Identification of Robust Strain Designs via Tandem pFBA/LMOMA phenotype prediction

Paulo Maia
SilicoLife, Lda
Braga, Portugal
pmaia@silicolife.com

Isabel Rocha
Centre Biological Engineering,
University of Minho
Braga, Portugal
irocha@deb.uminho.pt

Miguel Rocha
Centre Biological Engineering,
University of Minho
Braga, Portugal
mrocha@di.uminho.pt

ABSTRACT

The past two decades have witnessed great advances in the computational modeling and systems biology fields. Soon after the first models of metabolism were developed, methods for phenotype prediction were put forward, as well as strain optimization methods, within the field of Metabolic Engineering. Evolutionary computation has been on the front line, with the proposal of bilevel metaheuristics, where EC works over phenotype simulation, selecting the most promising solutions for bioengineering tasks.

Recently, Schuetz and co-workers proposed that the metabolism of bacteria operates close to the Pareto-optimal surface of a three-dimensional space defined by competing objectives. Albeit multi-objective strain optimization approaches focused on bioengineering objectives have been proposed, none tackles the multiobjective nature of the cellular objectives. In this work, we propose multi-objective evolutionary algorithms for strain optimization, where objective functions are defined based on distinct phenotype prediction methods, showing that those can lead to more robust designs, allowing to find solutions in more complex scenarios.

CCS CONCEPTS

•Theory of computation → Mathematical optimization; Evolutionary algorithms; Linear programming; *Quadratic programming;* **•Applied computing → Computational biology; Biological networks; Systems biology;**

KEYWORDS

Systems Biology, Evolutionary Algorithms, Metabolic Networks, Multiobjective optimization

ACM Reference format:
Paulo Maia, Isabel Rocha, and Miguel Rocha. 2017. Identification of Robust Strain Designs via Tandem pFBA/LMOMA phenotype prediction. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 8 pages.
DOI: http://dx.doi.org/10.1145/3067695.3082542

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3082542

1 INTRODUCTION

The concept of metabolic pathway manipulation towards desirable behavior is not new. Early methods relied mostly on the use of mutagenesis and strain selection [19]. However, with increasingly demanding industrial requirements, the need to resort to rational approaches became evident. The development of genetic engineering brought ways to more precisely modify specific genes/enzymes, thus paving the way towards the more rational introduction of direct genetic changes to create desirable strains [11].

Moreover, the recent advances in genome sequencing technologies which culminated in the development of next generation sequencing technologies [16] and semi-automated annotation techniques, made the availability of a large number of fully annotated microbial genomes a reality. This also accelerated the development of genome-scale metabolic models (GSMMs) for a large number of organisms [5]. The development of phenotype prediction methods supporting distinct genetic and environmental conditions, including the well-known method of Flux Balance Analysis [7, 14, 15], combined with GSMMs, brought powerful tools to predict the behavior of microbial strains and support rational ME efforts.

Backed by these efforts, the development of strain design methods, where bioengineering objectives could be rationally addressed, became paramount. In 2003, OptKnock was proposed [3], becoming the basis for a large number of constraint-based strain design methods. These approaches are able to propose genetic changes based on computational simulation and optimization methods. While these approaches have provided good results, they are still limited since they usually return a single solution to the problem. Among all, meta-heuristic CSOMs, mainly those based on Evolutionary Computation (EC) [13] provide the most diverse solutions, but typically those follow similar strategies to maximize the selected objective function. To overcome these limitations, information from multiple criteria is often included in a single objective function, which can introduce undesired biases in the sampling process. Multi-objective (MO) approaches search for optimal trade-offs of solutions instead of a single optimal solution, thus providing a valuable tool for expert researchers, allowing them to opt for compromise solutions believed to have better chances of working *in vivo*.

An analysis of available CSOMs reveals several shortcomings of the current methods. As an example, defining an objective function for a CSOM can be a difficult task. Because of this, and since models lack critical information to improve the quality of the predictions, the solutions proposed by most CSOMs are not only overly-optimistic, but sometimes physiologically impossible.

Indeed, assumptions regarding the cellular objectives of an organism when subjected to distinct conditions (environmental, genetic,

etc.) are still the object of active discussion. The most common approach is to consider the cell to be in a pseudo steady-state and, since the solution space for the metabolic fluxes of the cell is usually very large, constraint-based optimization approaches are often applied for simulating metabolic fluxes. Given this assumption, it is therefore plausible to predict cellular behaviour by solving optimization problems, as long as biologically realistic objective functions are put forward. Several methods have been developed following these trends. Among these, Flux Balance Analysis (FBA) [8] is the most widely used phenotype prediction algorithm, that uses a linear programming (LP) formulation for the maximization of growth (synthesis of biomass constituents) as the objective function, considering the biological assumption that unicellular organisms tend to maximize their growth as an evolutionary trend [6].

However, to predict the cellular behaviour of mutant organisms, such assumption is not widely accepted and, for that purpose, other methods have been proposed such as Minimization of Metabolic Adjustment (MOMA) [17] based on Quadratic Programming (QP), where the objective function is the minimization of flux variations relative to the wild-type. The hypothesis underlying MOMA is that fluxes in a perturbed cell (e.g. a mutant) will be redistributed in order to be as similar as possible to the wild-type [2].

In this work, we focused on variations of two of the most widely used phenotype prediction methods, the parsimonious enzyme usage FBA (pFBA) (a variation of FBA that minimizes the overall sum of enzyme-associated fluxes [9]) and LMOMA (a linear implementation of MOMA [1]). We analyze the influence of the simulation methods on the results of strain optimization metaheuristic algorithms and suggest a multi-objective approach capable of finding designs compliant with the cellular objectives assumed by the various phenotype prediction methods.

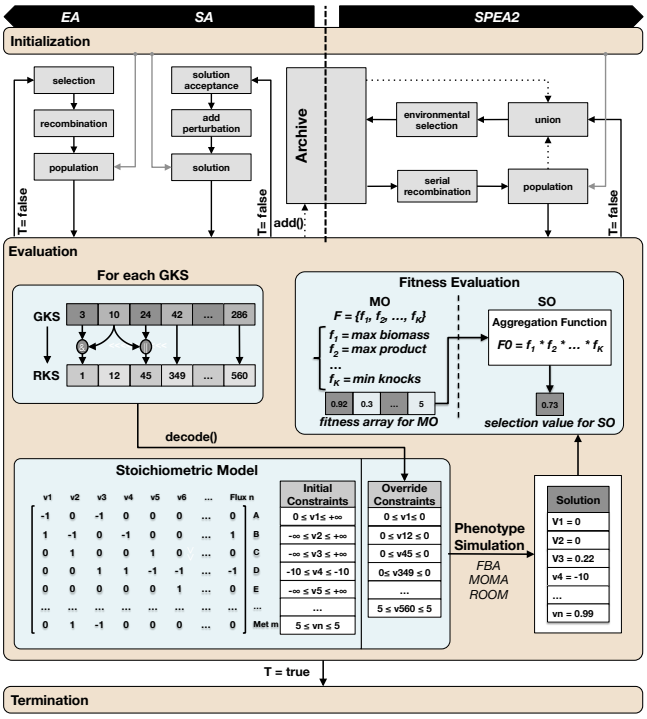

Figure 1: Overview of the developed algorithms. The upper region shows the major steps of the three algorithms. The evaluation box illustrates the processes of solution decoding, from Gene Knockout Sets (GKS) to Reaction Knockout Sets (RKS) (upper-left), phenotype prediction showing the added constraints (bottom) and fitness evaluation for both MO and SO cases (upper-right).

2 METHODS

In previous work by the authors, Evolutionary Algorithms (EA) and Simulated Annealing (SA) have been proposed to address strain optimization problems, selecting (near-)optimal sets of genes/ reactions to delete from a model, to overproduce a given compound, where both used the same variable size set-based representation [13]. Two types of reproduction operators were used: crossover (EA only) and mutation (both EA and SA). The first is inspired on *uniform crossover* and, regarding mutation, three operators were used: *random mutation*, *grow mutation* and *shrink mutation*. The details of both algorithms are depicted in Figure 1, and their full configuration can be obtained in [10].

In the first part of this work, the two metaheuristics (EAs and SAs) were executed using both pFBA and LMOMA as the phenotype prediction method. The output of the phenotype prediction is the set of flux values for all reactions in the model. These are used to compute the fitness value of the solution, using the Biomass-Product Coupled Yield (BPCY) [12] as objective function, given by $\frac{PB}{S}$, where P stands for the flux of the desired product; B for the biomass flux and S for the substrate uptake flux. In the EAs, the population size was set to 100 individuals. For analysis purposes, the resulting solution sets for the EA and SA algorithms were merged for each set of conditions. However, the convergence analysis is done separately (algorithm-dependent).

In the second part of this work, a multi-objective mechanism, capable of searching for genetic designs compliant with two or more phenotype prediction methods, was devised. In this work, the SPEA2 [21] was used following the structure depicted in Figure 1. SPEA2 uses an external archive that contains non-dominated solutions (called the external non-dominated set). At each generation, non-dominated individuals are copied from the population to this external set. For each individual in the archive, a strength value, proportional to the number of solutions in the archive it dominates, is computed. The fitness of each individual in the current population is computed according to the strengths of all external individuals that dominate it. This strategy is used to promote the convergence of the algorithm. The fact that the external non-dominated set is used in the selection process brings the problem that, if that set grows too much, the selection pressure might be reduced, thus slowing down the global search process. To prevent this, a clustering technique called "average linkage method" was adopted to prune the external non-dominated set, thus maintaining diversity.

For all the algorithms, each individual candidate solution encodes a set of identifiers for metabolic genes whose activity should be suppressed (knocked-out) from the GSMM. In the GSMM, this information is made available by means of Gene-Protein-Reaction

(GPR) associations which resorts to Boolean logic, where the relationships between reactions and their encoding genes are modeled as logical and/ or operations representing, among others, cases of protein complexes and isoenzymes, thus allowing, for instance, determination of the reactions inactivated after a set of gene deletions. In practice, each set of gene deletions encoded in a candidate solution is translated into the corresponding reaction deletions, which in turn are set as override constraints whose lower and upper limits are set to zero in the original model, thus simulating the effect of a gene deletion.

The configuration for SPEA2 follows the one used by the EA, using the same operators and termination criterion (other details are provided in [10]). The population and archive sizes were set to 100 individuals. The main difference concerns the evaluation of the solutions. In this case, each solution is decoded as before and simulated independently using the selected phenotype prediction methods, which in the experiments will be two: pFBA and LMOMA. These originate two distinct flux distributions which will be evaluated using BPCY. These two values, BPCY-pFBA and BPCY-LMOMA, make the two objective functions used by SPEA2.

