BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 21 2008, pages 2554–2556
doi:10.1093/bioinformatics/btn471

Systems biology

DESHARKY: automatic design of metabolic pathways for optimal cell growth

Guillermo Rodrigo[1], Javier Carrera[1,2], Kristala Jones Prather[3] and Alfonso Jaramillo[4,5,*]

[1]Instituto de Biologia Molecular y Celular de Plantas, CSIC, [2]Instituto de Aplicaciones en Tecnologias de la Informacion y las Comunicaciones Avanzadas (ITACA), Universidad Politecnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain, [3]Department of Chemical Engineering, Massachusetts Institute of Technology, Massachusetts Avenue 77, Cambridge MA 02139, USA, [4]Laboratoire de Biochimie, Ecole Polytechnique - CNRS, Route de Saclay, 91128 Palaiseau Cedex and [5]Epigenomics Project, Genopole, 523 Terrasses de l'Agora, 91034 Evry Cedex, France

Received on May 28, 2008; revised on August 14, 2008; accepted on September 02, 2008

Advance Access publication September 5, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The biological solution for synthesis or remediation of organic compounds using living organisms, particularly bacteria and yeast, has been promoted because of the cost reduction with respect to the non-living chemical approach. In that way, computational frameworks can profit from the previous knowledge stored in large databases of compounds, enzymes and reactions. In addition, the cell behavior can be studied by modeling the cellular context.

Results: We have implemented a Monte Carlo algorithm (DESHARKY) that finds a metabolic pathway from a target compound by exploring a database of enzymatic reactions. DESHARKY outputs a biochemical route to the host metabolism together with its impact in the cellular context by using mathematical models of the cell resources and metabolism. Furthermore, we provide the sequence of amino acids for the enzymes involved in the route closest phylogenetically to the considered organism. We provide examples of designed metabolic pathways with their genetic load characterizations. Here, we have used *Escherichia coli* as host organism. In addition, our bioinformatic tool can be applied for biodegradation or biosynthesis and its performance scales with the database size.

Availability: Software, a tutorial and examples are freely available and open source at http://soft.synth-bio.org/desharky.html

Contact: alfonso.jaramillo@polytechnique.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Biotechnology process development is frequently equated with the production of biologics, such as proteins and viral vaccines (Nielsen, 2001). Yet the use of biological systems for the production of small molecules goes back thousands of years and has been increasing since the discipline of metabolic engineering was defined 15 years ago (Bailey, 1991). Initially, metabolic engineering efforts were primarily focused on improving the productivity of naturally-occurring metabolites within an organism, such as for overexpressing glycolytic enzymes in yeast (Schaaff *et al*., 1989). More recently, the field has expanded to encompass a number of examples of introducing new enzyme activities into a host cell in order to produce non-natural products (Martin *et al*., 2003; Ro *et al*., 2006) or to engineer degradation of toxic compounds (Haro and de Lorenzo, 2001).

The use of automated techniques to design biological systems constitutes a breakthrough in biotechnology, and it has previously been applied to predict biodegradation pathways (Hou *et al*., 2003; Pazos *et al*., 2005). Interestingly, functional approaches (Hatzimanikatis *et al*., 2005; Hou *et al*., 2003; Li *et al*., 2004) could reveal novel pathways, but these are ultimately limited by the availability of naturally-occurring enzymes. In that sense, recent work shows how to construct biochemical pathways using atomic information (Arita, 2003, 2004), and this approach could be used to enlarge our enzyme database by adding abstract reactions corresponding to functional enzymes. This would allow the design of metabolic pathways that incorporate enzymes not found in nature but which could be engineered by directed evolution or using computational design (Rothlisberger *et al*., 2008). In this work we propose to go beyond by extending the design to biosynthesis and predicting the cell behavior when implementing a pathway in a given host using plasmids (Jones *et al*., 2000).

On the other hand, one of the major challenges in synthetic biology is engineering as far as possible orthogonal systems (Sprinzak and Elowitz, 2005). In that way, quantitative models provide fruitful insights. We propose the use of two different models to quantify the readjustment of fluxes (Varma and Palsson, 1994) and the consumption of cellular resources (Bremer and Dennis, 1996) that results from the expression of heterologous pathways. We select the growth rate as the control parameter for the cellular behavior evaluation. From the transcriptional approach, we consider a dynamical model involving RNAs, RNA polymerases, proteins and ribosomes (Carrera J. *et al*., manuscript in preparation). Accordingly, we compute the reduction in the growth rate due to the sequestration of RNA polymerases and ribosomes. On the other hand, since the cell is metabolically altered, we use Flux Balance Analysis (FBA) to predict the new growth rate. These two strategies give different predictions about the cell behavior, but they constitute two

[*] To whom correspondence should be addressed.

scores to be considered when implementing a designed pathway. Further approaches will use more complex models by integrating the metabolic and transcriptomic systems, and also taking advantage of databases of Gibbs free energies for all enzymatic reactions (Mavrovouniotis, 1991). Importantly, as the desired route could be not unique, we provide a methodology to rank different pathways according to their genetic loads.

