OPTSTRAIN: A HIERARCHICAL METABOLIC PATHWAY DISCOVERY AND DESIGN FRAMEWORK FOR MICROBIAL PRODUCTION SYSTEMS

Priti Pharkya and Costas D. Maranas*
The Pennsylvania State University
University Park, PA, 16802

Abstract

In this contribution, we will discuss a hierarchical procedure termed ***OptStrain*** aimed at guiding pathway modifications, through pathway additions and deletions, of microbial networks for the overproduction of targeted compounds. A comprehensive database of biotransformations, referred to as the Universal database (with over 5,700 reactions), is compiled and regularly updated by downloading and curating reactions from multiple biopathway database sources. Combinatorial optimization is then employed to elucidate the set(s) of non-native functionalities, extracted from this Universal database, to add to the examined production host for enabling the desired product formation. Subsequently, competing functionalities that divert flux away from the targeted product are identified and removed to ensure higher product yields coupled with growth. The range and utility of OptStrain is demonstrated by addressing two very diffferent product molecules, hydrogen and vanillin, which represent the extreme ends of the product size spectrum.

Keywords

Strain design, Genome-scale metabolic models, Optimization

Introduction

The recent availability of genome-scale models of microbial organisms has provided the pathway reconstructions necessary for developing computational methods aimed at identifying strain engineering strategies (Bailey 2001). These models, already available for *H. pylori* (Schilling et al. 2002), *E. coli* (Edwards and Palsson 2000; Reed et al. 2003), *S. cerevisiae* (Forster et al. 2003) and other microorganisms provide successively refined abstractions of the microbial metabolic capabilities. At the same time, individual reactions are deposited in databases such as KEGG, EMP, MetaCyc, and many more (Kanehisa et al. 2004; Karp et al. 2000; Selkov et al. 1998), forming encompassing and growing collections of the biotransformations for which we have direct or indirect evidence of existence in different species. This newly acquired plethora of data has brought to the forefront a number of computational and modeling challenges which form the scope of this article. Specifically, how can we systematically select from the thousands of functionalities catalogued in various biological databases, the appropriate set of pathways/genes to recombine into existing production systems such as *E. coli* so as to endow them with the desired new functionalities? Subsequently, how can we identify which competing functionalities to eliminate to ensure high product yield as well as viability?

Existing strategies and methods for accomplishing this goal include database queries to explore all feasible bioconversion routes from a substrate to a target compound from a given list of biochemical transformations (Mavrovouniotis et al. 1990; Seressiotis and Bailey 1988). More recently, elegant graph theoretic concepts (e.g., P-graphs (Fan et al. 2002) and k-shortest paths algorithm (Eppstein 1994)) were pioneered to identify novel biotransformation pathways based on the

* Corresponding author. Email: costas@psu.edu

tracing of atoms (Arita 2000; Arita 2004), enzyme function rules and thermodynamic feasibility constraints (Hatzimanikatis et al. 2003). Also an interesting heuristic search approach that uses the enzymatic biochemical reactions found in the KEGG database (Kanehisa et al. 2004) to construct a connected graph linking the substrate and the product metabolites was recently proposed (McShan et al. 2003). Most of these approaches, however, generate linear paths that link substrates to final products without ensuring that the rest of the metabolic network is balanced and that metabolic imperatives on cofactor usage/generation and energy balances are met.

In this paper, we will discuss a hierarchical optimization-based framework, ***OptStrain*** to identify stoichiometrically-balanced pathways to be generated upon recombination of non-native functionalities into a host organism to confer the desired phenotype. Candidate metabolic pathways are identified from an ever-expanding array of thousands (currently 5,738) of reactions pooled together from different stoichiometric models and publicly available databases such as KEGG (Kanehisa et al. 2004). Note that the identified pathways satisfy maximum yield considerations while the choice of substrates can be treated as optimization variables. Subsequently, gene deletions are identified (Burgard et al. 2003; Pharkya et al. 2003) in the augmented host networks to improve product yields by removing competing functionalities which decouple biochemical production and growth objectives. The breadth and scope of ***OptStrain*** is demonstrated by addressing in detail two different product molecules (i.e., hydrogen and vanillin).

The OptStrain Procedure

The OptStrain procedure is a four step procedure, each step of which introduces different computational challenges arising from the specific structure and size of the optimization problems that need to be solved.