Using a recent model of *Escherichia coli* K12 (iAF1260) [4], the experiments were setup considering two case studies for the production of lactate and succinate from glucose in aerobic conditions. For each algorithm, the execution was halted after 50000 function evaluations and the process was repeated 30 times.

3 RESULTS

3.1 Effects of phenotype prediction methods over strain optimization

A summary of the number of solutions generated by each of the algorithms in the first part is presented in Table 1. Only solutions where $BPCY \geq 1 \times 10^{-5}$ are shown. Note that this was a criterion used to take into account solutions where both the biomass and target compound fluxes are larger than zero.

Table 1: Size of the merged solution sets (EAs and SAs). Target products: Lac. - lactate, Suc. - succinate.

	Target Product	
Method	*Lac.*	*Suc.*
pFBA	292	143
LMOMA	661	1187

In an initial analysis of this table, the larger number of solutions reachable when LMOMA is the used prediction method is easily observable. This seems to imply that the space of BPCY-valid solutions is larger when LMOMA is used in the simulation, leading the algorithms to more rapidly finding interesting solutions.

To understand how different phenotype prediction methods affect the solutions reached, the convergences of the EA and SA, when using pFBA and LMOMA were analyzed separately. Figure 2 depicts the convergences in the two case studies.

In a first observation of the convergence plots, a smoother convergence when the LMOMA phenotype prediction method is being used becomes evident. When pFBA is the selected method, the convergence evolves in a stepped pattern, with no observable change in

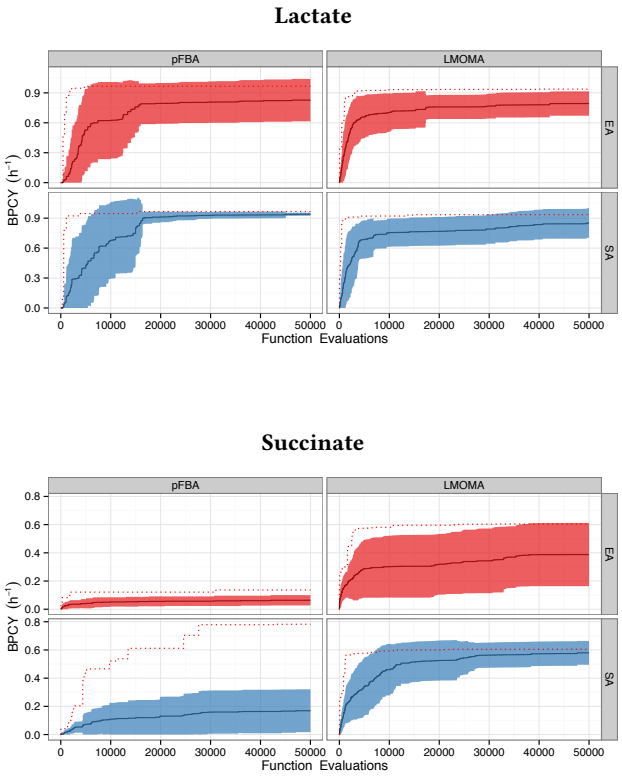

Figure 2: Convergence plots for the EA and SA algorithms applied to the production of lactate and succinate. The solid lines indicates the means of 30 runs, while the color-shaded areas indicate the standard deviation. The dashed lines represent the maximum value for each algorithm and problem.

fitness for several evaluation functions and larger fitness jumps in some steps. The LMOMA pattern is a smoother one, represented by slight but constant increases in fitness until convergence is attained. In the easier case study, lactate, these differences are not so easily observable, while in the more difficult one, succinate, this trend becomes evident. These trends are extensible to the additional case studies (in supplementary material) where, in some cases, these patterns are even more declared.

To evaluate the effect that these differences had in the phenotype (flux values) of the solutions reached by the algorithms with each of the phenotype prediction methods, an analysis on the flux distribution of such solutions was devised. A wild-type flux distribution was predicted using pFBA and taken as the reference flux distribution. The distribution of flux distances from the mutant phenotypes to this reference was then computed (Figure 3). To meet this end, the Jaccard distance for asymmetric binary attributes (d_J) was employed:

$$d_J = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}} \quad (1)$$

where M_{01} represents the total number of fluxes active in the mutant, but inactive in the wild-type; M_{10} is the total number of fluxes active in the wild-type but not in the mutant; and, M_{11} is the number of fluxes that are active in both the mutant and the wild type flux distributions. This metric only considers the flux differences as a binary array (on or off), thus ignoring the effective flux levels.