2 METHODS

2.1 Algorithm

We have developed a Monte Carlo algorithm (DESHARKY) with the aim of designing metabolic pathways. The purpose is to find a possible route connecting a given compound of interest with a metabolite from the considered hosting organism. These routes can be for biodegradation (reactant as source) or biosynthesis (product as source). For the source compound, we find the possible enzymatic reactions and select one among them with equitable probabilities. We repeat this process for the new source compound. Moreover, we consider with a given probability a move to go back, removing the previous reaction, to improve the convergence and to avoid long pathways. This probability is a function of the number of the already introduced steps, as the longer the pathway, the higher is the probability to go back, and here we have used a sigmoid function. We do not consider metabolic steps involving many compounds which are not specific to the hosting organism (here, one non-specific reactant and one product at most).

2.2 Transcription–translation model

The microbial production or degradation of chemical compounds usually requires the expression of foreign enzymes. This expression consumes cellular resources such as RNA polymerases and ribonucleotides for transcription, and ribosomes and amino acids for translation. Using previous knowledge on heterologous expression, we assume that RNA polymerases and ribosomes are the two critical pools. Using the experimental measurements of these resources in *Escherichia coli* (Bremer and Dennis, 1996), we have constructed a chassis model (Carrera *et al.*, manuscript in preparation), fitting those data with exponential equations (see caption of Fig. 1). Furthermore, we have modeled the total heterologous expression of RNA (RNA_h) by

$$\frac{\mathrm{d}}{dt}RNA_h = \phi C - \delta_r RNA_h, \tag{1}$$

and enzymes (ENZ_h) following

$$\frac{\mathrm{d}}{dt}ENZ_h = \psi RNA_h - \delta_e ENZ_h, \tag{2}$$

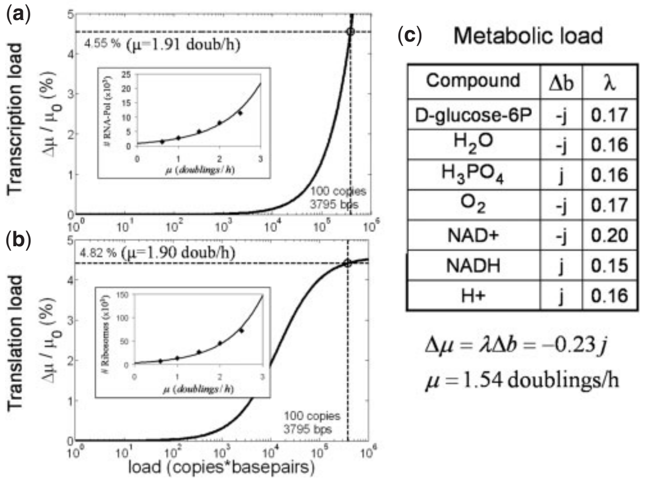
where ϕ is the average transcription rate, C the number of copies of external DNA, ψ the average translation rate, and δ_r and δ_e the degradation rates of the RNA and enzymes, respectively. Hence, a first order approach is to compute the consumption of cellular resources by the heterologous system ($RNAP_h = \phi C t_r$ and $RIB_h = \psi RNA_h t_p$, where t_r is the transcription time and t_p the translation time) and then to recompute the growth rate using the phenomenological chassis model (Fig. 1). We take the minimum value of μ throughout these resources.

2.3 Metabolic model

We have addressed the metabolic burden with FBA (Varma and Palsson, 1994). This linear program, in which we maximize the cell growth rate (μ), can be written as

$$\begin{gathered}\max\ \mu = cv \\ \text{subject to } Sv = b,\end{gathered} \tag{3}$$

where v are the cell metabolic fluxes, c their contributions to the growth rate, S the stoichiometry matrix, and b the uptake fluxes. Then, we have constructed the corresponding dual problem (Schrijver, 1998), which is equivalent to its primal, given by min $\mu = \lambda b$, subject to $\lambda S = c$, where λ, usually called shadow prices, are the contributions to the growth rate when perturbing the uptake fluxes ($\Delta\mu = \lambda\Delta b$). Therefore, we can precompute λ since it is a property of the host organism. In that way, the fact of introducing a new metabolic route in the host can be treated in a perturbative way. Then, $\Delta b = S^{*}j$ where S^{*} is the stoichiometry matrix for this pathway and j its flux.