Step 1: Curation of the database

The first step of the ***OptStrain*** procedure begins with the downloading and curation of reactions acquired from various sources in our Universal database. We have developed customized scripts using Perl (Brown 1999) to automatically download all reactions and parse the number of atoms of each element in every compound. Subsequently, the elementally unbalanced reactions are excluded from consideration. In addition, compounds with an unspecified number of repeat units, or unspecified alkyl groups R in their chemical formulae are removed from the downloaded sets. This step enables the formation of large-scale sets of functionalities to be used as recombination targets.

Step 2: Determination of the maximum yield

Once the reaction sets are determined, the second step is geared towards determining the maximum theoretical yield of the target product from a range of substrate choices, without restrictions on the number or origin of the reactions used. The maximum theoretical product yield is obtained for a unit uptake rate of substrate by maximizing the sum of all reaction fluxes producing minus those consuming the target metabolite, weighted with the stoichiometric coefficient of the target metabolite in these reactions. The maximization of this yield subject to stoichiometric constraints and transport conditions yields a Linear Programming (LP) problem, often encountered in Flux Balance Analysis frameworks (Varma and Palsson 1994).

Step 3: Identification of the minimum number of non-native reactions for a host organism.

The next step in ***OptStrain*** uses the knowledge of the maximum theoretical yield to determine the minimum number of non-native functionalities that need to be added into a specific host organism network. Mathematically, this is achieved by first introducing a set of binary variables y_j that serve as switches to turn the associated reaction fluxes v_j on or off. The corresponding constraints are imposed only on the reactions associated with heterologous genes such that if the reaction j is active, the associated binary variable y_j assumes a value of one and a value of zero if the reaction is inactive. This leads to a Mixed Integer Linear Programming (MILP) model for finding the minimum number of genes to be added into the host organism network while meeting the yield target for the desired product. Alternate optimal solutions can also be identified iteratively at this stage.

Step 4: Incorporating the non-native reactions into the host organism's stoichiometric model.

Upon identification of the appropriate host organism, the analysis proceeds with an organism-specific stoichiometric model augmented with the set of the identified non-native reactions. However, simply adding genes to a microbial production strain will not necessarily lead to the desired overproduction due to the fact that microbial metabolism is primed to be as responsive as possible to the imposed selection pressures (e.g., outgrow its competition). These survival objectives are typically in direct competition with the overproduction of targeted biochemicals. To combat this, we use our previously developed bilevel (Burgard et al. 2003; Pharkya et al. 2003) computational framework, OptKnock to eliminate all those functionalities which uncouple the cellular fitness objective, typically exemplified as the biomass yield, from the maximum yield of the product of interest.

Results

Computational results for microbial strain optimization focus on the production of hydrogen and vanillin. The hydrogen production case study underscores the importance of investigating multiple substrates and microbial hosts to pinpoint the optimal production environment. In contrast, in the vanillin study, identifying the smallest number of non-native reactions is found to be the key challenge for strain design.

Hydrogen Production Case Study

An efficient microbial hydrogen production strategy requires the selection of an optimal substrate and a microbial strain capable of forming hydrogen at high rates. First we solved the maximum yield LP formulation (Step 2) using all catalogued reactions which were balanced with respect to hydrogen, oxygen, nitrogen, sulfur, phosphorus and carbon (approximately 3,000 reactions) as recombination candidates. Different substrates such as pentose and hexose sugars as well as acetate, lactate, malate, glycerol, pyruvate, succinate and methanol were investigated. The highest hydrogen yield obtained for a methanol substrate was equal to 0.126 g/g substrate consumed. This is not surprising given that the hydrogen to carbon ratio for methanol is the highest at four to one. We decided to explore methanol and glucose further, motivated by the high yield on methanol and the favorable costs associated with the use of glucose. The next step in the ***OptStrain*** procedure (Step 3) entailed the determination of the minimum number of non-native functionalities for achieving the theoretical maximum yield in a host organism. We examined two different uptake scenarios: (i) glucose in *Escherichia coli* (an established production system) and (ii) methanol in *Methylobacterium extorquens* (a known methanol consumer). *E. coli* is a natural producer of hydrogen and therefore, no additional functionalities are required. In contrast, using Step 3 of ***OptStrain***, we discovered that *M. extorquens* cannot produce hydrogen by itself and identified that only a single reaction needs to be introduced into its stoichiometric model (Van Dien and Lidstrom 2002) to enable hydrogen production. Two such candidates are hydrogenase (E.C.# 1.12.7.2) which reduces protons to hydrogen or alternatively N_5, N_{10}-methenyltetrahydromethanopterin hydrogenase which catalyzes the following transformation:

E.C.# 1.12.98.2: 5,10-Methylenetrahydromethanopterin $\leftrightarrow$ 5,10-Methenyltetrahydromethanopterin + H_2.

