SAND2019-7077C

BioRetroSynthesis: A tool for identifying optimal metabolic pathways for production of a target compound

Leanne S. Whitmore[1,2], Anthe George[1,2] & Corey M. Hudson[1,2]

[1]Sandia National Labs, Livermore, CA, [2]Joint BioEnergy Institute, Emeryville, CA

Motivation

The biological production of high value chemical compounds and fuels from low value biological feedstocks - holds great promise for transforming the availability of more environmentally suitable chemicals. Unfortunately, biological production of compounds is still extremely inefficient. One of the most complex, time consuming, and expensive steps in bioengineering an organism is the selection of enzymes and reactions that will optimize the production of a target compound. Manual search of all microbial organisms for enzymes and reactions that are capable of synthesizing a compound is intractable and in many cases, infeasible and ultimately results in missing potential enzyme/reaction pairs.

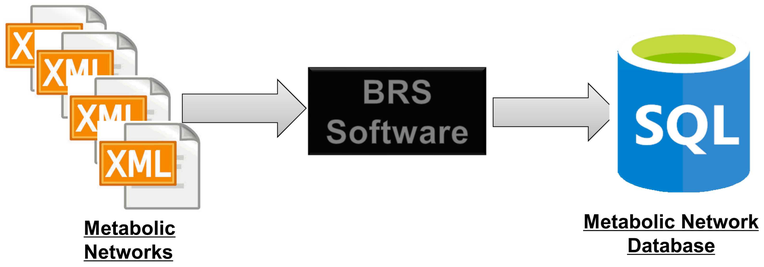
BioRetroSynthesis (BRS) reduces the complexity of this step in bioengineering in three ways:

1. Builds a comprehensive database of bacterial metabolic networks (including genes, compounds and reactions).
2. Identifies the minimal number of reactions and enzymes that need to be added to an organism to synthesize a select chemical compound.
3. Simulates metabolism using flux balance analysis (FBA) to identify theoretical yield of a compound and in the event that multiple paths were identified determines the optimal routes of the added reactions in production of a compound.

Overall, BRS streamlines an arduous and complex step of bioengineering a microbial organism, which will enable scientists to inexpensively expedite the production of important chemical compounds.

Metabolic database construction

BRS requires a repository of metabolic information for microbial organisms. BRS builds an SQL database of metabolic networks from their source Systems Biology Markup Language (SBML) files. We used 13,145 curated bacterial metabolic networks retrieved from Kbase (Arkin 2016) and reactions from the MetaCyc database (Caspi 2010) to build a metabolic database database.



Identifying minimal number of reactions using integer linear programs

BRS uses a novel integer-linear-program (ILP) algorithm to identify the minimum number of reactions (shortest path) to be added to an organism to synthesize a compound of interest.

Description of algorithm:

- All reactions in a metabolic database are variables (x_i).
- Variables are set as integers: bound to 0 and 1.
- Every compound in the database is represented by a mass balance equation. Equations are generated by the summation of reaction variables:
  - Variables have a -1 coefficients for reactions that consume the compound and 1 coefficients for reactions that produce the compound.
  - Each mass balance equation constraint is set equal to a different value dependent on whether the compound is **internal**, **external** or the **target** compound

$$\text{minimize } c^t x_i$$

$$\sum_{i=1}^{i} (X_i) = \infty \text{ For all internal compounds}$$

$$\sum_{i=1}^{i} (X_i) = 0, \infty \text{ For all external compounds}$$

$$\sum_{i=1}^{i} (X_i) = 1 \text{ For target compound}$$

$$x_i \in \{0,1\}$$

- **Internal compound** = *native* to the organism that the target compound is to be produced in.
- **External compound** = *non-native* to the organism that the target compound is to be produced in.
- **Target compound** = compound that is to be produced in a select organism
- **Objective Function (Z):** reactions needed to produce the compound. Weights (c^T) of 0 given to internal reactions and weights of 1 given to external reactions.
- **Multiple pathways**: To identify multiple pathways weights in **Z** for variables (reactions) found initially are increased. New weights (W) are calculated by dividing 1 by the number of reactions in the initial shortest path and then subtracting .01. W is then added new to initial weight. The problem is re-solved and if a new pathway is found of equal length to the shortest path, W is added to the initial weights of the new reactions identified in the new path and the problem is again solved until no further reactions are identified.

Other features of BRS

Flux balance analysis: Once pathway reactions have been identified, BRS simulates metabolism in the organism and predicts the efficiency of target compound production using Flux Balance Analysis. To implement FBA, BRS uses the already developed CobraPy (Ebrahim 2013) toolbox to easily and quickly predict maximum threshold yields of a compound.

Visualization: BRS visualization is done using GraphViz to generate graphs that depict the reactions and corresponding genes that need to be added to an organism. Further more if FBA has been implemented graphs can show the activity of reactions that were inserted into the select organism.

References:

Arkin, Adam P., et al. "The DOE Systems Biology Knowledgebase (KBase)." *bioRxiv* (2016): 096354.

Caspi, Ron, et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic acids research* 38.suppl 1 (2010): D473-D479.

