Retrosynthetic design of heterologous pathways

Pablo Carbonell, Anne-Gaëlle Planson, Jean-Loup Faulon

Retrosynthetic Design of Heterologous Pathways

Systems Metabolic Engineering pp 149-173

Part of the Methods in Molecular Biology book series (MIMB, volume 985)

- Pablo Carbonell (1)
- Anne-Gaëlle Planson (1)
- Jean-Loup Faulon (1) Email author (jean-loup.faulon@issb.genopole.fr)

1. Institute of Systems & Synthetic Biology (ISSB), Evry, France

Protocol

First Online:
17 January 2013

Abstract

Tools from metabolic engineering and synthetic biology are synergistically used in order to develop high-performance cell factories. However, the number of successful applications has been limited due to the complexity of exploring efficiently the metabolic space for the discovery of candidate heterologous pathways. To address this challenge, retrosynthetic biology provides an integrated framework to formalize and rationalize the problem of importing biosynthetic pathways into a chassis organism using methods at the interface from bottom-up and top-down strategies. Here, we describe step by step the process of implementing a retrosynthetic framework for the design of heterologous biosynthetic pathways in a chassis organism. The method consists of the following steps: choosing the chassis and the target, selection of an *in silico* model for the chassis, definition of the metabolic space, pathway enumeration, gene selection, estimation of yields, toxicity prediction of pathway metabolites, definition of an objective function to select the best pathway candidates, and pathway implementation and verification.

Key words

Synthetic biology Metabolic engineering Retrosynthesis Metabolic pathway

Retrosynthetic design of heterologous pathways

Pablo Carbonell, Anne-Gaëlle Planson, and Jean-Loup Faulon

Abstract

Tools from metabolic engineering and synthetic biology are synergistically used in order to develop high-performance cell factories. However, the number of successful applications has been limited due to the complexity of exploring efficiently the metabolic space for the discovery of candidate heterologous pathways. To address this challenge, retrosynthetic biology provides an integrated framework to formalize and rationalize the problem of importing biosynthetic pathways into a chassis organism using methods at the interface from bottom-up and top-down strategies. Here, we describe step by step the process of implementing a retrosynthetic framework for the design of heterologous biosynthetic pathways in a chassis organism. The method consists of the following steps: choosing the chassis and the target; selection of an *in silico* model for the chassis; definition of the metabolic space; pathway enumeration; gene selection; estimation of yields; toxicity prediction of pathway metabolites; definition of an objective function to select the best pathway candidates; and pathway implementation and verification.

Key words: synthetic biology, metabolic engineering, retrosynthesis, metabolic pathway

1. Introduction

Production of value-added compounds such as drugs or biofuels in chassis organisms often requires importing heterologous genes to build efficient biosynthetic pathways. Computational techniques have provided useful methods for the development of such cell factories through a streamlined process for modeling and design of a complete pathway at each step of its construction. Recent advances in systems and synthetic biology are further enabling a systematic practice through an integrated computational framework for rational

biosynthetic pathway design. Nowadays, genome-scale metabolic network reconstructions providing an accurate *in silico* model of the metabolism are available for many industrial chassis organisms ***(1)***. In such models, the effects on steady-state fluxes of metabolic interventions like enhancement of substrate uptake, supplements addition, reduction of undesirable by-products fluxes, introduction of heterologous pathways, or product export to the extracellular medium ***(2)*** can be predicted with a remarkable degree of agreement with experimental observations ***(3)***. Furthermore, an increasingly formalization of the space of biochemical transformations in metabolic networks is allowing the designer to explore creative ways to implement alternative biosynthetic pathways ***(4)***.

To that end, metabolic modeling for heterologous pathway design can be done from two complementary approaches: a topological approach using hypergraphs, where catalytic reactions are hyperedges connecting node substrates to products; and a steady-state approach, where stoichiometry of reactions is used in order to study the properties of all feasible equilibrium states. Knowledge-based comparative analysis, graph search algorithms, and constraint-based models are alternative approaches used in order to infer meaningful pathways in metabolic networks, even if there is missing information on the enzymes.

