A study on the robustness of strain optimization algorithms

Paulo Vilaça, Paulo Maia, and Miguel Rocha

Abstract In recent years, there have been considerable advances in the use of genome-scale metabolic models to provide accurate phenotype simulation methods, which in turn enabled the development of efficient strain optimization algorithms for Metabolic Engineering. In this work, we address some of the limitations of previous studies regarding strain optimization algorithms, mainly its use of Flux Balance Analysis in the simulation layer. We perform a thorough analysis of previous results by relying on Flux Variability Analysis and on alternative methods for phenotype simulation, such as ROOM. This last method is also used in the simulation layer, as a basis for optimization, and the results obtained are also the target of thorough analysis and comparison with previous ones.

1 Introduction

The recent advances on genome sequencing techniques have led to the knowledge of the complete genetic information of a large number of organisms over the last few years. Together with the development of novel methods in the fields of Bioinformatics and Systems Biology, this data allowed, among many other applications, the reconstruction of genome-scale metabolic models for some organisms [6], mostly microbes with an industrial interest in Biotechnology. Within Metabolic Engineering (ME) [4], one of the applications of these models is to allow the simulation of the phenotype of these microbes, under different environmental conditions (e.g, nutrients, aerobic/ anaerobic conditions). Also, it is possible to predict the phenotypes of mutant strains (e.g. gene knockouts). In fact, several distinct constraint-based methods have been developed that rely only on the information about the metabolic ca-

Paulo Vilaça, Paulo Maia
CEB-IBB / CCTC, University of Minho, Portugal

Miguel Rocha
CCTC, University of Minho, Portugal e-mail: mrocha@di.uminho.pt

pacities of an organism to reach phenotype predictions. Among those, Flux Balance Analysis (FBA) [2] and Regulatory on-off minimization of metabolic flux changes (ROOM) [10] have reached a remarkable level of success.

The combination of reliable models with efficient simulation methods has been the basis for different strain optimization algorithms. Their goal is to find the set of genetic modifications to apply to a given strain, to achieve an aim, typically related with the industrial production of a metabolite of interest. Indeed, one of the major recent trends in industry has been the replacement of traditional industries (e.g. chemical industry) by Biotechnology, as a way to produce numerous important products, but this typically requires to retrofit the original strain. In previous work, an approach based in the use of metaheuristics, such as Evolutionary Algorithms and Simulated Annealing, has been proposed to solve the optimization task of reaching an optimal (or near optimal) subset of reaction deletions to optimize an objective function related with the production of a given compound [9]. The idea is to force the microbes to synthesize a desired product, while keeping it viable.

While good results have been obtained, there are still some limitations that need to be addressed. Some of those are related to the fact that all previous work has relied on the use of FBA to provide the phenotype simulation layer. This brings two major problems: (i) FBA relies on solving a constraint-based optimization problem that is formulated using linear programming (LP). However, it considers only one optimal solution, while the problem can have alternative optimal solutions. Thus, the phenotype taken as the result is only one of the possible alternatives. This can have an impact on the optimization results. (ii) Other methods have been proposed for phenotype simulation, claiming better results when simulating mutant strains (e.g. ROOM). However, these methods have not yet been used as a basis for strain optimization, mainly due to their computational burden.

In this work, the aim is to shed a new light over strain optimization, by addressing two tasks: (i) to re-analyse some previously published results, checking the robustness of the solutions found, under two perspectives: checking the impact of the multiple optima issue (using Flux Variabilty Analysis) and simulating those solutions with another method (ROOM); (ii) use ROOM as the mutant phenotype prediction method within strain optimization and comparing the results with those previously obtained with FBA. The ultimate goal will be to gain an insight on these approaches that allows to improve the robustness of the underlying algorithms.

2 Methods

2.1 Flux-balance and flux variability analysis

The Flux Balance Analysis (FBA) [2] approach is based on a steady state approximation to the concentrations of internal metabolites, which reduces the corresponding mass balances to a set of linear homogeneous equations. For a network of M

metabolites and N reactions, this is expressed as: $\sum_{j=1}^{N} S_{ij} v_j = 0$, where S_{ij} is the stoichiometric coefficient for metabolite i in reaction j and v_j is the flux over the reaction j. The maximum/minimum values of the fluxes can be set by additional constraints in the form $\alpha_j \leq v_j \leq \beta_j$, also used to specify nutrient availability.

