Final Report

Department of Energy Grant DE-SC0004878

"Curation and Computational Design of Bioenergy-Related Metabolic Pathways"

Peter D. Karp, Mario Latendresse, Ron Caspi

SRI International

Project Summary

Pathway Tools is a systems-biology software package written by SRI International (SRI) that produces Pathway/Genome Databases (PGDBs) for organisms with a sequenced genome. Pathway Tools also provides a wide range of capabilities for analyzing predicted metabolic networks and user-generated omics data. More than 5,000 academic, industrial, and government groups have licensed Pathway Tools. This user community includes researchers at all three DOE bioenergy centers, as well as academic and industrial metabolic engineering (ME) groups. An integral part of the Pathway Tools software is MetaCyc, a large, multiorganism database of metabolic pathways and enzymes that SRI and its academic collaborators manually curate.

This project included two main goals:

I. Enhance the MetaCyc content of bioenergy-related enzymes and pathways.

II. Develop computational tools for engineering metabolic pathways that satisfy specified design goals, in particular for bioenergy-related pathways.

In part I, SRI proposed to significantly expand the coverage of bioenergy-related metabolic information in MetaCyc, followed by the generation of organism-specific PGDBs for all energy-relevant organisms sequenced at the DOE Joint Genome Institute (JGI). Part I objectives included:

1: Expand the content of MetaCyc to include bioenergy-related enzymes and pathways.

2: Enhance the Pathway Tools software to enable display of complex polymer degradation processes.

3: Create new PGDBs for the energy-related organisms sequenced by JGI, update existing PGDBs with new MetaCyc content, and make these data available to JBEI via the BioCyc website.

In part II, SRI proposed to develop an efficient computational tool for the engineering of metabolic pathways. Part II objectives included:

4: Develop computational tools for generating metabolic pathways that satisfy specified design goals, enabling users to specify parameters such as starting and ending compounds, and preferred or disallowed intermediate compounds. The pathways were to be generated using metabolic reactions from a reference database (DB).

5: Develop computational tools for ranking the pathways generated in objective (4) according to their optimality. The ranking criteria include stoichiometric yield, the number and cost of additional inputs and the cofactor compounds required by the pathway, pathway length, and pathway energetics.

6: Develop tools for visualizing generated pathways to facilitate the evaluation of a large space of generated pathways.

Results

Part I.

Objective 1. To expand the content of MetaCyc to include bioenergy-related enzymes and pathways, our efforts added 49 new pathways to MetaCyc, including the following.

Lignocellulose degradation: we added pathways for the degradation of most important lignocellulosic compounds, including cellulose, rhamnogalacturonan, mannan, xylan, arabinan, xyloglucan, glucuronoarabinoxylan, pectin, xylose, 1,5-anhydrofructose, furfural, 5-hydroxymethylfurfural, vanillin, and vanilate. We also curated pathways for the degradation of marine biopolymers such as agarose, carrageenan, chitin, alginate, and porphyran.

Aliphatic carbon production: we added important pathways for the biological production of compounds such as alkanes, very long chain fatty acids, terminal olefins, lipids, and alcohols.

Hydrogen production: we added pathways describing naturally occurring and bioengineered hydrogen production, and curated the relevant enzymes from different species of archaea, bacteria, and algae. MetaCyc now contains pathways describing hydrogen production by 15 key species.

Algal oil production: we added pathways describing the production of the lipids botryococcene and methylated squalene by the green alga *Botryococcus braunii*, the long chain fatty acid docosahexanoate by the microalga *Isochrysis galbana*, and the polar acyl lipid diacylglyceryl-*N,N,N*-trimethylhomoserine by the unicellular alga *Ochromonas danica* and the green alga *Chlamydomonas reinhardtii*.

In addition, we also added a number of engineered (artificial) pathways for biofuels production to MetaCyc because such pathways are required to increase production rates of potential biofuels, and are of great interest to metabolic engineers. Among these are pathways for the production of long chain fatty acid esters, methyl ketone, hexanol, isopropanol, butanol, methylbutanol, isobutanol, ethylene, and isoprene.

Objective 2. To enable Pathway Tools to display and edit complex polymer degradation processes, we developed an interface between Pathway Tools and the GlycanBuilder software. In addition, we introduced support for the symbolic representation of the glycans recommended by the Consortium for Functional Glycomics (CFG) that uses the Glyco-CT format. These enhancements enabled developing a new type of pathway diagram that displays the structure of complex carbohydrates by using symbolic representation, and that allows precisely locating the sites attacked by the different enzymes by using color-coded arrows pointing to the cleaved bonds within the polymer structure. This type of diagram is much more suitable for describing the degradation of such polymers, which often consists of multiple types of enzymes

simultaneously attacking in parallel different types of bonds within the polymer. Using these new capabilities, we curated several new lignocellulose degradation pathways of complex molecules, and converted existing pathways of this type to use the new format.

