Optimizing genetic manipulation of microbial organisms for production of multiple target chemical compounds

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

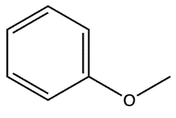
SAND2019-7076C

Abstract

Biological compound production for industrial and economical purposes is an interesting and complex problem. Ideally, minimal genetic manipulation to the industrial microbial organism increases the likelihood of successful target compound production as well as decreases cost. Determining which gene additions would allow production of multiple target compounds is a difficult and a nearly impossible task without the help of computational tools. RetSynth is a tool we developed that uses a novel constraint-based approach, which works through dynamic and recursive manipulation of constraints followed by integer-linear programming, to discover the minimal number of reactions/genes (optimal solution) that when added to a chassis organism results in production of a target compound. Unique to RetSynth, all optimal solutions, all combinations of the minimal number of reactions/genes that can produce the target chemical, are discovered allowing the user more reaction/gene options. Additionally, sub-optimal solutions to target compound production, solutions which require more reactions than the minimal number, can also be identified. By obtaining all optimal and sub-optimal solutions for multiple target chemicals we can identify genes that overlap in the solutions and therefore would be ideal gene additions to the host organism. Adding the gene that appears in the greatest number of solutions into the chassis organism would lessen the amount of downstream genetic manipulation required for production of many target compounds. We identified optimal and sub optimal solutions in *Escherichia Coli K12*, *Psuedomonas Putida KT-2440*, and *Streptomyces Venezuele ATCC* for 1979 hydrocarbons from the MetaCyc database. From these solutions we discovered genes that are present the highest number of target production solutions and would therefore make an optimal genetic manipulation to each of the chassis organisms.

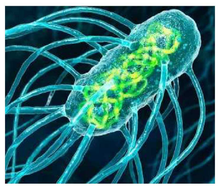
RetSynth workflow

Target compound

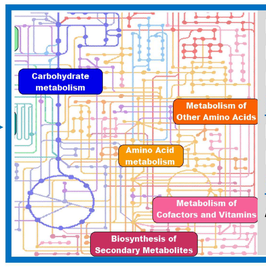


anisole

Chassis organism



Metabolic Database



Kbase https://kbase.us/
MetaCyc https://metacyc.org/
Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/
Metabolic In Silico Network Expansion Databases http://minedatabase.mcs.anl.gov/#/home
ATLAS of Biochemistry http://lcsb-databases.epfl.ch/atlas/

Solve for external pathway (optimal solution) using novel integer program

- C stoichiometric matrix mxr, with m compounds and r reactions, describing availability of metabolites and reactions in the database
- D is a vector combining vectors N, V, A which represent availability of native metabolites in the chassis organism, non-native metabolites and target compound, respectively
- I is a vector of constraints for native reactions in the chassis organism
- E is a vector of constraints for available reactions of size m
- *Solve integer program minimizing external reactions added to system to produce target compound*
- **Multiple Solutions:** To calculate all optimal solutions a penalty function is added to variables (reactions) that are are already part of discovered solution
  - Penalty function $= 1 + \frac{1}{2\beta^*}$. β^* = number of reaction steps in optimal solutions
  - Add weights to each variable in optimal solution and continue to resolve problem until $\mathbf{t}^T\mathbf{x} = \beta^*\left(1 + \frac{1}{2\beta^*}\right)$
- **Sub-optimal Solutions:** Solutions that are greater in reaction number than the optimal solutions but can still produce target compound.
  - To calculate sub optimal solutions constraints are added to the integer program to prevent optimal solutions from being identified forcing the program to seek alternative solutions.
  - O = reactions in β^*, Y = reactions not in β^*, MS = all optimal solutions

