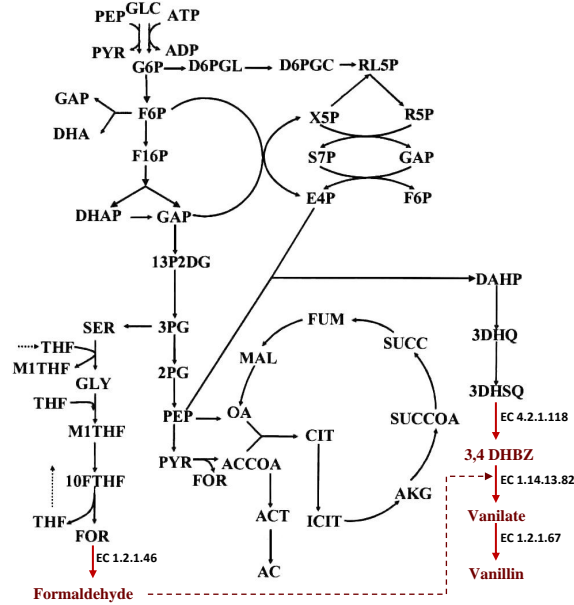
Fig. 6 The added pathway for the vanillin production.

Table 2 The 95% confidence interval of results obtained in the optimization process.

	EA	SA
Fitness (BPCY)	[0.17; 0.181]	[0.177; 0.189]
Biomass	[0.309; 0.351]	[0.323; 0.437]
Product	[5.264; 5.478]	[4.822; 5.407]
Number of knockouts	[8.78; 11.421]	[11.134; 14.332]
Number of added reactions	[8.741; 9.259]	[8.409; 9.058]

0.035). Although the vanillin production is lower in our case, the BPCY value increased significantly given that the biomass is much higher, which mean that our strain has a larger growth rate.

Afterwards, we focused in increasing the production of vanillin, without considering the biomass formation as a priority. Considering this, the tests were repeated with a new objective function, by maximizing the flux of the product, ensuring a minimum limit of biomass production (5% of the wild type value). The results shown in Table 3 contain solutions considering new pathways, where the production of vanillin is higher than the obtained in [8].

The smaller set of added reactions suggested by the optimization process include the reactions with KEGG (http://www.genome.jp/kegg) ids: R01216, R01627, R05273, R05274. A supplementary file containing the full results obtained in the experiments summarized here is given in http://darwin.di.uminho.pt/pacbb2012/.

Table 3 Best results of strain optimization for vanillin production using the Yield objective function for each algorithm (EA and SA).

	Product	Biomass	No. added reactions	No. Knockouts
	6.948	0.022	20	4
EA	6.945	0.022	19	7
	6.944	0.023	20	4
	6.948	0.022	17	6
SA	6.948	0.022	18	4
	6.948	0.022	19	5

5 Conclusion

This paper presents methods for the simulation of strains by adding external reactions to the metabolic model, aiming to produce a desired product or to fill gaps. In this approach, information is added to the stoichiometry model regarding new reactions, thus making an extension to the initial model. Methods for strain optimization were developed, using EAs and SA, to find a sets of external reactions to be added and the necessary knockouts to maximize an objective function, typically related to the production of a compound of interest.

To provide these features to the scientific community, a plug-in has been developed for the OptFlux ME platform that allows simple and intuitive phenotype simulation and strain optimization with the addition of external reactions to the metabolic model. Thus, the tool set available for ME experts has been enlarged with useful techniques.

Future work will be devoted to the validation of these methods with other real world case studies.

Acknowledgements

References

1. Thomas Bäck. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, Dortmund, Germany, 1996.
2. Jeremy S Edwards, Markus Covert, and Bernhard Palsson. Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol*, 4(3):133–40, March 2002.

3. M Hucka, A Finney, H M Sauro, and H Bolouri et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, March 2003.
4. Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. Advances in flux balance analysis. *Curr Opin Biotechnol*, 14(5):491–6, October 2003.
5. S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983.
6. J. Nielsen. Metabolic engineering. *Applied Microbiology and Biotechnology*, 55(3):263–283, 2001.
7. Kiran Raosaheb Patil, Isabel Rocha, Jochen Förster, and Jens Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308, 2005.
8. Priti Pharkya, Anthony P Burgard, and Costas D Maranas. Optstrain: a computational framework for redesign of microbial production systems. *Genome Res*, 14(11):2367–76, November 2004.
9. I. Rocha, P. Maia, P. Evangelista, P. Vilaca, S. 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 Syst Biol*, 4:45, 2010.
10. Miguel Rocha, Paulo Maia, Rui Mendes, José P Pinto, Eugénio C Ferreira, Jens Nielsen, Kiran Raosaheb Patil, and Isabel Rocha. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics*, 9:499, 2008.
11. V. Satish Kumar, M. S. Dasika, and C. D. Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8:212, 2007.
12. Daniel Segrè, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112–7, November 2002.
13. Tomer Shlomi, Omer Berkman, and Eytan Ruppin. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21):7695–700, May 2005.