The set of linear equations obtained usually leads to an under-determined system, for which there exists an infinite number of feasible flux distributions that satisfy the constraints. However, if a given linear function over the fluxes is chosen to be maximized, it is possible to obtain a single solution by applying standard algorithms (e.g. *simplex*) for LP. The most common flux chosen for maximization is the biomass, based on the premise that microorganisms have maximized their growth along natural evolution, a premise that has been validated experimentally [1].

Flux Variability Analysis (FVA) is a technique that also relies on LP, exploring the space of all possible solutions that comply to a given set of constraints. The idea is to calculate the limits for a given flux in the model, given the set of constraints as in FBA. To explore the space of possible values of a flux within the space of optimal solutions of an FBA instance, the following steps are executed: (i) run the LP as before, maximizing the biomass flux (FBA); (ii) add a constraint stating biomass is greater or equal to the value reached in FBA; (iii) run two LP problems maximizing and minimizing the target flux. In this work, the minimization of the target flux will be used, since this provides a worst-case scenario for the desired product.

2.2 Regulatory on-off minimization of metabolic flux changes

An alternative to FBA for the phenotype simulation is the Regulatory on-off minimization of metabolic flux changes (ROOM) method. ROOM is appropriate only for the simulation of mutants, since it calculates the solution with minimum number of significant changes in the value of the fluxes from the mutant strain, relative to the original wild-type solution (obtained with FBA). The method is implemented based on a mixed integer linear programming (MILP) formulation. The full details on the mathematical formulation can be found in the original paper [10]. The authors provide experimental evidence of the better accuracy of this method for the phenotype prediction of knock-out mutants.

2.3 Simulated annealing for strain optimization

The problem addressed in this work consists in selecting, from a set of reactions in a microbe's genome-scale model, a subset to be deleted to maximize a given objective function. The encoding of a solution is achieved by a variable size set-based representation, where each solution consists of a set of integer values representing the reactions that will be deleted, with a value between 1 and N, where N is the number of genes in the model. For all reactions deleted, the flux will be constrained

to 0, therefore disabling it from the metabolic model. The process proceeds with the simulation of the mutant using the chosen phenotype simulation method (FBA or ROOM). The output of both methods is the set of fluxes for all reactions, that are then used to compute the fitness value, given by an appropriate objective function. The objective function used is the Biomass-Product Coupled Yield (BPCY) [5], given by: $BPCY = \frac{PG}{S}$, where P stands for the flux representing the excreted product; G for the organism's growth rate (biomass flux) and S for the substrate intake flux. Besides optimizing for the production of the desired product, this function also allows to select for mutants that exhibit high growth rates. To address this task, we will use Simulated Annealing (SA) as proposed previously in [9], where full details can be found.

3 Experiments and results

3.1 Case studies and experimental setup

The implementation of the proposed algorithms was performed by the authors in *Java*, within the OptFlux open-source ME platform (http://www.optflux.org) [8].

Two case studies were considered, both considering the microorganism *Escherichia coli*. The aim is to produce succinate and lactate with glucose as the limiting substrate. The lactate is split into aerobic and anaerobic conditions, i.e. allowing (or not) the uptake of oxygen from the media. Succinate is one of the key intermediates in cellular metabolism and therefore an important case study for ME [3]. It has been used to synthesize polymers, as additives and flavouring agents in foods, supplements for pharmaceuticals, or surfactants. Lactate and its derivatives have been used in a wide range of food-processing and industrial applications like meat preservation, cosmetics, oral and health care products. The genome-scale model used is given in [7] and includes a total of $N = 1075$ fluxes and $M = 761$ metabolites. A number of pre-processing steps were conducted to simplify the model and reduce the number of targets for the optimization (see [9] for details) leaving the simplified model with $N = 550$ and $M = 332$; 227 essential reactions are identified, leaving 323 variables to be considered when performing strain optimization.

3.2 Results

3.2.1 Re-analysing solutions from FBA

The first task was to consider a large set of solutions for strain optimization problems, obtained using FBA as the phenotype simulation method. These solutions were analysed by running FVA for the target product flux, thus addressing the is-

sue of their robustness to multiple optima in the LP solutions. The set of solutions analysed came from three sources: experiments run for this study and previous results obtained by the authors in [9] and [11]. The selected set of solutions was simulated using FVA, minimizing the target flux (succinate or lactate production). This provides the minimum predicted production value that can be obtained by the mutant. As a measure of robustness, the value of maximum loss was calculated, taking into account the original FBA value (used in the optimization to evaluate the solution), here denoted as $FBAProdValue$, and the minimum limit calculated by the FVA for the product flux, denoted as $FVAMinValue$: $MaxLoss = (FBAProdValue - FVAMinValue)/FBAProdValue$

Table 1 Results for the FVA analysis.