Several metabolic databases with rich information are available: one of the most comprehensive is MetaCyc and its associated BioCyc collection of pathway/genome databases ***(5)***; similarly, KEGG is a database resource that integrates genomics, chemical and systemic functional information ***(6)***; BRENDA is another database that contains one of the most complete collections of enzyme functional data ***(7)***. Gaps or incomplete knowledge, however, are still present in many cases, especially when looking for novel ways to synthesize compounds. In this regard, computational approaches can provide new alternatives by predicting putative heterologous pathways producing the target compound. In order to successfully handle this challenging task, the design process needs to be rationalized by

following the principles of synthetic biology: modeling of the biological system of interest, modular design through standardization, goal-oriented optimization, and experimental validation. To contribute to this endeavor, we present here a retrosynthetic design approach that aims to provide a streamlined methodology for addressing the general problem of obtaining successful high-yield production of target compounds in cell factories.

1.1 The retrosynthetic framework for heterologous pathway design

Retrosynthesis algorithms are applied to metabolic networks in order to perform a backwards search from the target compound to the host metabolites through the iterative application of a defined set of biochemical transformation rules. Depending on the level of atomic resolution of those rules, recruited enzymes in the biosynthetic pathways may involve novel compound intermediates and putative reactions with unknown efficiency. A successful expression of those genes in the chassis organism needs to be addressed. Therefore, subsequent optimization of the engineered strain through genetic, metabolic and enzyme design approaches would be usually necessary in order to maximize production yields of the target.

We present here a unified framework that combines several techniques involved in the design of heterologous biosynthetic pathways through a retrosynthetic biology approach, enabling by these means the flexible design of industrial microorganisms for the efficient on-demand production of chemical compounds of interest. The method for retrosynthetic design of heterologous pathways consists of the following steps:

1. First, the problem is defined by choosing the chassis organism and the target compound.
2. Second, an *in silico* reconstructed model of the organism containing at least the stoichiometric reactions involved in its metabolism is defined from biological databases and literature.

3. Third, the metabolic space is constructed from all known metabolic reactions and expanded to putative promiscuous reactions.
4. Fourth, heterologous pathways producing the target compound from endogenous metabolites in the chassis are enumerated using a retrosynthetic algorithm.
5. Fifth, gene sequences encoding heterologous enzymes are chosen in order to maximize gene expression and enzyme performance in the chassis organism.
6. Sixth, steady state fluxes for each pathway are estimated through flux balance analysis.
7. Seventh, toxicity of intermediate metabolites are estimated by using a QSAR model.
8. Eight, a cost function is defined for the pathway and the best pathways are chosen.
9. Ninth, selected pathways are implemented and their efficiency is verified.

2. Materials

The following list provides a review of the main metabolic engineering tools used at each step of the heterologous pathway design, including some specific tools for retrosynthesis design:

1. Choosing the chassis: a host organism optimized for metabolic engineering, such as strains from *Escherichia coli*, *Bacillus subtilis* or *Saccharomyces cerivisiae*.
2. Selecting an *in silico* model of the chassis organism: from repositories of *in silico* model organisms like the databases BIGG ***(8)*** or BioModels ***(9)***.
3. Construction of the metabolic space: metabolic databases such as MetaCyc ***(5)*** or KEGG ***(6)*** and enzymatic activity databases such as Brenda ***(7)***.
4. Pathway enumeration: software MetaHype ***(10)***.
5. Gene selection: genomics databases (e.g. UniProt, NCBI Entrez) focused on protein (enzyme) families.

6. Flux balance analysis: metabolic analysis software (e.g. COBRA ***(11)***, OptFlux ***(12)***, COPASI ***(13)***).
7. Metabolite toxicity data for the chassis, either experimental or predicted (e.g. EcoliTox ***(14)***).
8. Definition of a final cost function (e.g. RetroPath ***(15)***).
9. Experimental implementation:
    a) Molecular biology reagents for PCR and cloning.
    b) Bacterial strains for cloning and expression.
    c) Expression vectors.
    d) Growth media.
    e) Analytical techniques for protein identification (electrophoresis gel).
    f) Chromatography system.
    g) Analytical system for metabolite identification.

