Leveraging Ensemble Information of Evolving Populations in Genetic Algorithms to Identify Incomplete Metabolic Pathways

Eddy J Bautista
Department of Chemical and Biomolecular Engineering
University of Connecticut, Storrs, CT, USA
191 Auditorium Road
Storrs, CT 06226 USA
eddy.bautista@uconn.edu

Ranjan Srivastava
Department of Chemical and Biomolecular Engineering
University of Connecticut, Storrs, CT, USA
191 Auditorium Road
Storrs, CT 06226 USA
srivasta@engr.uconn.edu

ABSTRACT
Genome-scale metabolic models are powerful tools in helping to understand the metabolism of living organisms. They can be applied to the biomedical and biotechnological arenas. The use of such models enables fundamental understanding of metabolism and identification of drug targets in pathogenic microorganisms. They also facilitate metabolic engineering of recombinant organisms to make products useful to society. These mathematical models of metabolism are created based upon the genome annotation of the organism of interest. However, development of high quality versions of these models is non-trivial due to incomplete knowledge regarding gene function, as well as errors in genome annotations. Models developed under such circumstances display "metabolic inconsistency" and are mathematically infeasible. Genetic algorithms may be used to resolve these inconsistencies. In the process, it is possible to take advantage of the ensemble information inherent to the evolving population to gain additional biologically relevant insight. Specifically, it is possible to identify the most pathologic metabolic inconsistencies in an organism, facilitating experimental design and hypothesis development.

Categories and Subject Descriptors
J.3 [**Life and Medical Science**]: Biology and genetics

General Terms
Algorithms.

Keywords
Metabolic modeling, computational biology, metabolic engineering, genetic algorithms, genome-scale modeling, flux balance analysis, systems biology

1. INTRODUCTION
Genome-scale metabolic models represent a powerful tool for modeling metabolism. The constraint-based approach known as flux balance analysis (FBA) has been particularly fruitful [1]. FBA has been successfully used for fundamental research [2], for recombinant protein and metabolite production [3], and for biomedical development, with applications in drug target identification [4] and investigating human metabolic diseases [5]. FBA models are constructed through an iterative process where reactions are assigned to the annotated genes in a genome sequence; the biomass composition and energy requirements are determined; and additional constraints such as mass balance, thermodynamic, environmental and other physico-chemicals are specified. Gaps in the metabolic network are identified and filled using literature and database information resulting in a quantitative stoichiometric system model. Because the model is underdetermined, an objective function is postulated, such as maximization of growth rate, to determine the optimal distribution of metabolic resources or fluxes.

During the reconstruction processes, the most time consuming portion is identifying and resolving the gaps in the metabolic network. Here we discuss an approach in which a genetic algorithm (GA) is integrated with FBA to identify and resolve ill formed mass balance constraints that result in erroneous models.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

2. METHODS
2.1 Genetic Algorithm/Flux Balance Analysis
Flux balance analysis or FBA is a methodology used to model and predict metabolic phenotypes [1]. We have developed an approach that integrates a GA with FBA, referred to as GAFBA, to identify problematic metabolites that are responsible for inaccurate model results. . Inaccuracies generally arise because the mass balance constraints of the problematic metabolites cannot be closed. As a result, there are some additional reactions required for the model to fulfill the mass balance of the relaxed constraints. GAFBA identifies the minimum number of constraints to be relaxed to achieve a feasible solution to the FBA optimization problem. In this work, flux balance analysis (FBA) was applied to the creation of a genome-scale model of *Mycoplasma gallisepticum*. The objective function chosen was maximization of growth rate. Experimental constraints applied were based on measured rates of glucose uptake and lactate production.

To determine the minimum number of relaxed constraints, a genetic algorithm was implemented using an *elite selection* strategy. An initial population of chromosomes was generated. Each chromosome represented a potential metabolic model of the microorganism. The chromosomes were binary encoded where each of the genes represented a mass balance constraint. If the mass balance constraint for a metabolite was relaxed the gene was assigned a value of zero. If it was enforced, the value assigned was one. The objective function for GAFBA was to minimize the number of mass balance constraints relaxed as determined by summing the gene values in each chromosome. However, FBA

model simulations that returned no feasible solution were automatically given a score of -1 regardless of the sum of the gene values. In this way, infeasible results were deprioritized. The following parameters were used: a population of 30 chromosomes was used; the crossover probability was set to 70%; and the mutation probability was set to 1%. The GAFBA algorithm was run for 2,000 generations per simulation. Each simulation took approximately seven hours on a 3.33 GHz Intel Core 2 Duo CPU/ 4 GB workstation. A total of 40 simulations were carried out.

A list of the problematic metabolites was generated upon completion of each run. Manual review of the metabolites, as well as some of the upstream and/or downstream metabolites, in the pathway of the relaxed mass balance constraint was required.