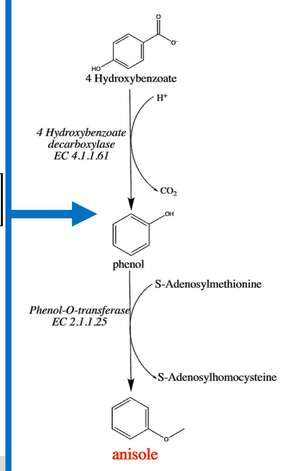
$$\min \mathbf{t}^{\mathrm{T}}\mathbf{x}$$
$$s.t\ Cx \le d$$
$$for\ each\ \beta^*\ in\ MS:$$
$$P^{\mathrm{T}}x \le \beta^* - 1$$
$$\mathbf{x} \in \{0,1\}^r$$

$$N = \begin{bmatrix}\infty\\ \infty\\ \vdots\\ \infty\end{bmatrix} V = \begin{bmatrix}0\\ 0\\ \vdots\\ 0\end{bmatrix} A = [1] \quad D = \begin{bmatrix}n\\ v\\ a\end{bmatrix}$$

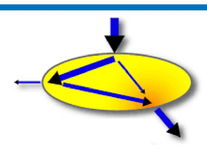
$$I = \begin{bmatrix}0\\ 0\\ \vdots\\ 0\end{bmatrix} E = \begin{bmatrix}1\\ 1\\ \vdots\\ 1\end{bmatrix} \quad \mathbf{t} = \begin{bmatrix}E\\ I\end{bmatrix}$$

$$Y = \begin{bmatrix}0\\ 0\\ \vdots\\ 0\end{bmatrix} O = \begin{bmatrix}1\\ 1\\ \vdots\\ 1\end{bmatrix} P = \begin{bmatrix}S\\ O\end{bmatrix}$$

Optimal solution



Flux Balance Analysis



Once optimal solutions have been identified flux balance analysis (FBA) is used to simulate metabolic activity of the organism to determine theoretical yields of the target compound with the added pathways. Additionally, gene and reaction knockouts are performed to see if theoretical yields can be increased by further genetic manipulation.

Download RetSynth at https://github.com/sandialabs/RetSynth

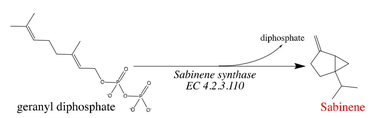
This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies and Vehicle Technologies Offices. Sandia National Laboratories is a multimission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

Identifying optimal genetic manipulations

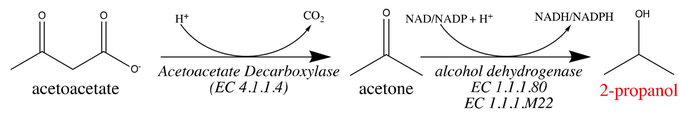
Validation of RetSynth

To validate RetSyth we searched for pathways to target compound production for which there was already experimental pathways developed in *Escherichia Coli DH1*. This allowed us to compare our tools results to experimentally developed pathways to prove our algorithm was effective at identifying reasonable pathways to production.