3. Methods

The design methodology for any metabolic engineering application starts with the selection of a chassis organism along with an associated genome-scale *in silico* model of its metabolic network (*see* **Note 1**). The retrosynthetic approach offers as well the possibility of performing an additional preliminary modeling step to expand the starting metabolic reaction space and, thus, increasing the possibility of discovering novel biosynthetic routes. These prior steps will provide the designer with a detailed knowledge base about the metabolic system that can be used advantageously at later stages of the design. In the same fashion, target compounds need to be defined at this stage.

A basic methodology for retrosynthetic design of heterologous pathways will consist of the following steps (**Fig. 1**): 1) choosing the chassis; 2) selecting an *in silico* model for the

chassis; 3) definition of the metabolic space; 4) pathway enumeration; 5) gene selection; 6) estimation of yields; 7) toxicity prediction of pathway metabolites; 8) definition of an objective function to select the best pathway candidates; 9) pathway implementation and verification.

3.1. Choosing the chassis

An early decision that necessarily influences the rest of the retrosynthetic design process is the choice of the chassis organism where the desired compound will be produced. For example, in order to increase the production of a compound naturally produced in plants, its biosynthetic pathway, if known, is imported into an industrial chassis organism. Factors that need to be considered when choosing the chassis include the following:

1. The extent and level of curation of the organism's metabolic pathways in databases.
2. The availability of a genome-wide reconstructed *in silico* model that has been experimentally verified ***(16)***, and that is ready to be used in constraint-based modeling to quantitatively estimate steady-state fluxes (*see* **Note 2**).
3. The availability of information about toxicity effects of heterologous metabolite intermediates in the organism.
4. The fact that biosynthetic pathways may involve large enzymatic complexes (such as polyketide synthases or non-ribosomal synthases for secondary metabolite synthesis) ***(17)***.
5. Similarly, specific redox reactions catalyzed by the CYP450, which is often needed in the last steps of metabolite synthesis, add another layer of complexity because of the difficulty to model these reactions and often a need to optimize further its catalytic activity through protein engineering ***(18)***.

3.2. Selecting an in silico model for the chassis

In silico organisms models are currently available for many industrial strains, including strains evolved for efficient production in *Escherichia coli*, *Saccharomyces cerevisiae* or *Bacillus Subtilis* ***(19-21)***. Most of them have been deposited in open databases such as BIGG ***(8)*** or BioModels ***(9)*** and numerous tools exist for their analysis and simulation ***(3)***.

Example of chassis selection for production of resveratrol in *E. coli*. Resveratrol (3,5,4'-trihydroxy-trans-stilbene) is a plant phenolic compound with important associated health benefits like prevention of cardiovascular diseases, cancer and promotion of longevity in several animal systems ***(22)***. Resveratrol, however, is only found in a limited number of plant species, including grape (*Vitis sp.*) and peanut (*Aracis hypogaea*). Because of its beneficial properties, there is an increasing interest in the optimization of the production of resveratrol in microorganisms ***(23,24)***. Interestingly, *E. coli* provides an industrial chassis organism with one of the best characterized *in silico* models ***(19)***. As shown in **Fig. 2**, production of resveratrol is derived from phenylalanine that is transformed into cinnamic acid by phenyalanine ammonia lyase (PAL, EC 4.3.1.24). Next, cinnamic acid is transformed into 4-coumaric acid by cinnamate-4-hydroxylase (C4H, EC 1.14.13.11), which is further transformed into coumaroyl-CoA by the 4-coumarate:coenzyme A (CoA) ligase (4CL, EC 6.2.1.12). Then, the stilbene synthase (STS, EC 2.3.1.95) condenses the coumaroyl-CoA and three units of malonyl-CoA to form resveratrol ***(23)***. In the rest of this chapter, we will present the different steps of a retrosynthetic design methodology for metabolic engineering of resveratrol production in *E. coli*.