Compound	Δb	λ
D-glucose-6P	-j	0.17
H_2O	-j	0.16
H_3PO_4	j	0.16
O_2	-j	0.17
NAD+	-j	0.20
NADH	j	0.15
H+	j	0.16

$$\Delta\mu = \lambda\Delta b = -0.23j$$

$$\mu = 1.54 \text{ doublings/h}$$

Fig. 1. Genetic load characterization of the glucaric acid biosynthesis pathway (see Table 1). In (**a**) transcription load assuming a plasmid copy number of 100. In the inset, amount of RNA polymerases as a function of the cell growth rate given by $RNAP = 910exp(1.06\mu)$ (diamonds are experimental measurements). In (**b**) translation load. In the inset, amount of ribosomes as a function of the cell growth rate given by $RIB = 3690exp(1.23\mu)$ (diamonds are experimental measurements). In (**c**) metabolic load: list of the shadow prices for all cofactors required in that pathway and the source compound (D-Glucose-6-Phosphate).

2.4 Implementation

DESHARKY is implemented in C/C++, it is easily compiled, and it runs in UNIX environments (e.g. in Linux or in Windows using Cygwin). Here we have taken *E. coli* as the cell model. We have used an extended description of *E. coli* metabolism involving 1039 compounds, including extracellular compounds, and 2381 biochemical reactions (Schuetz *et al.*, 2007). We provide the KEGG (Kanehisa and Goto, 2000) databases for chemical compounds and enzymatic reactions in a depured format. There are 14 965 chemical compounds, of which 826 are present in the host, 4942 enzymes, of which 2350 have available their sequence, and 7400 enzymatic reactions from 650 organisms. Also we consider a set of compounds eventually in the medium that can be used as substrates by the cell. To enlarge the capabilities of the algorithm, we can assume reversible reactions. In addition, we can introduce reactions which are not found in KEGG. The input of our algorithm is the target compound. The output is the designed metabolic pathway together with the quatification of the transcription, translation and metabolic load. In addition, we provide the sequence of amino acids of the enzymes involved in the pathway. These sequences are the closest phylogenetically to *E. coli* according to the KEGG classification of organisms.

Here we have assumed an initial growth rate of $\mu_0 = 2$ doublings/h, a transcription kinetics of $\phi = 0.1$ RNA polymerases/s, a translation kinetics of $\psi = 0.4$ ribosomes/s, a number of DNA copies for the enzymes of $C = 100$, a transcription velocity of $1/t_r = 45$ nt/s, a translation velocity of $1/t_p = 16$ aa/s, and a metabolic pathway flux of $j = 1$ mmol/gDW/h.

Table 1. Examples of metabolic pathways designed with DESHARKY

Pathway	Host Metabolite[a]	Enzymes[b]	Trans. Load[c](%)	Met. Load[d](%)
Biodegradation of toluene	*p*-Benzenediol	1.14.13.7, 1.14.13, 1.1.1.97, 1.2.1.7, 1.14.13.24, 4.1.1.62	6.65	14.90
Biodegradation of phenol	4-Hydroxy-2-Oxopentanoate	1.14.13.7, 1.13.11.2, 1.2.1.32, 5.3.2, 4.1.1	5.96	−22.82
Biosynthesis of sorbitol	D-Fructose-6-Phosphate	1.1.1.14, 2.7.1.1	5.12	21.50
Biosynthesis of glucaric acid	D-Glucose-6-Phosphate	5.5.1.4, 3.1.3.25, 1.13.99.1, 1.1.1.203	4.82	11.17

[a] Metabolite present in the hosting organism which is the initial substrate in case of biosynthesis and final product in case of biodegradation.
[b] EC number of the enzymes involved in the designed pathway. Note that the enzyme 1.1.1.203 for glucaric acid production has no data available in KEGG.
[c] Transcriptomic load: cell growth rate reduction due to the consumption of cellular resources, such as RNA polymerases or ribosomes, to express the enzymes.
[d] Metabolic load: cell growth rate reduction due to the readjustment of the cellular metabolic fluxes when implementing a new biochemical route.