Sabinene



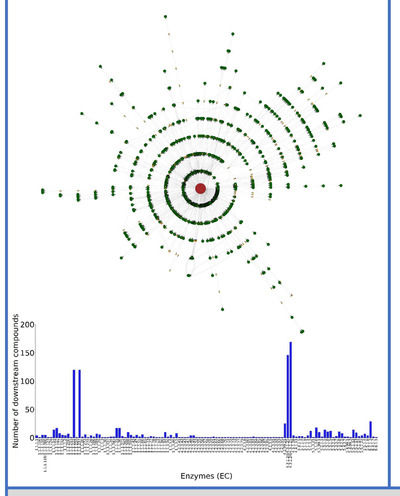
2-propanol



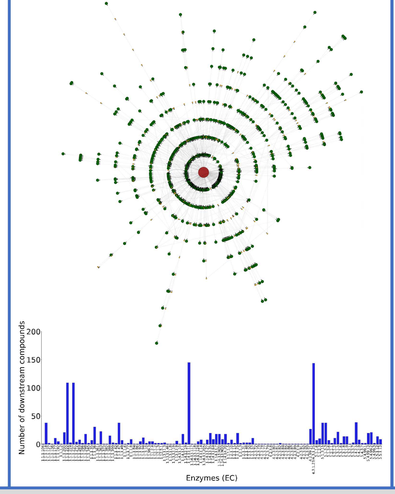
Optimal genetic manipulations

- To find the optimal initial genetic manipulation (the gene/enzyme if added that would lead to the highest number of down stream target production if further genetic manipulation was added) we used RetSynth to identify optimal and sub-optimal pathways for 1979 hydrocarbon molecules in host organisms *Escherichia Coli K12*, *Pseudomonas putida KT2440*, and *Streptomyces venezuelae ATCC 10712* . It was then determined which gene/enzyme was apart of the most number of targets' pathways to production and therefore would be a beneficial genetic manipulation.
- Graphs below depict the optimal and sub optimal solutions that were identified for the 1979 for each of the chassis organisms.
  - Brown (center node) – represents the chassis organism, Green nodes – represent the target compounds, Orange nodes – represent intermediate compounds, Black edges – represent reactions

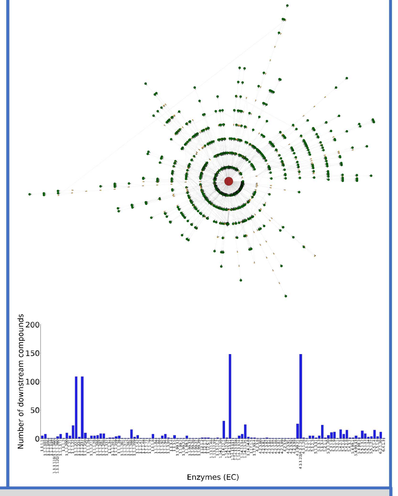
Escherichia Coli K-12



Psuedomonas Putida KT-2440



Steptomyces Venezuele ATCC



Conclusions

Of the three chassis organisms *Pseudomonas Putida* can with the optimal pathways can produce the most target compounds (904) closely followed by *Escherichia Coli* (866) and *Streptomyces Venezuele* (866).

- Hydroxyphenylpyruvate reducatases (EC 1.1.1.222 and 1.1.1.237), which catalyze the conversion of P-hydroxyphenylpyruvate and 3,4-dihydroxyphenylpyruvate to (R)-3-(4-hydroxyphenyl)lactate and (R)-3- (3,4-dihydroxyphenyl)lactate respectively, when added to all each of the 3 chassis organisms can lead to the downstream production of ~100 hydrocarbons. These enzymes are natively found in the organism *Solenostemon scutellariodes.*
- Tyrosine ammonia-lyase (EC 4.3.1.23/EC 4.3.1.25) which catalyzes the reaction of tyrosine to 4-coumarate can also lead to the production of ~150 down stream compound products in each of the chassis organisms. This enzyme can be found naturally in *Rhodotorula glutinis*.
- Cinnamate 4-hydroxylase (EC 1.14.13.11), which catalyzes cinnamate to 4-courmate when added can lead to the production of ~150 downstream targets in chassis organisms *Pseudomonas* and *Streptomyces* and is naturally found in *Arabidopsis thaliana.*
- Phenylalanine ammonia-lyase (EC 4.3.1.24/EC 4.3.1.25) which catalyzes the reaction of phenylalanine to cinnamate can also lead to the production of ~400 down stream compound products in each of the chassis organisms. This enzyme can be found naturally in a number of species including *Streptomyces maritimus, Arabidopsis thaliana* and *Ustilago maydis 521*.

Overall by using RetSynth we were able to identify the described enzymes that if added to the respective chassis organisms would reduce the genetic manipulation that would be required by researchers to produce a multitude of different hydrocarbon targets.