Case study	Succinate (aerobic)	Lactate (aerobic)	Lactate (anaerobic)
Number solutions	65	77	48
Mean MaxLoss	0.2%	92.0%	92.5%
Mean FBAProdValue	5.82	15.97	17.35
Mean FVAMinValue	5.81	1.20	0.96
$FVAMinValue < 25\%\ FBAProdValue$	0%	92%	94%
$FVAMinValue > 75\%\ FBAProdValue$	100%	8%	6%

Table 1 summarizes the results obtained for the 3 case studies. The first row shows the number of solutions analysed, then the mean values for the $MaxLoss$, $FBAProdValue$ and $FVAMinValue$ are shown and the last two rows show the percentage of solutions where the value is smaller than 25% of the FBA value and larger than 75%. The results show a huge difference between the two case studies. In fact, solutions for succinate production optimization seem very robust; indeed, all solutions have a $MaxLoss$ of less than 5% and more than 95% of the solutions have a value of zero. This means that, in this case, FBA does not have alternative optimum solutions that can lower the product value significantly. On the other hand, in the lactate case studies, the scenario is the reverse. In fact, more than 90% of the solutions analysed have a drop of 100% or very near, which means that the great majority of the solutions are not robust, existing alternative solutions where the production of the target metabolite is very low (or even non existent in many cases).

The next step was to take each solution (reaction deletion list) and perform the simulation of the respective mutant using the ROOM algorithm. The aim was to check if the results obtained were near or if there were significant differences. For each solution and each method (FBA and ROOM) the values obtained for the biomass flux and for the target product flux were collected. As a measure of the deviation between both methods, the relative differences were calculated, by subtracting the two values (FBA and ROOM) and dividing by the original FBA value. This process was repeated both for the biomass and product fluxes. Table 2 shows the results of these experiments. These show that the values obtained by FBA and ROOM are generally in agreement in two of the cases: succinate and lactate (anaer-

obic), but are very distant in the lactate case study with aerobic conditions. In the first two, the solutions seem to be robust to the phenotype simulation methods, while in the latter the results are quite different.

Table 2 Results for the analysis of solutions obtained using FBA in the optimization, simulated now with ROOM.

Case study	Succinate (aerobic)	Lactate (aerobic)	Lactate (anaerobic)
Mean relative diff. biomass	-27.0%	-95.2%	-11.3%
Mean relative diff. product	+0.6%	-93.6%	-7.4%
Mean biomass flux (FBA)	0.575	0.195	0.144
Mean biomass flux (ROOM)	0.427	0.017	0.128
Mean product flux (FBA)	5.82	15.97	17.36
Mean product flux (ROOM)	5.83	0.91	15.71

3.2.2 Using ROOM for strain optimization

A natural follow-up of the previous experiments is to use the ROOM algorithm as the phenotype simulation method within strain optimization algorithms. This task was addressed here, although with some limitations given the high computational demands, since MILP problems needed by ROOM are harder to solve than the LP used in FBA. SA was the optimization algorithm chosen for the job and the configuration proposed in [9] was kept. The termination criteria was based on 50000 fitness evaluations. For each configuration, the process was repeated for 10 runs, given the computational constraints. Also, based on the results of the previous section, experiments were only run for two case studies: the succinate and the lactate (anaerobic). The same set of experiments was also done with FBA as the simulation method to enrich the comparative analysis.

The main results for the optimization with both methods are provided in Table 3. From this table, we see that the results are quite comparable being within the same range of values in most cases. Also, we decided to conduct a robustness analysis for the ROOM results, similar to the one conducted in the previous section. Therefore, we re-analysed the solutions using FVA and also simulating with FBA. The metrics used are similar to the ones defined above (reversing the roles of ROOM and FBA), and the results are given in Tables 4 and 5, respectively. From those tables, we can conclude that, unlike the previous section, the results on the lactate case study now seem much more robust in both FBA and FVA analysis. This shows that the optimization using the ROOM phenotype simulation approach leads the optimization algorithm to very different solutions in both case studies. Also, it is also clear that it is not easy to know *a priori* what is the best optimization algorithm for a given task.