3.3. Definition of the retrosynthetic metabolic space

The power of retrosynthesis for heterologous pathway design resides in the way representations of chemical biotransformations can provide a generalization of important

chemical features. The most valuable information (but also the most challenging) obtained from retrosynthesis analysis is the identification of putative metabolic pathways involving promiscuous biochemical transformations that often had not yet been well annotated. Several techniques have been proposed in order to expand the metabolic reaction space to contain such putative reactions. The basic idea is to use a set of reaction rules from which not only known reactions can be generated but also novel reactions. To define reaction rules, several representations have been used, such as those derived from bond-electron matrices ***(25)*** (*see* **Note 3**), or on the smallest molecular substructure that can be modified through the transformation ***(26)***. The combinatorial complexity associated with such representations is a major issue that we have recently addressed by proposing a tradeoff solution based on molecular signatures ***(15)***. The main advantage of the molecular signature method relies on the control of the complexity of the pathway search through the selection of the level of specificity in the reaction representation.

For any reaction found through reaction rules, there are two essential questions to consider: a) Is it a putative reaction or has it been reported previously in the databases? b) Is there any known enzyme sequence annotated for such biochemical transformation? In addition, as much information as possible about the selected reaction needs to be collected, including if it has been observed for some enzyme as a promiscuous or side reaction, or what are their kinetics constants, cell localization, or phylogenetic diversity. Main sources for such information are enzymatic databases like BRENDA ***(7)*** and metabolic databases like MetaCyc ***(5)*** or KEGG ***(6)***.

Example of metabolic space expansion. 4-coumarate:CoA ligase (4CL), an enzyme involved in the phenylpropanoid biosynthesis, attaches 4-coumaric acid to the pantetheine group of Coenzyme-A (CoA) to produce 4-coumaroyl-CoA, the precursor of resveratrol. Besides this native reaction, however, 4CL is reportedly able to catalyze promiscuously,

among others, reactions producing caffeoyl-CoA from caffeate, feruloyl-CoA from ferulate, sinapoyl-CoA from sinapate, cinnamoyl-CoA from cinnamic acid ***(27)***. Interestingly, substrate cinnamic acid is transformed into cinnamoyl-CoA, which may also serve as a precursor for the production of resveratrol. As shown in **Fig. 3**, both reactions can be derived from a single reaction rule, since the net balance of bonds that are formed and broken is identical for both reactions.

3.4. Enumerating heterologous pathways

The metabolic space that has been defined in the previous step consists of both endogenous and heterogeneous reactions. In order to produce exogenous compounds, the corresponding metabolic routes containing heterologous enzymes that start from endogenous metabolites must be found (*see* **Note 4**). Two methods can be applied in order to list all possible pathways leading to the target compound ***(10)*** (*see* **Note 5**): steady state and topological methods.

Pathway enumeration through the steady-state approach. The problem of enumerating heterologous pathways can be approached by using the well-known metabolic engineering technique of computing elementary modes in a metabolic network ***(28)***. By definition, any pathway producing a target compound can be formed by some positive linear combination of the elementary modes. Here, we are focusing on a specialized version of elementary modes studies. Namely, we are interested in finding all pathways connecting endogenous metabolites to the target compound through heterologous enzymes. In particular, it is essential to correctly define what are the inputs, outputs, and the stoichiometric matrix of the metabolic system as follows:

1. Input: any metabolite that can be produced in the chassis organism;
2. Output: any metabolite that is produced and is not further consumed by the heterologous network;

3. Stoichiometric matrix: it is given by the reactions that are heterologous to the chassis organism.

Special care needs to be taken with those endogenous metabolites that are also produced in the heterologous network (usually co-factors and currency metabolites such as ATP, NAD and protons, which participate in a large number of reactions ***(29)***, but also by-products of the biosynthesis), since they will appear in the system defined above as both inputs and outputs, generating thus elementary modes containing loops. An easy way to prevent the enumeration of these return loops is by replacing endogenous products by a generic end node sink (*see* **Note 6**).

Under this set up, each elementary mode will correspond to a pathway that produces heterologous compounds. Because the number of pathways needs to be kept minimal, all pathways of interest producing a target compound should be contained in the elementary modes. In addition, elementary modes containing loops should not be considered as pathway candidates. Several software packages are available that compute the elementary modes of a given metabolic network. A popular implementation is Metatool ***(28)***.

Pathway enumeration through the topological approach. In the topological approach, each reaction is represented by a node of a hypergraph that is connected through hyperedges to the substrates (*see* **Note 7**). The strategy used in the hypergraph approach for pathway enumeration is the application of a recursive backward algorithm that traverses the network starting from the target in order to search for all possible pathways connecting the target to the source (*see* **Note 8**).

