Sandia National Laboratories

Exceptional service in the national interest

SAND2016-6910C

A Novel Retrosynthesis Software for Identifying Optimal Pathways for Biological Production of Fuel Compounds

Sarah Deng (Mission San Jose High School), Leanne Whitmore, and Corey Hudson
Sandia National Laboratories, 8633 Systems Biology Department
July 27,2016

Abstract

The Co-Optima project aims at co-optimizing fuels and advanced engines. The Low-Greenhouse Gas Teams objective within this project is to identify and biologically produce low-greenhouse gas compounds that can be added as blendstocks to existing petroleum-based fuels. Our current work focuses on determining the most efficient paths to biological compounds production. We are creating software that uses genome-scale metabolic information for a diverse group of bacteria from the Department of Energy Bioinformatic tool, KBase (https://kbase.us/), to identify efficient synthetic biological paths to target molecule production.

Introduction

As a nonrenewable resource, oil is something that needs to be used conservatively. As of 2016, there is an estimated 1624.6 billion barrels of oil left, which can power the world for approximately another 51.3 years. Fossil fuel use is the primary contributor to fluorinated gas emissions, making up about 65% of global emissions. Decreasing emissions by even small factors can make drastic differences in the condition of the environment. By increasing efficiency, it is possible to make the most out of the oil and also benefitting the environment by possibly reducing emissions. Using the BiocompoundML software, we have predicted high research octane numbers (RON) for 114 compounds (Table 1). High RON values indicate increased tolerance to combustion, resulting in more efficient fuel (Figure 1). Our focus now is how to optimize production of these compounds biologically. To address this issue, we are developing a retro-synthesis tool which will, given a bacterial organism you want to use for production of target compounds, predict the best synthesis pathways for the target compound, outputting necessary information such as genes that may need to be introduced into organism for production.

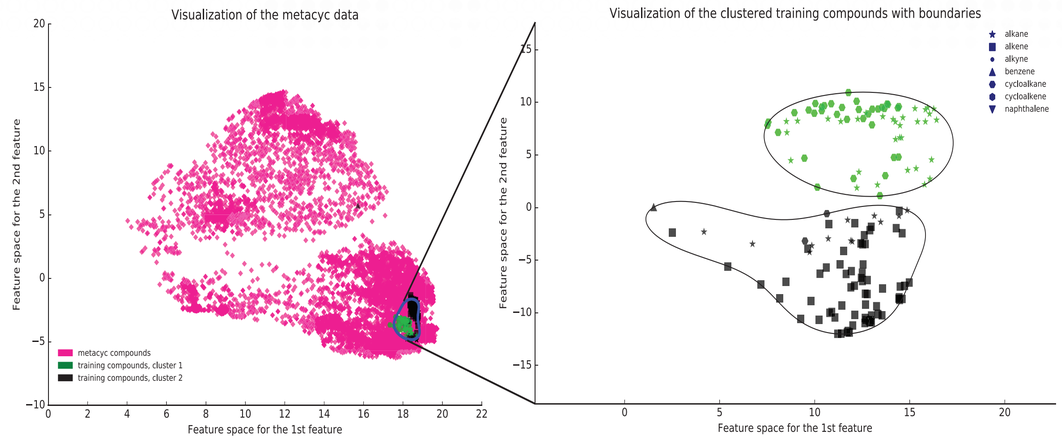

Figure 1: MetaCyc compounds and clustering of RON training data(Whitmore L. et. al unpublished)

Compound	Probability Low RON	Probability High RON	CAS	PubChem	Formula
butyl acetate	0.010	0.990	123-86-4	31272	C6H12O2
1,4-benzoquinone	0.023	0.977	106-51-4	4650	C6H4O2
fumarate	0.023	0.977	110-17-8	5460307	C4H2O4
ethanol	0.025	0.975	64-17-5	702	C2H6O
diacetyl	0.027	0.973	431-03-8	650	C4H6O2
1-O-methylsalicylate	0.029	0.971	119-36-8	4133	C8H8O3
2-methylbutanol	0.029	0.971	137-32-6	8723	C5H12O
anisole	0.029	0.971	100-66-3	7519	C7H8O
ethyl acetate	0.031	0.969	141-78-6	8857	C4H8O2
2-methylbut-3-en-2-ol	0.033	0.967	115-18-4	8257	C5H10O
methylglyoxal	0.033	0.967	78-98-8	880	C3H4O2
patulin	0.035	0.965	149-29-1	4696	C7H6O4
3-methylbutanol	0.035	0.965	123-51-3	31260	C5H12O
cyclopentanone	0.035	0.965	120-92-3	8452	C5H8O
acetoin	0.037	0.963	513-86-0	179	C4H8O2
1,3-propanediol	0.037	0.963	504-63-2	10442	C3H8O2
(-)-camphor	0.037	0.963	464-48-2	444294	C10H16O
(R)-mevalonate	0.039	0.961	150-97-0	5288798	C6H11O4
2-butyne-1,4-diol	0.039	0.961	110-65-6	8066	C4H6O2
2-methylphenol	0.043	0.957	95-48-7	335	C7H8O

Table 1: List of MetaCyc compounds predicted to have high RON value (Whitmore L. et. al unpublished)

Methods

Dijkstra

- Shortest path algorithm
- Starts with tree from initial vertex to every other vertex
- Choses shortest path
- Repeat until all vertexes reached
- No repeated paths
- Result = shortest overall path

Bellman-Ford

- Shortest path algorithm
- Start by considering all paths
- Initial value for all vertexes = infinity
- Relax edges
- Pick the shortest paths of relaxed edges
- Result = shortest overall path

Flux Balance Analysis (FBA)

- Constraint-based modeling approach simulating metabolism for genome wide reconstructions of metabolic networks
- Formulates a systems of equations describing the production and consumption of each metabolite in network as a dot product of a matrix of stoichiometric coefficients (S) and vector of unknown reaction fluxes (v), set equal to 0, to simulate the assumption of steady-state
  - Sv=0
- Linear programming is used to solve for reaction fluxes, v

Results

1. Software will begin by users inputting organism of interest, source compounds (media), and target compound
2. Utilizing the database KBase, specifically metabolic networks the software will:
   1. First identify if compound can be produced in organism
   2. If not, identify the lowest number of genes (shortest path) that needs to be added to organism for production
3. FBA will be used to simulate metabolism identifying shortest pathways needed to best maximize production of target compound and growth of organism

Discussion

The functions of the software can also be applied to a variety of fields, such as synthesis of materials previously detrimental to the environment like plastic compounds. Overall, by being able to efficiently biologically synthesize molecule, we can greatly decrease emissions and creation of unnecessary wastes.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration