A COMPUTATIONAL DRIVEN DESIGN METHOD FOR ENGINEERING BACTERIAL SYNTHETIC BIOLOGICAL SYSTEMS

Bello Suleiman* and David Gilbert

- bello.suleiman@brunel.ac.uk
- david.gilbert@brunel.ac.uk

Computational Design Laboratory, Synthetic Biology Theme
Institute for Environment, Health and Societies , Brunel University, Uxbridge, UB8 3PH, UK
www.brunel.ac.uk/research/Institutes/Institute-of-Environment-Health-and-Societies/Synthetic-Biology

Brunel University London

Introduction

The Challenge.

The design and optimization of microbial strains with improved metabolite production as targets while preserving optimal biomass is a challenge in synthetic biology.

Solution: Design by Optimization.

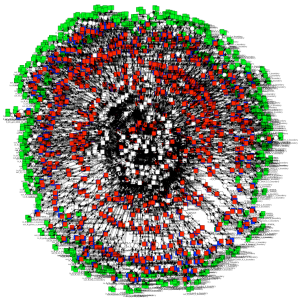
We have developed a **library of biological components** from **naturally evolved strain variants** of Escherichia coli that act as a framework for generating reliable genome-scale models which act as designs at the metabolic level. Using **Flux Balance Analyses**, we take advantage of the stoichiometry of those models and analyze the metabolic capabilities of a microbial strain. Due to their complexities in both regulatory information and topology, a robust computational design approach is needed to predict the effects genetic modifications have on those models. Optimizing these genome scale designs for a desired phenotypic behavior requires **transcriptional information** integrated via gene-reaction-rules. This allows the use of an **evolutionary programming** based approach to optimization and quickly identify gene knock-in/out strategies.

DESIGN: Genome Scale Metabolic Models (GEM)

A genome-scale metabolic network is primarily reconstructed from the information that is present in its genome and the literature. This reconstruction involves steps such as (More detailed):

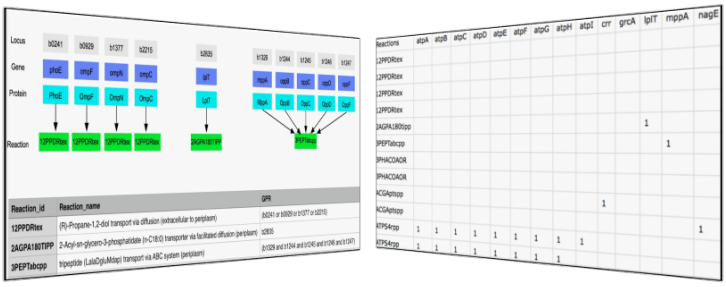
1. Metabolic reactions identification as well as their stoichiometry, direction and compartments i.e. cytosol, periplasm and external.
2. Identification of metabolite charges and chemical formulas.
3. Determination of **biomass composition**.
4. Definition of model constraints.
5. Identifying relationships between genes, the enzymes they transcribe and the reactions those enzymes catalyze.

This information forms the groundwork for the generation of computable stoichiometric model of metabolism. Utilizing constraint-based modelling approaches, GEM can be used as basis of a design to engineer the metabolic systems of an organism.



Genome Level Representation

Gene-Protein-Reaction associations.



Biological reactions are catalyzed by enzymes thus the state of a reaction in a network is controlled by the expression of associated genes. The expression, translation and transcription of genes implies the feasibility of the reactions they encode for. A protein can be made up of subunits. In cases that involve multiple genes and proteins, their relationship can be described using Boolean Logic. The relationship between genes, proteins and their catalyzed reactions can occur in several instances.

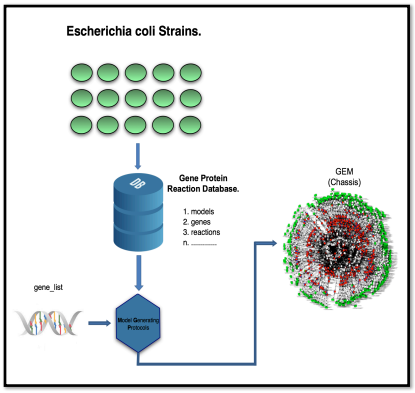
Logic Table

The **Model Generating Protocols** uses the logic table to decide which reactions are chosen during the GEM generation. The logic table is generated from the gene-reaction rules of each metabolic reaction in the Database. Each record is accompanied by "1" which shows the gene present for the reaction to catalyze. Consecutive ones (1) indicate an "AND" relationship. When reactions are represented more than once in the table it indicates the presence of an "OR" relationship. With these two methods a wide array of relationships can be expressed logically in this way. Using the table above, and given a set of genes, the system can determine if a reaction has met its logical gene criteria and either accept or deny candidate reactions from being included in a model.

Gene reaction rule example:

ATPS4rpp: (atpA and atpB and atpC and atpD and atpE and atpF and atpG and atpH and atpI) or (atpA and atpB and atpC and atpD and atpE and atpF and atpG and atpH)

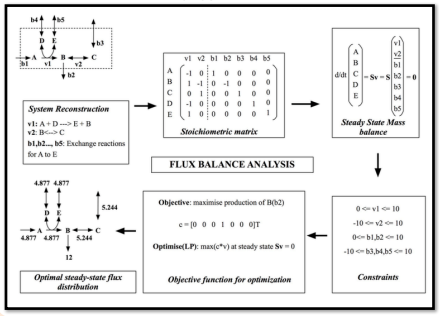
Strain Generation



We currently use **Escherichia coli** because it is the most detailed and complete metabolic reconstruction of any organism to date. By integrating a library of **naturally evolved strain variants** from several published genome-scale metabolic models into one resource, we can compare components i.e. reactions, metabolites and genes, across different strains. The database resource holds the the union of the gene, protein and reaction sets of all the strains of a specie. The library would grow over time as more knowledge is learnt about the organism. Given an Ecoli gene list, we can generate the corresponding genome scale model using metabolic components from every strain in the database resource. This GEM forms the **chassis strain** that will undergo target driven optimization.