The main advantage of this approach is computational efficiency. Another remarkable feature of the topological approach is that it allows for supplements and bootstraps molecules identification. Supplements are compounds that provide new biosynthetic pathways if added to the medium because they act as precursors of the target compounds. Bootstraps are

compounds that are needed to be present in the medium at least in small amounts in order to allow the reactions in the pathway to start producing the target compound. A software tool for enumerating pathways using the topological approach is Metahype ***(10)***.

Example of pathway enumeration of resveratrol producing pathways in *E. coli*. Starting from precursors in *E. coli*, five alternative viable pathways producing resveratrol are identified by the retrosynthetic approach (shown in **Fig. 4**). One of the pathways consists of three enzymatic steps, while the rest contain four enzymes. The question that is investigated about these pathways in the next sections is how to prioritize them depending on their expected performance, as a preliminary step before selecting the ones that would eventually be implemented.

3.5. Gene compatibility for expression in the host

Heterologous enzymes of the pathway need to be successfully expressed. Gene compatibility with the expression host is crucial, although it remains still a challenging task. Facilities proposing gene synthesis with codon optimization ***(30)*** have been developed in the past few years and are often used for metabolic engineering. In order to select the gene, several strategies are possible:

1. Rare codons, and GC content are known parameters that can influence gene expression, as well as RNA secondary structure ***(30)***. In addition, other parameters such as sequence length or hydrophobicity might also influence a successful gene expression.
2. Homology search of heterologous genes: A blast search of the National Center of Biotechnology Information (NCBI) nucleotide data bank can identify sequences predicted to encode the enzyme having the desired activity. Phylogenetic trees can be built to identify groups of the different enzymes identified ***(31)***. Minimizing the phylogenetic

distance between the chassis organism and the organism where the gene is endogenous can also help in order to choose the homologue enzyme.

3. Scoring gene compatibility: An adequate strategy for gene sequence selection is to associate a score to each sequence, so that only sequences with top score are further considered. In a simple approach, the score can be built as a weighted sum of the considered factors, such as GC content or phylogenetic distance. Because of the multiplicity of factors that can influence enzyme expression, it might be difficult to blindly assign weighting priorities to each factor. One possible approach to address this issue is to build a statistical learning predictor based on techniques such as multilinear regression, support vector machines, or decision trees ***(32)***. The training set consists of the selected sequence properties with positive data formed by the list of enzyme sequences in the chassis, while the negative set has to be chosen as a significantly diverse selection of heterologous enzyme sequences (*see* **Note 9**).

Example of gene selection in resveratrol production. Besides its production from stilbene synthase (STS), production of resveratrol has also been observed as a cross-reaction from chalcone synthase (CHS, EC 2.3.1.74) ***(33)***. Interestingly, CHS is ubiquitous in plants and is also found in bacteria, while STS is only found in plant species that accumulate resveratrol and other related compounds ***(22)***. Selecting a prokaryotic CHS gene from an organism closer to *E. coli* could, thus, ease its successful expression. To that end, the retrosynthesis methodology can be applied in order to select best gene candidates expressing CHS enzyme, with the goal of converting it into an efficient resveratrol-producing factory. **Table 1** provides the score of CHS genes as candidates for promiscuously producing resveratrol in *E. coli*.

3.6. Estimating yield and drains

The next step in pathway design corresponds to metabolic analysis of the enumerated pathways in order to get an estimation of growth and yield of the target product associated with each metabolic intervention. This step is typically performed for the steady state through flux balance analysis (FBA) ***(3)***. To that end, an *in silico* model of the metabolic network of the chassis organism, normally given in SBML representation, needs to be obtained from the literature or from databases such as BIGG. Several software packages for FBA like the COBRA toolbox ***(11)*** are available. The following steps should be followed to perform flux balance analysis of the heterologous pathways:

1. Consistency check between the *in silico* SBML model and the metabolic database: Because they contain only reactions reconstructed from high-throughput *omics* data, genome-wide reconstructed *in silico* models do not contain such level of detail in the pathways as the one in metabolic databases like MetaCyc or KEGG. Therefore, a minimum level of consistency needs to be guaranteed between the metabolic network model from databases described in **Section 3.3** and the *in silico* SBML model for FBA, since it might happen otherwise that the heterologous pathway obtained from pathway enumeration in **Section 3.4** is fully or partially disconnected from the chassis in the *in silico* model. Therefore, it is necessary to verify that the endogenous precursors in the pathway are present in the *in silico* model, either because they are being produced by some enzymatic reaction or by adding the ones missing to the medium.