Vanillin Production Case Study

Vanillin is an important flavor and aroma molecule. In this case study, we identify metabolic network redesign strategies for the *de novo* production of vanillin from glucose in *E. coli*. Using ***OptStrain***, we first determined the maximum theoretical yield of vanillin from glucose to be 0.63 g/g glucose by solving the LP optimization over 4,270 candidate reactions balanced with respect to all elements but hydrogen (Step 2). We next identified that the minimum number of non-native reactions that must be recombined into *E. coli* to endow it with the pathways necessary to achieve the maximum yield is three (Step 3). Numerous alternative pathways, differing only in their cofactor usage, which satisfy both the optimality criteria of yield and minimality of recombined reactions, were identified. For example, one such pathway uses the following three non-native reactions:
(i) E.C.# 1.2.1.46: Formate+NADH+H^+ $\leftrightarrow$ Formaldehyde + NAD^+ + H_2O,
(ii) E.C.# 1.2.3.12: 3, 4-dihydroxybenzoate + NAD^+ + H_2O + Formaldehyde $\leftrightarrow$ Vanillate + O_2 + NADH,
(iii) E.C.# 1.2.1.67: Vanillate + NADH + H^+ $\leftrightarrow$ Vanillin + NAD^+ + H_2O.

Interestingly, these steps are essentially the same as those used in the experimental study by Li and Frost (Li and Frost 1998) to convert glucose to vanillin in recombinant *E. coli* cells demonstrating that the computational procedure can indeed uncover relevant engineering strategies. Note, however, that the reported experimental yield of 0.15 g/g glucose is far below the maximum theoretical yield (i.e., 0.63 g/g glucose) of the network indicating the potential for considerable improvement.

This motivates examining whether it is possible to reach higher yields of vanillin by systematically pruning the metabolic network using OptKnock (Step 4). Here the most recent genome-scale model of *E. coli* metabolism (Reed et al. 2003), augmented with the three functionalities identified above, was integrated into the OptKnock framework to determine the set(s) of reactions whose deletion would force a strong coupling between growth and vanillin production. The highest vanillin-yielding quadruple knockout strategy is discussed next for a basis glucose uptake rate of 10 mmol/gDW/hr. A substantially high level of vanillin production is predicted in this mutant network. Note also that an anaerobic growth environment was selected by OptKnock for overproducing vanillin. This strategy leads to the production of 6.17 mmol/gDW/hr of vanillin or 0.52 g/g glucose at the maximum growth rate of 0.056 hr^{-1}. The OptKnock framework suggested the deletion of phosphoenolpyruvate carboxylase (E.C.# 4.1.1.31), lactate dehydrogenase (E.C.# 1.1.1.28), pyruvate formate lyase (E.C.# 2.3.1.54), and acetaldehyde dehydrogenase (E.C.# 1.2.1.10). These deletions eliminate the competing byproduct production

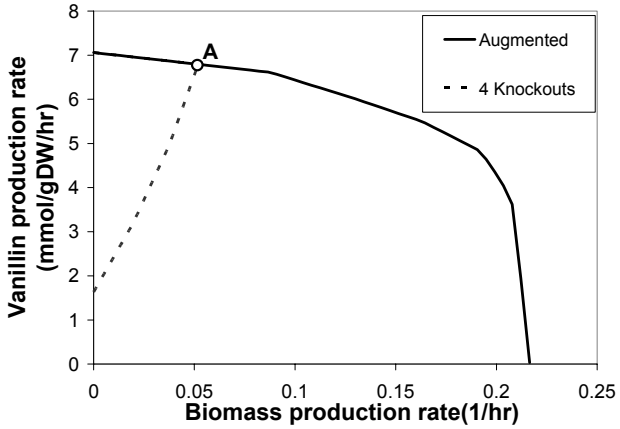

Figure 1: Vanillin production envelope of the augmented E. coli metabolic network for a basis 10 mmol/gDW/hr uptake rate of glucose. Point A denotes the maximum growth point for the four reaction deletion mutant network. In contrast to the wild-type network for which vanillin production is not guaranteed at any rate of biomass production, the mutant networks require significant vanillin yields to achieve high levels of biomass production.

routes for ethanol, formate, and lactate. Furthermore, a surprising network flux redistribution involves the utilization of a group of reactions from one-carbon metabolism to form 10-formyltetrahydrofolate, which is subsequently converted to formaldehyde. Figure 1 compares the vanillin production envelopes, obtained by maximizing and minimizing vanillin formation at different biomass production rates for the wild-type and the mutant networks. These deletions endow the network with high levels of vanillin production under any growth conditions.

Discussion

The ***OptStrain*** framework is aimed at systematically reshaping whole genome-scale metabolic networks of microbial systems for the overproduction of not only small but also complex molecules. We have so far examined a number of different products (e.g., 1, 3 propanediol, inositol, pyruvate, etc.) using a variety of hosts (i.e., *E. coli, Clostridium acetobutylicum, M. extorquens*). The two case studies, hydrogen and vanillin, discussed earlier show that ***OptStrain*** can address the range of challenges associated with strain redesign. At the same time, it is important to emphasize that the validity and relevance of the results obtained with the ***OptStrain*** framework are dependent on the level of completeness and accuracy of the reaction databases and microbial metabolic models considered. We have identified numerous instances of unbalanced reactions and ambiguous reaction directionality in the reaction databases that we mined. Careful curation of the downloaded reactions preceded all of our case studies. Whenever the balanceability of a reaction could not be restored, the reaction was removed from consideration. The purely stoichiometric representation of metabolic pathways in microbial models can lead to unrealistic flux distributions by not accounting for kinetic barriers and regulatory interactions (e.g., allosteric regulation). Despite these simplifications, ***OptStrain*** has already provided useful insight into microbial host redesign in many cases and, more importantly, established for the first time an integrated framework open to future modeling improvements.

Acknowledgement

Financial support by the NSF Award BES0120277 is gratefully acknowledged.

References

Arita M. 2000. Metabolic construction using shortest paths. Simulation Practice and Theory 8:109-125.

Arita M. 2004. The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A 101(6):1543-7.

Bailey JE. 2001. Complex biology with no parameters. Nat Biotechnol 19(6):503-4.

Brown M. 1999. Perl programmer's reference. Berkeley, Calif.: Osborne/McGraw-Hill. xix, 380 p.

Burgard AP, Pharkya P, Maranas CD. 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84(6):647-57.

Edwards JS, Palsson BO. 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97(10):5528-33.

Eppstein D. Finding the k shortest paths; 1994; Santa Fe. p 154-165.

Fan LT, Bertok B, Friedler F. 2002. A graph-theoretic method to identify candidate mechanisms for deriving the rate law of a catalytic reaction. Comput Chem 26(3):265-92.

Forster J, Famili I, Fu P, Palsson BO, Nielsen J. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13(2):244-53.

Hatzimanikatis V, Li C, Ionita JA, Broadbelt LJ. 2003. A Computational Framework for the Discovery of Novel Biobased Chemicals. presented at Biochemical Engineering XIII Conference, Session 2:Boulder, CO.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res 32 Database issue:D277-80.

Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. 2000. The EcoCyc and MetaCyc databases. Nucleic Acids Res 28(1):56-9.

Li K, Frost JW. 1998. Synthesis of vanillin from glucose. Journal of American Chemical Society 120:10545-10546.

Mavrovouniotis M, Stephanopoulos G, Stephanopoulos G. 1990. Computer-Aided Synthesis of Biochemical Pathways. Biotechnol Bioeng 36:1119-1132.

McShan DC, Rao S, Shah I. 2003. PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics 19(13):1692-8.

Pharkya P, Burgard AP, Maranas CD. 2003. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. Biotechnol Bioeng 84(7):887-99.

Reed JL, Vo TD, Schilling CH, Palsson BO. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54.

Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO. 2002. Genome-scale metabolic model of Helicobacter pylori 26695. J Bacteriol 184(16):4582-93.

Selkov E, Jr., Grechkin Y, Mikhailova N, Selkov E. 1998. MPW: the Metabolic Pathways Database. Nucleic Acids Res 26(1):43-5.

Seressiotis A, Bailey JE. 1988. MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. Biotechnol Bioeng 31:587-602.

Van Dien SJ, Lidstrom ME. 2002. Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph Methylobacterium extorquens AM1, with application to reconstruction of C(3) and C(4) metabolism. Biotechnol Bioeng 78(3):296-312.

Varma A, Palsson BO. 1994. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Bio/Technology 12:994-998.