Ebrahim, Ali, et al. "COBRApy: constraints-based reconstruction and analysis for python." *BMC systems biology* 7.1 (2013): 74.

Zheng , et al. "Metabolic engineering of Escherichia coli for high-specificity production of isoprenol and prenol as next generation of biofuels." *Biotechnology for biofuels* 6.1 (2013).

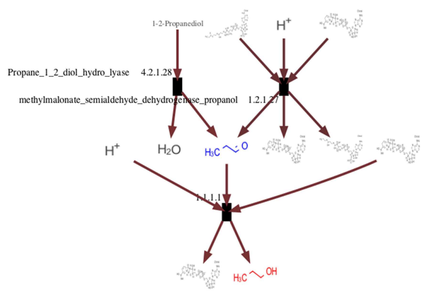
Tranh, Cong T., et al. "Redesigning Escherichia coli metabolism for anaerobic production of isobutanol." Applied and environmental microbiology 77.14 (2011): 4894-4904, which also discusses the theoretical yields

Atsumi, Shota, et al. "Metabolic engineering of Escherichia coli for 1-butanol production." *Metabolic engineering* 10.6 (2008): 305-311.

Validation of BRS

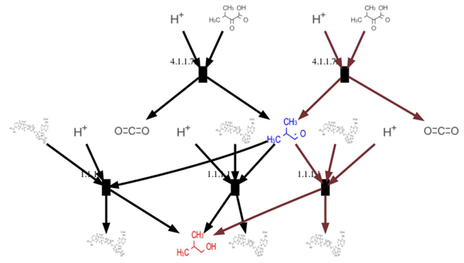
To validate the objective function (the minimal number of reactions) we searched for pathways of target compounds in organisms (E. Coli DH1) which have already been genetically modified and reported on to produce the target compound.

1-propanol



This route uses either 1,2-Propanediol or Propionyl CoA as precursor and runs through two separate enzymes: E. coli can be engineered to overproduce 1,2-Propanediol (Jain, Rachit, et al. "Engineering microaerobic metabolism of E. coli for 1, 2-propanediol production." *Journal of industrial microbiology & biotechnology*42.7 (2015): 1049-1055.)

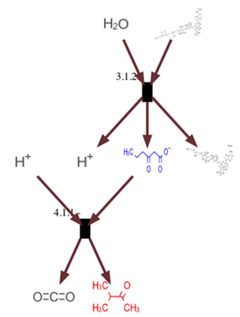
Isobutanol



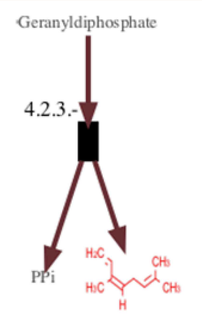
Production of isobutanl from 3-methyl-2-oxobutanoate (also known as a-ketoisovalerate) is discussed in Trinh (2011). This happens through multiple potential pathways. These are through either *alsS* + *adhE* genes or *kivd* + *adh2* genes. A variety of other genes can be added to maximize the production of a-ketoisovalerate precursor.

Compounds with novel production pathways

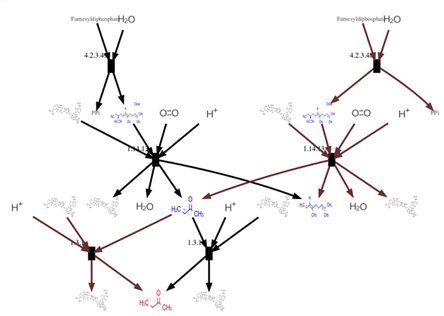
3-methyl-2-butanone



beta ocimene



2-butanone



Compounds	E.C. Number	Name
3-methyl-2-butanone	3.1.2.-	3-oxoacyl-CoA hydrolase
	4.1.1.-	methylketone synthase
beta ocimene	4.2.3.-	δ-selinene synthase
butan-2-one	4.2.3.49	terpene synthase
	1.14.13.-	DMNT synthase
	1.3.1.-	alkenal/one oxidoreductase

Theoretical Yields

BRS can proceed to simulate metabolism of the organism with the identified pathways on a glucose media predicting theoretical yield of a compound and, in the invent that multiple pathways were identified, which pathway will perform optimally.

Target Compound	Organism	Glucose Uptake	Target Production	Theoretical Yield
3-methyl-2-butanone	Escherichia coli	Glucose Flux: -100.0	Target Flux: 88.1	0.88 mol 3-methyl-2-butanone/mol glucose
beta ocimene	Escherichia coli	Glucose Flux: -100.0	Target Flux: 33.04	0.33 mol beta ocimene/mol glucose
butan-2-one	Escherichia coli	Glucose Flux: -100.0	Target Flux: 21.74	0.22 mol butan-2-one/mol glucose

Future Work

Future plans for the development of this software include building a more comprehensive database which includes not only biological reactions but chemical reactions as well thereby allowing for the user to obtain information about both biological and chemical reactions needed to produce a compound of interest.

Funding & Acknowledgements

This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies and Vehicle Technologies Offices.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. **SAND Number:**