2.2 Frequency analysis of relaxed metabolic constraints

When minimizing the number of relaxed constraints, the GA generated a population of models represented by chromosomes indicating which metabolic constraints had to be relaxed. It was possible to determine the frequency with which a metabolic constraint was relaxed across the population of chromosomes. The metabolic constraints with the highest frequency of relaxation were likely the most pathologic to the generation of a feasible metabolic model. As a further point of reference, as populations evolved, frequencies were tracked across generations. Doing this provided further information regarding how critical a particular metabolic constraint was to the viability of the model.

The process was begun by randomly selecting the final populations of two different simulations, *m* and *n*, from GAFBA for comparison. Each simulation consisted of evolving model populations for 2,000 generations. Constraints that were consistently relaxed in the top 20 chromosomes of each of the final populations were selected. Relaxed metabolic constraints that arose in both of the top 20 chromosomes of each population were considered the most pathologic metabolites. Therefore, they were the first metabolites to be analyzed.

3. RESULTS

The initial model of *M. gallisepticum* consisted of 520 metabolites and 485 reactions. Implementing the model resulted in no feasible solution to the optimization problem due to inconsistent constraints. The model was then analyzed using GAFBA to identify problematic metabolites. Randomly selected simulations were used for carrying out frequency analysis.

At simulation number three, 23 metabolites were shared between the pool of the best 20 chromosomes. The best model only had one more metabolite dropped. Eleven metabolites were present in a previous simulation. Many of these metabolites were amino acids lacking known transport reaction to enter to the cell. Through a literature review and based on experimental conditions, it was possible to infer that some of the problematic metabolites were being transported into the cell from the environment.

Nine shared metabolites were found at simulation number 17. Six of them were already present in a previous simulation. One metabolite was selected for a deep study, and it was discovered that two reactions that were associated with it had been marked as irreversible in the annotation when experimental evidence clearly indicated they were reversible.

After simulation number 27, five metabolites were found common to all the best 20 chromosomes. Four of them had appeared in previous simulations. At this point sodium ion and L-phosphatidate were selected for analysis. It was observed that by solving the mass balance of sodium ion, the mas balance of L-phosphatidate could be closed since the metabolites where connected through the choline pathway. The issue with sodium ion was that the transport of sodium ion was coupled with the transport of choline. Choline was one of the precursors of phosphatidyl choline, which is a component of the biomass of *M. gallisepticum.* Thus, the transported of choline was required, but it was set to zero to avoid accumulation of sodium ion in the cytosol. A Na-ATPase transport reaction for volume regulation in *M. gallisepticum* was added based on literature data[6].

At simulation number 36, NAD+ was the only problematic metabolite in the best chromosome. It was present in all the top 20 chromosomes and it had already appeared in a previous simulation. The model had a partial biosynthesis pathway for NAD+. Based on the observation that *M. gallisepticum* had two reactions of the partial pathway and the requirement of the starting component in other relatives, it was hypothesized that NAD+ synthase reaction was present in *M. gallisepticum*.

Finally, at simulation 40 the final model was obtained. The best model, represented by simulation 40 was able to close all the mass balances constraints and had a growth rate value of $0.3258 \pm 0.12\ h^{-1}$ compared with the experimental value of $0.244 \pm 0.03\ h^{-1}$. The final *M. gallisepticum* model accounted for 441 metabolites and 395 reactions.

4. CONCLUSIONS

The importance of development, curation, and implementation of high quality genome-scale metabolic models will continue to grow. Development of strategies to facilitate these processes will be of extraordinary value. Of particular interest will be approaches that reduce costs and time associated with experimental. The use of GAs for optimization purposes is well established. However, in this case, the ensemble information of the population being evolved turns out to be biologically relevant. It allows the researcher to focus on the most critical metabolites whose resolution will likely give the greatest return on investment of time, effort, and money.

5. ACKNOWLEDGMENTS

Support for this work was provided by NIH Grant 1R03LM009753-01.

6. REFERENCES

[1] Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, New York, 2006.

[2] Yang, K., Ma, W., Liang, H., Ouyang, Q., Tang, C. and Lai, L. Dynamic simulations on the arachidonic acid metabolic network. *PLoS Computational Biology*, 3, 3 (Mar 23 2007), e55.

[3] Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J. and Lee, S. Y. Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, 23, 4 (Aug 2012), 617-623.

[4] Raman, K., Rajagopalan, P. and Chandra, N. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS Computational Biology*, 1, 5 (Oct 2005), e46.

[5] Mardinoglu, A. and Nielsen, J. Systems medicine and metabolic modelling. *Journal of Internal Medicine*, 271, 2 (Feb 2012), 142-154.

[6] Shirvan, M. H., Schuldiner, S. and Rottem, S. Role of Na+ cycle in cell volume regulation of *Mycoplasma gallisepticum*. *Journal of Bacteriology*, 171, 8 (Aug 1989), 4410-4416.