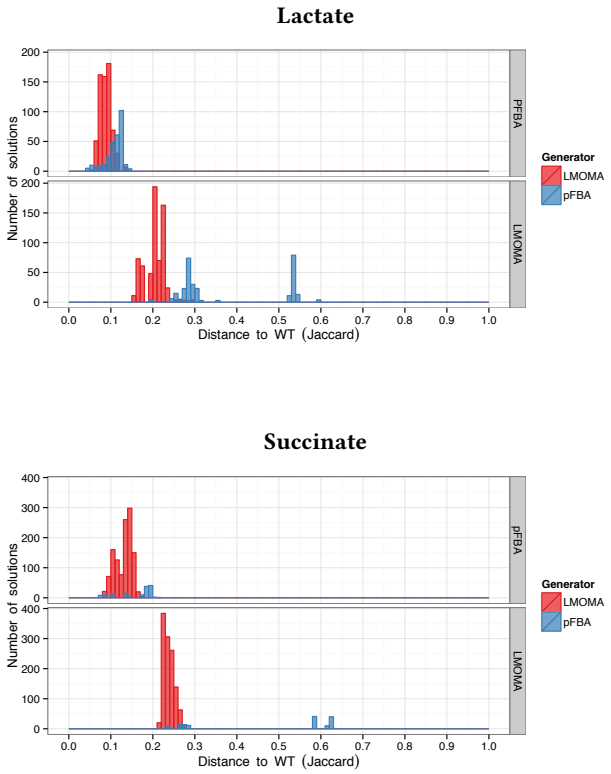

Figure 3: Distribution of the Jaccard distances from the solutions to the wild-type flux distributions for the production of lactate and succinate. Red bins and blue bins represent solutions generated by LMOMA and pFBA, respectively. Solutions were re-simulated with pFBA (top in each plot) and LMOMA (bottom in each plot).

Every solution generated by the EA and SA, while using LMOMA (LMOMA-generated) was re-simulated using pFBA (top histogram in each chart), while every solution generated using pFBA (pFBA-generated) was re-simulated using LMOMA (bottom histogram in each chart). By visually inspecting the histograms, some observations are possible:

(1) Overall, the solutions simulated by LMOMA are usually farther from the wild-type than the ones simulated by pFBA;
(2) When re-simulated with pFBA, the LMOMA-generated solutions, are generally closer to the wild-type than the pFBA ones;
(3) When re-simulated with LMOMA, the pFBA-generated solutions are generally much farther from the wild-type than the LMOMA ones. Some of them even have a $d_J > 0.5$.

These facts can be dissected and analyzed in more detail. The formulation of the pFBA procedure helps explaining the observation 1 given that, for a given flux space that maximizes biomass, it will return the flux distribution that minimizes the overall sum of fluxes. On the other hand, the fact that LMOMA solutions are closer to the wild-type than pFBA ones, even when simulated with pFBA (observation 2), can be attributed to the fact that the LMOMA objective function tries to minimize the distance between the wild-type and the mutant flux distributions (i.e., there is a bias in the LMOMA optimization towards this objective). This is important, because we assume that solutions that are closer to the wild-type are more likely to work in reality [17]. This is observable for simulations with pFBA and LMOMA. On the other hand, the solutions simulated with pFBA are generally closer to the wild-type, which can be attributed to the pFBA objective function that minimizes the overall sum of fluxes. This means that pFBA simulations are probably closer to other phenotype prediction methods such as ROOM [18] than LMOMA. Finally, in observation 3, the pFBA-generated solutions are clearly modifying a higher number of fluxes when they are simulated with LMOMA. This can be explained by the fact that the pFBA procedure does not have any bias towards flux distributions that are closer to the wild-type.

Another question that quickly arises is how many solutions, generated by each of the methods, are actually valid when simulated with the other. Venn-like diagrams are presented in Figure 4 to provide a first answer to this question.

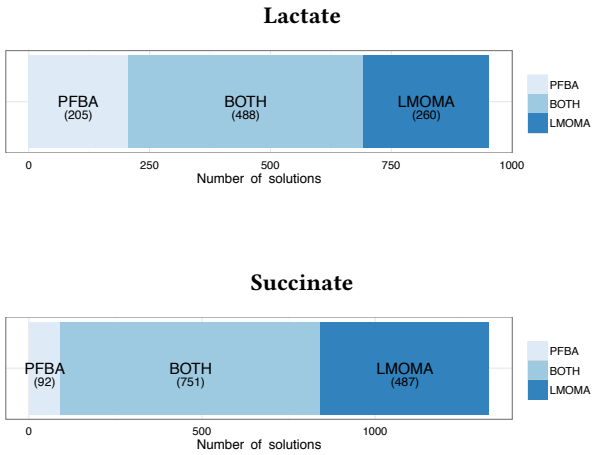

Figure 4: Venn-like diagrams for checking the inter-method validity of the solutions. From left to right, solutions that are BPCY-valid for: only pFBA, both pFBA and LMOMA, and only LMOMA.

It is easily observable that a large number of solutions (about half) are not BPCY-valid for both methods. From these, the majority are BPCY-valid for LMOMA only. This confirms that the

phenotype prediction method is a determining factor not only on the performance of the strain optimization algorithms, but also in the sets of solutions they yield. If we assume that the likelihood of these solutions working in reality increases if they are valid using different phenotype prediction methods methods, then most of the solutions found are not robust.

While it is also clear that, in these case studies, there is a good set of solutions that are BPCY-valid for both methods, the quality of these solutions is not addressed, since this analysis includes solutions whose BPCY values are close to zero ($\geq 1 \times 10^{-5}$).

To better understand how the fitnesses vary as a function of the phenotype prediction methods, Figures 5 and 6 are put forward, where the BPCY-values are taken into account. In Figure 5 , solutions generated by the strain optimization algorithms when one of the phenotype prediction methods was used were re-simulated with the other, their BPCY values were calculated for both cases and represented in the form of boxplots.