3 RESULTS AND DISCUSSION

We have applied DESHARKY to design several metabolic pathways including biodegradation of toluene or phenol and bioproduction of sorbitol and glucaric acid (Table 1). For instance, the microbial production of glucaric acid is important for therapeutic purposes including cholesterol reduction and cancer chemotherapy, and for the synthesis of new nylons and hyperbranched polyesters. In Figure 1 we show the transcription, translation and metabolic load for this pathway, and in the Supplementary Figure S1 we depict the biochemical transformations and the list of genes encoding the corresponding enzymes. In addition, in the Supplementary Material we have compared the biodegradation pathways we found with those obtained from UM-BBD (Hou *et al*., 2003) showing alternative routes.

Our tool uses a heuristic algorithm based on Monte Carlo to find a possible route connecting a specified target metabolite with the host metabolism, instead of using a pathway selection by enumeration of all possible metabolic routes (Arita, 2003; Eppstein, 1998). DESHARKY finds a proper pathway and computes its associated genetic load in a few seconds. In addition, our software can be used in distributed computing to sample most of the solution space. For illustration purposes, we show in the Supplementary Material all possible biodegradation routes for phenol. Here, we have assumed non-weighted reactions for the heuristic procedure and we compute the genetic load *a posteriori* using the transcription and metabolic models. Alternatively, a global optimization could be addressed by considering the load of each reaction during the heuristic procedure (Croes *et al*., 2006).

ACKNOWLEDGEMENTS

We thank T. S. Moon for his help, and B. Canton and D. Endy for their fruitful comments on the chassis model.

Funding: Generalitat Valenciana G.R. (BFPI 2007/160 to G.R.); Spanish Ministry of Education (TIN 2006-12860); MIT-France program, Structural Funds ERDF; EU grant BioModularH2 (FP6-NEST-043340).

Conflict of Interest: none declared.

REFERENCES

Arita,M. (2003) In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Res.*, **13**, 2455–2466.

Arita,M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl Acad. Sci. USA*, **101**, 1543–1547.

Bailey,J.E. (1991) Toward a science of metabolic engineering. *Science*, **252**, 1668–1675.

Bremer,H. and Dennis,P.P. (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In Neidhardt,F.C. *et al*. (eds.) *Escherichia coli and Salmonella*, Vol. 2, 2nd edn. ASM Press, Washington, D.C, pp. 1553–1569.

Croes,D. *et al*. (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

Eppstein,D. (1998) Finding the k shortest paths. *SIAM J. Comput.*, **28**, 652–673.

Haro,M.-A. and de Lorenzo,V. (2001) Metabolic engineering of bacteria for environmental applications: construction of *Pseudomonas* strains for biodegradation of 2-chlorotoluene. *J. Biotechnol.*, **85**, 103–113.

Hatzimanikatis,V. *et al*. (2005) Broadbelt. Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.

Hou,B.K. *et al*. (2003) Microbial pathway prediction: a functional group approach. *J. Chem. Inf. Comput. Sci.*, **43**, 1051–1057.

Jones,K.L. *et al*. (2000) Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metabolic Engineering*, **2**, 328–338.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Li,C. *et al*. (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.*, **59**, 5051–5060.

Martin,J.J. *et al*. (2003) Engineering a mevalonate pathway in *escherichia coli* for production of terpenoids. *Nat. Biotech.*, **21**, 796–802.

Mavrovouniotis,M.L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.*, **266**, 14440–14445.

Nielsen,J. (2001) Metabolic engineering. *Appl. Microbiol. Biotechnol.*, **55**, 263–283.

Pazos,F. *et al*. (2005) MetaRouter: bioinformatics for bioremediation. *Nucleic Acids Res.*, **33**, D588–D592.

Ro,D.K. *et al*. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940–943.

Rothlisberger,D. *et al*. (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, **453**, 190–195.

Schaaff,I. *et al*. (1989) Overproduction of glycolytic enzymes in yeast. *Yeast*, **5**, 285–290.

Schrijver,A. (1998) *Theory of Linear and Integer Programming*. John Wiley & Sons, New York.

Schuetz,R. *et al*. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molec. Syst. Biol.*, **3**, 119.

Sprinzak,D. and Elowitz,M.B. (2005) Reconstruction of genetic circuits. *Nature*, **438**, 443–448.

Varma,A. and Palsson,B.O. (1994) Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, **12**, 994–998.