Flux Balance Analyses

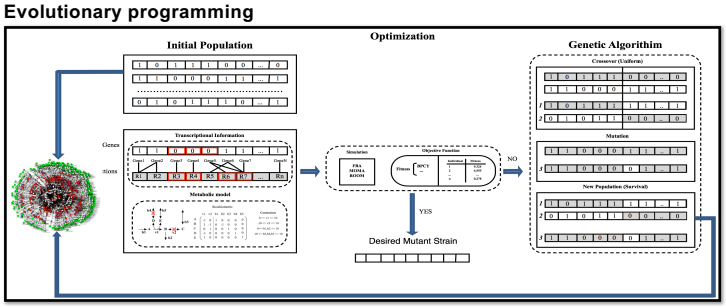
FBA



FBA is a technique used in biotechnology and systems biology to analyze the complete metabolic genotype of a microbial strain. It converts each metabolite of a metabolic network into mathematical coefficients using linear programming. It then examines the relationships between metabolites in a network to find solutions that satisfy some optimal behavior of the metabolic network at steady state. The constraints are often the maximization of biomass yield of an organism and bounds on flux values of reactions. This allows an attainable set of generalized predictions for the growth of the organism, metabolite concentration levels and product output inside the cell.

Target driven optimization

Evolutionary programming



The generated GEM is called the **chassis strain**. The default state of the chassis is saved. All reactions and their lower and upper boundaries are stored and used to return the chassis back to its original state. Below are the steps the protocol takes to implement the genetic algorithm as a method to solve the problem at hand.

1. Setup steps:
   1. The initial chassis gene list is encoded as an array. The array is propagated into the population as individuals. Individuals have randomly generated binary off and on states. The list of randomly generated individuals make up the **initial state of the population**. Each individual represents a chromosome that could be the desired mutant strain solution of the problem at hand.
   2. The **transcriptional information** is used to determine which genes are active or not, which in turn controls which reactions are active as well. Reactions that aren't in the list are deactivated by setting their lower and upper bounds to zero. This prevents those reactions from carrying any flux during Flux balance analyses. The fitness is calculated using the Biomass Product Coupled Yield (BPCY) equation (Patil et al., 2005).
2. Variants in the population are created by crossover and mutation reproduction operations. These operations are responsible for genetic sharing and variation.
3. Once each member of the population undergoes the reproduction operations, they are analyzed utilizing **Flux Balance Analyses**. Their fitness values are stored and a **fitness evaluation** is conducted to determine if the desired mutant strain has been found. After the evaluation the population is reduced to its initial population size by selecting the individuals with the best fitness value.
4. Steps 2 and 3 are repeated until the number of generations allocated are exhausted.

The final output would be an array of binary integers indicating what genes to turn off in a set of genes in order to manipulate the cellular topology and achieve particular phenotypic behavior. It is important to note that the optimization can be conducted in different environmental condition and on different carbon sources. The organism can be designed to maximize the production of any of its **naturally produced compounds.**

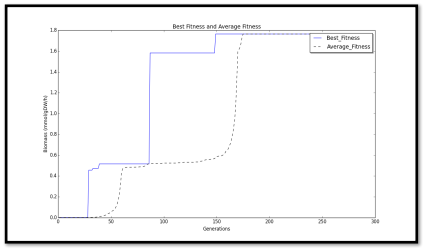
Fitness Score

Coupling the target production yield with the biomass production ensures that the cell must produce the target molecule and the biomass components required for growth to achieve its Fitness Value. The Fitness score is calculated using a combination of an objective value and reaction fluxes that make up the Biomass Product Coupled Yield (BPCY) (Patil et al., 2005).

$$BPCY = \frac{PG}{S}$$

Where **P** is the flux of the product produced, **G** is the flux of the biomass function (growth rate), and **S** is the flux of the desired substrate consumed.

Fitness Evaluation



The best fitness of each generation holds its unique set of chromosome values. The differences between them is as important as the overall best fitness of the run. Genetic contrast help indicate unique and important genes required for different phenotypic behaviors. Analyzing those distinctions could be the key to further our understanding of the metabolic capabilities of a microbial organism. This design methodology can be applied to **other bacteria** as well e.g. shigella, yeast.

Conclusions

The need for prediction with reasonable certainty before organic implementation becomes apparent when we weigh the risks of uncertainty. A methodology that uses the wealth of biological data available to create predictive models and optimizes them with the objective of obtaining desirable phenotypes could be implemented in tools and would make a significant contribution to the efficiency and reliability of the Synthetic Biology ecosystem. **Future work (alleles and foreign gene transfer, other bacterial species.)**

References

1. Feist, Adam MPalsson, 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26, 659–667. doi:10.1038/nbt1401
2. Patil, K., Rocha, I., J., Nielsen, J., 2005. Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics 6, 308. doi:10.1186/1471-2105-6-308