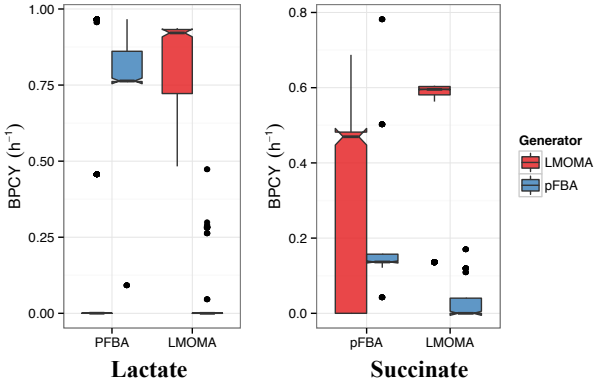

Figure 5: BPCY (*mmol product . mmol substrate*$^{-1}$ *. h*$^{-1}$) boxplots for the Lactate and Succinate case studies. Solutions generated with pFBA (blue,right) and LMOMA (red,left) are re-simulated using both methods (x-axis).

From the boxplots it is clear that the distribution of the BPCY values of the solutions generated when using one of the phenotype prediction methods changes dramatically when using the other. Remarkably, in the succinate case study, the average BPCY of the LMOMA-generated solutions when simulated with pFBA is superior to the average of the pFBA-generated solutions.

This fact is further supported by the scatter plots presented in Figure 6, which allow the visualization of the BPCY obtained using the two different methods for individual solutions. In these plots, particular attention should be paid to the LMOMA solutions in the top right region of the plots. In the perspective of this work, these will be the desired solutions since they provide good results using both prediction methods, being considered more reliable.

Albeit being curious, this can be partially attributed to the larger/less constrained LMOMA solution space. That is, in cases where few valid FBA solutions exist, and there is the necessity for a specific (and restricted) combination of knockouts to guarantee the production of a desired compound, that specific combination might

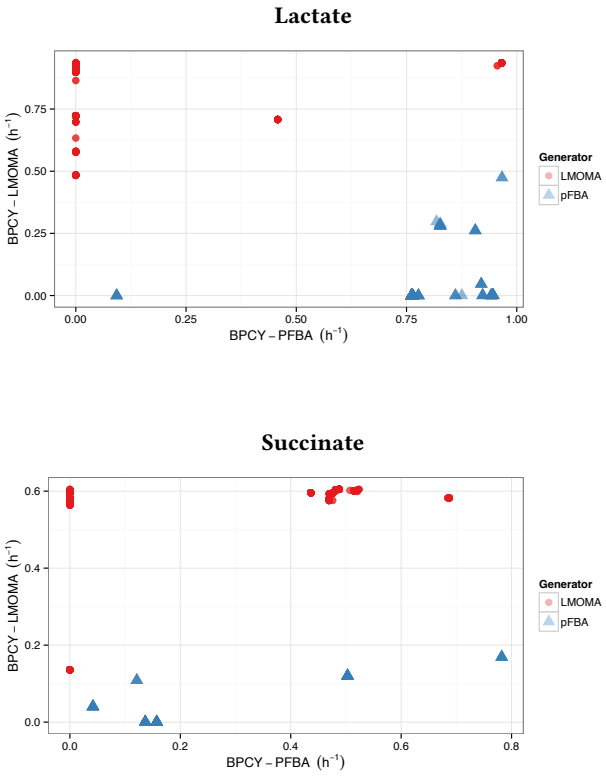

Figure 6: Scatter plots showing BPCY (*mmol product . mmol substrate*$^{-1}$ *. h*$^{-1}$) values obtained by LMOMA(red) and pFBA(blue) generated solutions when simulated using pFBA (x-axis) and LMOMA (y-axis) for the lactate and succinate case studies.

be hard to reach using FBA, thus rendering the optimization process close to a random sampling while no valid solution is found. Alternatively, LMOMA solutions can spread the flux by multiple reactions reaching a multitude of valid solutions from the early stages of the optimization, i.e, with few knockouts (this effect can be observed in the convergence plots). Some of these solutions or areas of the LMOMA solution space are BPCY-valid for FBA as well, as shown by Figures 5 and 6. This supports the rationale that LMOMA-based optimization can be used to guide the FBA-based optimization, which was used as one of the pivotal reasonings behind the development of the tandem optimization approach detailed in the next section.

3.2 Robust strain optimization by means of tandem phenotype prediction

We applied our MO approach to all previously presented case studies and compared the results. Thus, the analysis will be focused on the aerobic production of succinate and lactate using glucose as the carbon source. The number of solutions found by our method in

the two case studies are the following: 1184 in lactate, and 709 in succinate. The generally larger number of BPCY-valid solutions is easily perceptible when compared to the last section results.

Figure 7 represents the Jaccard distance of the mutant flux distributions (solutions found by the tandem approach) to the wild-type flux distribution. From the histograms, it is possible to conclude that, while the average distance of the LMOMA-based flux distributions when simulated with both pFBA and LMOMA has not decreased significantly in comparison with the EA/SA approaches, the outliers found in the pFBA-based flux distributions when simulated with LMOMA are not present anymore.