The full results of this study can be checked in two files given as supplementary material available in the site: http://www.optflux.org/suppmaterial.

Table 3 Results for the the optimization using with ROOM compared with simulation using FBA.

Simulation Method	Case Study	Number Knockouts	Mean BPCY	Mean Biomass	Mean Product
ROOM	Succinate	3	0.146	0.706	2.26
ROOM	Succinate	6	0.301	0.509	5.90
ROOM	Succinate	12	0.321	0.527	6.09
ROOM	Lactate (anae.)	3	0.153	0.152	10.59
ROOM	Lactate (anae.)	6	0.229	0.146	15.81
FBA	Succinate	3	0.059	0.859	0.693
FBA	Succinate	6	0.235	0.689	3.74
FBA	Succinate	12	0.340	0.539	6.33
FBA	Lactate (anae.)	3	0.153	0.162	10.61
FBA	Lactate (anae.)	6	0.204	0.153	14.16

Table 4 Results for the FVA analysis.

Case study	Succinate (aerobic)	Lactate (anaerobic)
Mean MaxLoss	63.3%	0.0%
Mean ROOMProdValue	4.75	13.20
Mean FVAMinValue	1.28	13.20
$FVAMinValue < 25\%\ ROOM_{P}rodValue$	16%	0%
$FVAMinValue > 75\%\ ROOM_{P}rodValue$	27%	100%

Table 5 Results for the analysis of solutions obtained using ROOM in the optimization, simulated now with FBA.

Case study	Succinate (aerobic)	Lactate (anaerobic)
Mean relative diff. biomass	13.1	-10.2%
Mean relative diff. product	-63.3%	21.8%
Mean biomass flux (ROOM)	0.580	0.149
Mean biomass flux (FBA)	0.655	0.134
Mean product flux (ROOM)	4.75	13.2
Mean product flux (FBA)	1.28	13.0

4 Conclusions

In this work, we addressed the issue of robustness in strain optimization algorithms by re-analysing previous results with alternative simulation methods. The results show that this is an important question to address, since for many of the previous results, the solutions do not seem to be robust when other simulation methods are used. Thus, it is highly recommended that this type of analysis is conducted as a post-processing step of strain optimization methods. This work lays the basis to create a workflow for this task, although this still needs the be further refined in

the future. Also, the first results for strain optimization algorithms using a method alternative to FBA (in this case, ROOM) were provided, being the first study that conducts this type of research. The results show that there is no rule stating which is the best method to use and, in practice, the best alternative is to use more than one alternative and perform a careful post-processing of the results.

In further work, the development of methods that can incorporate the robustness of the solutions within the evaluation function of the metaheuristics will be explored. Although this can increase the computational effort of the algorithms it can be an alternative worth to be explored.

Acknowledgements This work is supported by Portuguese FCT - project MIT-PT/BS-BB/0082/2008.

References

1. R.U. Ibarra, J.S. Edwards, and B.G. Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420:186–189, 2002.
2. K.J. Kauffman, P. Prakash, and J.S. Edwards. Advances in flux balance analysis. *Curr Opin Biotechnol*, 14:491–496, 2003.
3. S.Y. Lee, S.H. Hong, and S.Y. Moon. In silico metabolic pathway analysis and design: succinic acid production by metabolically engineered escherichia coli as an example. *Genome Informatics*, 13:214–223, 2002.
4. J. Nielsen. Metabolic engineering. *Appl Microbiol Biotechnol*, 55:263–283, 2001.
5. K. Patil, I. Rocha, J. Forster, and J. Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6(308), 2005.
6. K.R. Patil, M. Akesson, and J. Nielsen. Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotechnol*, 15:64–69, 2004.
7. J.L. Reed, T.D. Vo, C.H. Schilling, and B.O. Palsson. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biology*, 4(9):R54.1–R54.12, 2003.
8. I. Rocha, P. Maia, P. Evangelista, P. Vilaa, 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 Systems Biology*, 4(45), 2010.
9. 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*, 9, 2008.
10. T. Shlomi, O. Berkman, and E. Ruppin. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *PNAS*, 102(21):7695–7700, 2005.
11. P. Vilaça, P. Maia, I. Rocha, and M. Rocha. Metaheuristics for strain optimization using transcriptional information enriched metabolic models. In Clara Pizzuti, Marylyn D. Ritchie, and Mario Giacobini, editors, *EvoBIO*, volume 6023 of *LNCS*, pages 205–216. Springer, 2010.