Objective 3. We included all of the bioenergy-related microbial genomes listed on the Bioenergy Science Center KnowledgeBase (at http://cricket.ornl.gov/cgi-bin/microb.cgi) in the PGDB collection hosted on the BioCyc.org website. These include biomass degraders, fuel producers, endophytes, model organisms, and four other species. These PGDBs are available for online analysis as well as for download using multiple mechanisms, such as flat data files and the Pathway Tools PGDB Registry.

Part II.

Objective 4.

The subsequent sections describe two new computational techniques developed and implemented in Pathway Tools under this project.

A computational technique, based on linear programming, to infer the atom mappings of biochemical reactions.

The complete atom mapping of a chemical reaction is a bijection of the reactant atoms to the product atoms that specifies the terminus in the product of each reactant atom. We created a new method for computing atom mappings based on the minimum weighted edit-distance (MWED) metric. MWED models can be efficiently formulated as mixed-integer linear programs (MILPs).

An MILP models all possible valid atom mappings, using linear constraints, with the objective of minimizing the sum of the propensity values of the bonds broken or made. A chemist selected the propensity values such that a bond that is likely to form or break has a low value compared to a bond that is unlikely to form or break. For more than 95% of the reactions of MetaCyc, only one optimal solution is found (that is, only one atom mapping is found).

We have demonstrated this technique on the MetaCyc database, for which 87% of the atom mappings could be computed in less than 10 seconds. We have also shown that the error rate was 0.9% (22 reactions) by comparing the generated atom mappings to 2,446 atom mappings found in the manually curated Kyoto Encyclopedia of Genes and Genomes (KEGG) RPAIR database.

The technique is used on a daily basis to infer the atom mappings for any reaction entered by curators into the MetaCyc database. As of September 2014, this technique has computed atom mappings for more than 11,200 reactions found in MetaCyc. The atom mappings are illustrated by coloring conserved structural moieties on the reaction web pages at BioCyc.org, where they allow users to more quickly grasp the chemical transformation occurring within a reaction. Any database, not only MetaCyc, at BioCyc.org can refer to these atom mappings when displaying a reaction. The atom-mapping data are also available as downloadable data files for use by external scientists.

A paper was published describing the computational technique: Latendresse *et al*., “Accurate Atom-Mapping Computation for Biochemical Reactions,” *JCIM* 2012.

An algorithm to compute the optimal metabolic route in a network of reactions of an

organism based on atoms conserved, using the atom mappings and considering the insertion of new genes.

The algorithm assumes that the user begins with a source (feedstock) compound and a target (product) compound. The aim is finding the optimal metabolic routes that connect those compounds within a genome-scale metabolic-reaction network. The algorithm further assumes that the reaction network is annotated with atom mappings, which describe the path of each atom in a metabolic reaction from a reactant compound to a product compound. The algorithm is based on a branch-and-bound searching algorithm. The optimality criterion is the sum of the cost of the reactions used in the route plus the cost of the atoms lost from the source compound to the target compound. The number of atoms lost along the route is computed by using the reaction atom mappings described in the previous section.

The method allows supplementing the starting reaction network of an organism with additional reactions from an external reaction library, which, for our algorithm, is the MetaCyc database. More specifically, the algorithm takes the following parameters as input: a network of reactions of an organism; a start compound; a goal compound; a list of compounds and side compounds to avoid in the routes to find; the cost of losing one atom along the route from the source compound to the target compound; the cost of using a reaction in the organism's metabolic network in the resulting route; and the cost of using a reaction from MetaCyc in the resulting route (if the tool is used with MetaCyc).

We evaluated the algorithm on five metabolic-engineering problems from the literature; for one problem, the published solution was equivalent to the optimal route found by our algorithm; for the remaining four problems, our algorithm found the published solution as one of its best-scored solutions. These problems were each solved in less than five seconds of computational time.

This algorithm is implemented as the RouteSearch tool available on the website BioCyc.org and in the downloadable version of Pathway Tools. RouteSearch includes a graphical user interface that speeds user understanding of its search results. The display shows the top optimal routes with the compounds, reactions, and enzymes involved, and shows the atom mappings at each step of the route.

A paper was published describing the RouteSearch algorithm and tool: Latendresse *et al*., "Optimal Metabolic Route Search Based on Atom Mappings," *Bioinformatics*, 2014.

We also implemented in Pathway Tools the group contribution method of Jankowski *et al.* to estimate both the Gibbs free energy of formation of compounds and the change in Gibbs free energies of metabolic reactions. We applied the algorithm to all compounds and reactions of MetaCyc; the results are available on the BioCyc.org website on the compound and reaction web pages. This implementation will periodically be used to recompute the Gibbs free energies of compounds and reactions for future releases of MetaCyc.