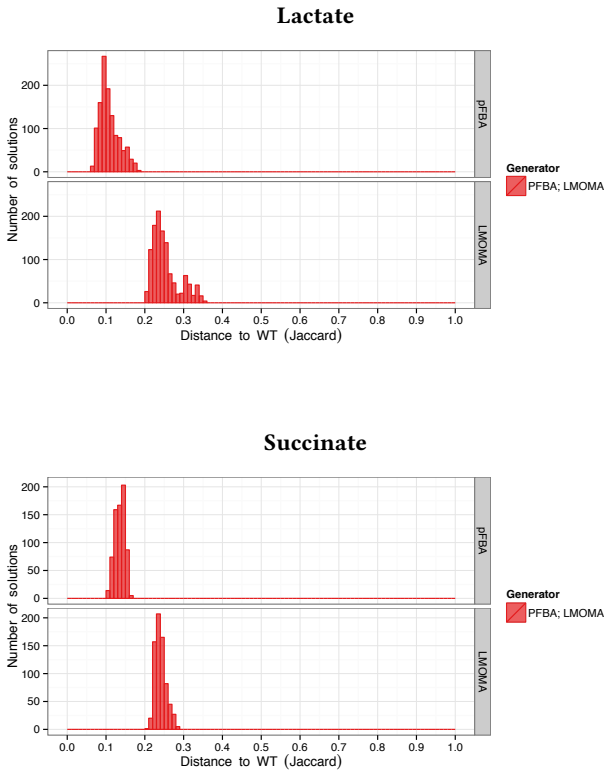

Figure 7: Distribution of the Jaccard distances from the solutions flux distributions to the wild-type flux distributions for the aerobic production of lactate and succinate from glucose. All the solution were re-simulated with pFBA (top in each plot) and LMOMA (bottom in each plot).

This observation suggests that the current solutions are closer to each other in terms of flux distributions. Notwithstanding, as stated in the previous section, if we assume that the likelihood of the solutions working in reality increases, if they are valid for different phenotype prediction methods, no conclusions can be derived about the inter-method validity of these solutions. To access the validity in both pFBA and LMOMA phenotype prediction methods, Figure 8 is introduced. The results presented in the Venn-like diagrams are self-explanatory, with all solutions but 2 in the succinate case study and 1 in the lactate case study being BPCY-valid for both methods. This result is extremely positive by itself, however, the precise performance of these solutions remains to be evaluated.

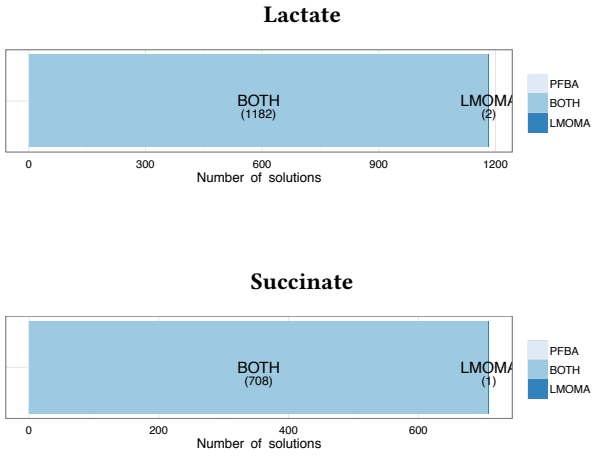

Figure 8: Venn-like diagrams for checking the inter-method validity of the solutions. From left to right, solutions that are BPCY-valid for: only pFBA, both pFBA and LMOMA, and only LMOMA.

The hypothesis raised in the previous section, that the LMOMA-based optimization could be used to guide the pFBA-based optimization, is now revisited here. The corresponding boxplots were generated and are presented in Figure 9.

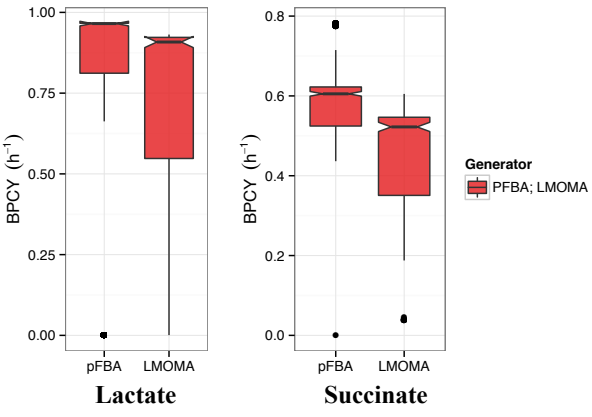

Figure 9: BPCY Boxplots for the Lactate and Succinate case studies obtained by the tandem approach. Solutions are re-simulated with both methods (x-axis).

By analyzing the boxplots it is now evident that the results have improved greatly in comparison with the previous approach. In

the Lactate case study, while in the former approach the pFBA-generated solutions were generally not valid with LMOMA and the LMOMA-generated solutions were not valid with pFBA, here, the solutions are not only valid, but the average of their BPCY values is better, in particular for the pFBA method.

Even more interesting are the results of the Succinate case study. In the previous section, we pointed out the curious results found for this example, where the LMOMA-generated solutions achieved better BPCY values when simulated with pFBA than the pFBA-generated solutions themselves. It is clear that the tandem approach is able to find still better solutions that are valid with pFBA, than in the previous approach, with an average BPCY of 0.6 $mmol\ product\ .\ mmol\ substrate^{-1}\ .\ h^{-1}$. This improvement in the results can also be witnessed in the scatter plots presented in Figure 10 where a large portion of the solutions are located on the top-right region of the plots.

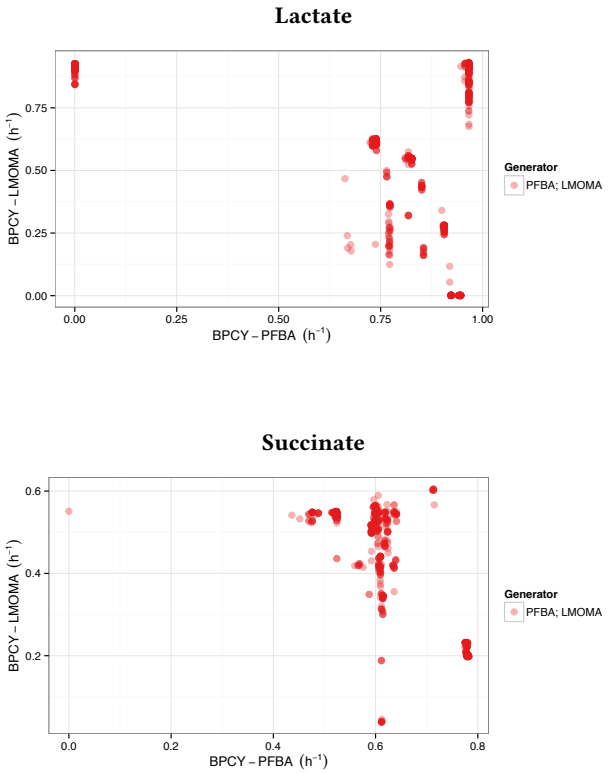

Figure 10: Scatter plots showing BPCY values obtained by the tandem approach generated solutions when simulated using pFBA (x-axis) and LMOMA (y-axis) for the lactate and succinate case studies.

One of our claims is that, in very constrained FBA solution spaces where few BPCY-valid solutions exist, i.e., where to reach solutions than can couple biomass growth and target overproduction a large number of specific deletions is required, the LMOMA-based optimization process can act as chaperone for the FBA-based optimization. To help illustrate this process, Figure 11 is put forward.

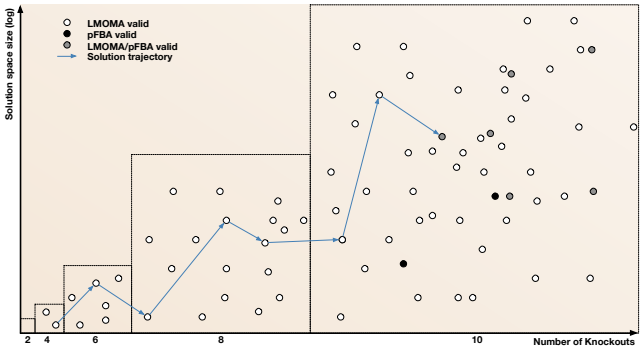

Figure 11: Illustration of LMOMA-pFBA tandem optimization. Solution spaces with few knockouts contain no valid pFBA solutions while some LMOMA solutions can be found. In the much larger, many-knockouts, solution space, pFBA, LMOMA and pFBA/LMOMA solutions are found.

The early stage LMOMA solutions allow the algorithm to initialize convergence towards interesting regions of the solution space, by using solely LMOMA solutions' fitnesses. For higher numbers of deletions, despite being scarce, valid pFBA solutions exist. However, the probability of a valid combination of deletions being found by an evolutionary heuristic using only pFBA as the phenotype prediction method is low. When (and if) the LMOMA and pFBA feasible solution spaces intersect, the tandem optimization approach starts to attribute more value to solutions that are valid for both methods. This is a natural outcome of the dominance property of the underlying MO approach.

In this context, our interpretation of robustness is two-fold. First, we introduce a new concept of robustness in which solutions that are predicted by more than one phenotype prediction method are more robust, since they comply with more than one assumption regarding the behavior of the organism when subjected to perturbations (multi-method robustness). While we will not provide any further tests supporting this claim, this robustness is a natural consequence of the objective functions of our tandem approach, which are sufficiently detailed in our previous analyses.

Secondly, we argue that the tandem optimization process is able to attain solutions that are also robust in the LP (FBA) solution cone (FBA-robustness). The problems associated with competing pathways not being accounted for by strain optimization algorithms were first brought to light by Tepper and Schlomi in [20] where they introduced the concept of robust solutions.

The FBA-robustness is tested in a 2-step approach, first a regular FBA phenotype prediction is performed, maximizing the biomass (bio_{step1}), while subjected to the genetic and environmental conditions of the solutions. Next, FBA is performed with the objective of minimizing the production of the target compound, but an extra constraint - $bio_{step2} \geq bio_{step1} * (1 - \alpha)$ where $\alpha = 0.00001$- is added to the problem. If FBA is still able to predict the production of the target compound in these conditions, we consider the solution to be FBA-robust.

To analyze the validity of this claim, we observed the FBA-robustness of the solutions reached by the tandem algorithm. The results of this analysis show 76% in lactate and over 99% in succinate. Thus, most of the solutions found by the tandem algorithm are FBA-robust. The percentage of FBA-robust solutions found by our method is in the same range of the previous methods, however our method is able to find a much larger number of solutions. Thus, as a consequence, a higher number of FBA-robust solutions is made available to the researchers.

4 CONCLUSIONS

From the results of this work, we have confirmed that the results of strain optimization meta-heuristics are highly dependent on the phenotype prediction methods, and specifically, the use of FBA/pFBA leads to sub-optimal results in more challenging tasks.

A new tandem optimization approach capable of finding robust strain designs compliant with multiple phenotype prediction methods is proposed, to address these limitations. Several advantages emerge from using this tandem approach. First, the algorithm helps uncovering pFBA solutions that would otherwise be difficult to find by traditional approaches. Secondly, the majority of these solutions are both FBA-robust and multi-method robust.

Arguably, a valid alternative would be to ignore FBA/pFBA as a phenotype prediction method for perturbed/mutant organisms and use MOMA/LMOMA. However, LMOMA designs suffer from some limitations. Given that the objective function in MOMA/LMOMA is to minimize the distance to the wild-type flux distribution and since it is not bound to the maximization of biomass constraint, MOMA can artificially activate/deactivate a large number of reactions to reach this minimum value. This results in flux distributions with a large number of minimally activated fluxes, which is unlikely to be biologically sound. Furthermore, because of this, the analysis of the flux distribution of MOMA/LMOMA solutions is a challenging task, whereas analyzing pFBA flux distributions is an amenable one. The solutions attained by our tandem algorithm provide the advantages of both approaches with none of the shortcomings.

ACKNOWLEDGMENTS

The authors acknowledge the project *DD-Decaf - Bioinformatics Services for Data-Driven Design of Cell Factories and Communities*, funded by Horizon 2020 (LEIT Biotechnology, ref. H2020-LEIT-BIO-2015-1 686070-1). This study was also supported by the Portuguese FCT under the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte (NORTE-01-0145-FEDER-000004) funded by ErDF under the scope of Norte2020.

REFERENCES

[1] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard Ø Palsson, and Markus J Herrgard. 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols* 2, 3 (2007), 727–738.

[2] Ana Rita Brochado, Sergej Andrejev, Costas D Maranas, and Kiran R Patil. 2012. Impact of stoichiometry representation on simulation of genotype-phenotype relationships in metabolic networks. *PLoS computational biology* 8, 11 (2012), e1002758.

[3] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering* 84, 6 (2003), 647–657. http://www3.interscience.wiley.com/journal/106556773/abstractpapers2://publication/doi/10.1002/bit.10803

[4] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology* 3, 1 (2007).

[5] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* 28, 9 (Sept. 2010), 977–82. http://dx.doi.org/10.1038/nbt.1672

[6] Rafael U Ibarra, Jeremy S Edwards, and Bernhard O Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420, 6912 (2002), 186–189.

[7] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. 2003. Advances in flux balance analysis. *Current Opinion in Biotechnology* 14, 5 (Oct. 2003), 491–496. DOI: http://dx.doi.org/10.1016/j.copbio.2003.08.001

[8] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. 2003. Advances in flux balance analysis. *Current opinion in biotechnology* 14, 5 (2003), 491–496.

[9] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, and others. 2010. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology* 6, 1 (2010).

[10] Paulo Maia, Isabel Rocha, and Miguel Rocha. 2013. An integrated framework for strain optimization. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*. IEEE, 198–205.

[11] Jens Nielsen. 2001. Metabolic engineering. *Applied Microbiology and Biotechnology* 55, 3 (2001), 263–283.

[12] Kiran Raosaheb Patil, Isabel Rocha, Jochen Förster, and Jens Nielsen. 2005. Evolutionary programming as a platform for in silico metabolic engineering. *BMC bioinformatics* 6, 1 (2005), 308.

[13] M. Rocha, P. Maia, R. Mendes, J. Pinto, E. Ferreira, J. Nielsen, K. Patil, and I. Rocha. 2008. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC bioinformatics* 9, 1 (2008), 499.

[14] Joanne M. Savinell and Bernhard O. Palsson. 1992. Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *Journal of Theoretical Biology* 155, 2 (March 1992), 201–214. DOI: http://dx.doi.org/10.1016/S0022-5193(05)80595-8

[15] Joanne M. Savinell and Bernhard O. Palsson. 1992. Optimal selection of metabolic fluxes for in vivo measurement. II. Application to Escherichia coli and hybridoma cell metabolism. *Journal of Theoretical Biology* 155, 2 (March 1992), 215–242. http://www.sciencedirect.com/science/article/pii/S002251930580596X

[16] Stephan C Schuster. 2008. Next-generation sequencing transforms today's biology. *Nature methods* 5, 1 (2008), 16–18.

[17] Daniel Segre, Dennis Vitkup, and George M Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* 99, 23 (2002), 15112–15117.

[18] Tomer Shlomi, Omer Berkman, and Eytan Ruppin. 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America* 102, 21 (2005), 7695–7700.

[19] George Stephanopoulos, Aristos A. Aristidou, and Jens Nielsen. 1998. *Metabolic Engineering: Principles and Methodologies*. http://www.google.pt/books?hl=pt-PT&lr=&id=9mGzkso4NVQC&pgis=1

[20] Naama Tepper and Tomer Shlomi. 2010. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26, 4 (2010), 536–543.

[21] E. Zitzler, M. Laumanns, L. Thiele, and others. 2001. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. *EUROGEN* (2001), 95–100.