2. Inserting the heterologous pathway and their corresponding transport and exchange processes: Next, the heterologous pathway has to be imported into the model by adding as many reactions as enzymatic steps are present in the pathway. The target product is exported out of the cell through the use of the relevant transport reactions (*see* **Note 10**). In addition, any side product of the reaction steps, which is not being further degraded by

any reaction has to be exported out of the cell. *In silico* models generally make distinction between the exchange reaction between the extracellular medium and the periplasmic space, and the net exchange reaction of the metabolite with the system (*see* **Note 11**). For all reactions present in the model, constraints need to be defined by placing bounds in their reaction fluxes. For a nonreversible reaction, a bound between 0 and infinity might suffice, although more accurate constraints could be used based on empirical data. Similarly, bounds in exchange reactions have to be defined.

3. Setting the objective of the flux balance analysis: Constraint-based FBA is a technique that allows computing the optimal steady state fluxes in the *in silico* once bounds in fluxes and an overall objective function has been established. Generally, the objective function is defined by a linear combination of fluxes experimentally determined to correlate with biomass growth. In the case of organisms engineered through genetic modification to produce a target compound in order to estimate the effect of the pathway insertion several goals can be established: a) comparison of the optimal growth before and after pathway insertion; b) computation of maximum yield (generally leading to zero growth) by setting the goal to produce the desired compound; c) computation of the optimal biomass-product coupled yield goal **(34)**. These optimal values provide an overall overview of what can be achieved by the modified strain as a cell factory of the desired compound.

Example of yield estimation of resveratrol in *E. coli*. Using the COBRA toolbox, fluxes maximizing resveratrol and biomass yields were computed for the 5 alternative pathways in **Fig. 4**, as shown in **Table 2**. Optimal flux for biomass in wild type strain is 0.737 (a.u.). A similar value is obtained in the strain that has been engineered with Pathway 1, while the maximum yield for resveratrol is 1.951. The strain with Pathway 2 shows an increase in biomass (1.111), indicating that some of the by-products can be used to increase growth. Maximum production of resveratrol (3.589) is significantly increased in this pathway.

Pathway 3 and 4 correspond to biomass and resveratrol yields that are in between the maximum values obtained by Pathway 1 and 2. Finally, Pathway 5 is the one that yields the maximum production of resveratrol (3.951).

3.7. Toxicity effects of heterologous pathways

In metabolic engineering, the importance of compound toxicity has been pointed it out by several authors ***(35,36)***. Indeed, for pathway performance the less toxic molecule is usually desired, while conversely the highly toxic molecule is wanted when producing therapeutics such as antimicrobials. In both cases, detailed information on compound toxicity is important in the design of metabolic pathways and toxicity is a parameter that should be included in the computer-aided pathway design framework ***(15)***. In order to establish a reference database of toxicity data in the chassis organism, a library of MIC (minimal inhibitory concentration) or IC_{50} (half maximal inhibitory concentration) experimental values should be built ***(14)***. These experimental values can then be used to develop a quantitative structure-activity relationship (QSAR) ***(37)*** model for toxicity of chemicals towards the chassis organism. The process consists of the following steps:

1. Firstly, a library should be designed in a way to provide a representative set of chemicals with maximal chemical diversity. The selection of compounds can be done using a method based on the optimal clustering of the chemical space determined by the distances defined as the chemical dissimilarity between compounds (*see* **Note 12**). A significant region of the chemical space has to be covered in order to maximize the spectrum of toxicity values.
2. The bacteria used for the toxicity assay must have been identified at the genus and species level. Standardization for accuracy of results and reproducibility is crucial, and as an

example one important parameter to control is the inoculum size (usually 5.10^5 colony-forming units (cfu) ml^{-1} for broth dilution).

3. A fresh pure culture of *E. coli* strain (such as *E. coli* ATCC 25922 usually chosen for toxicity assay) is used for the inoculum. Bacteria are grown in liquid medium at 37°C, and bacterial growth determined at the stationary phase after incubation for a defined period (for example 18 hours). Toxicity assay can be performed in 96-well microtiter plates. The chemicals are screened by serial dilution to assess their toxicity towards *E. coli*. Controls are important to add for each experiment. A positive control (as a triplicate) consists to bacterial culture without any chemical, and a negative control (as a triplicate) to monitor the absence of contamination consists to the media only.

4. Bacterial cell growth is monitored by measuring the turbidity (at 600 nm) at the stationary phase and then data analyses are carried out to determine MIC or IC_{50}. Dose-response curves are built for each compound and fit using the sigmoidal equation:

$$y = \frac{OD_{\max}}{1+\left(\frac{C}{IC_{50}}\right)^{p}} \quad (1)$$

where the ODmax represents the maximal OD_{600nm}, C is the concentration of the compound, IC_{50} is the molecule concentration that inhibits 50% of the bacterial growth, and *p* the Hill slope describing the steepness of the curve. Only IC_{50} extracted from the curve fitting having a coefficient of determination $R^2>0.9$ should be used in the training set. The Levenberg-Marquardt least squares fitting algorithm for the sigmoidal curve can be used.

5. In order to predict toxicity values of intermediates, a quantitative structure-activity relationship (QSAR) model of toxicity has to be developed from the experimental dataset by using a statistical software package like the pls library in R ***(38)***. Descriptors for the compounds in the dataset might be chosen in the same fashion as the ones selected for

performing the clustering of chemicals. For regularization purposes, IC_{50} values should be transformed into $\log(IC_{50})$. Two basic statistical methods can then be applied in order to build the QSAR model:

a) Principal component analysis (PCA) in order to reduce the dimension of the signature vectors, by keeping only components whose variance is above some given cutoff ratio of the total variance in the set.
b) Model fitting by the partial least squares (PLS) regression method ***(38)***. PLS decomposes the principal components of the molecular signature descriptors into several latent variables that correlate best with the toxicity values $\log(IC_{50})$.

6. Validation of the QSAR is typically accomplished through two steps:

a) Internal validation through the leave-one-out method;
b) External validation by using a list of experimental values that have not been used before for training and validation.

Example of toxicity estimation of resveratrol intermediates in *E. coli*. Table 3 lists predicted toxicity values for the intermediate heterologous metabolites involved in the production of resveratrol, as computed by the QSAR model of the EcoliTox web server for prediction of toxicity in *E. coli* ***(14)***. Typically, inhibition values for *E. coli* endogenous metabolites are found between IC_{50} = 0.1 g/1 and 50 g/l. Predicted values for the resveratrol intermediates were also found within this range. Based on these estimations, high inhibition effects might not be expected due to the insertion of the pathways. Cinnamoyl-CoA, the heterologous intermediate of Pathway 4 and Pathway 5 in **Fig. 4**, is the most toxic compound in the list.

3.8. Defining a cost function for the pathways

As presented in previous sections, several aspects are to be considered when estimating the cost of pathway insertion. A quantitative definition of a cost function associated with the pathway provides the possibility of ranking enumerated pathways, so that top pathways can be selected by the designer for implementation. The process is as follows:

1. A simplified scheme consists on dividing the effects of pathway insertion into three main factors: enzyme compatibility and reaction efficiency (**Section 3.3** and **Section 3.5**), expected yield (**Section 3.6**), and metabolite toxicity (**Section 3.7**). A possible definition of the pathway cost function is as follows ***(15)***:

$$W(c,\rho) = -\lambda_{flux} v_c(\rho) + \lambda_{path} \sum_{r\in\rho}^{N} K(S(r)) + \lambda_{tox} \sum_{r\in\rho}^{N} \sum_{p\in r} T(p)$$
$$\lambda_{flux}, \lambda_{path}, \lambda_{tox} \geq 0 \qquad (2)$$
$$\lambda_{flux} + \lambda_{path} + \lambda_{tox} = 1$$

where $\nu_c(r)$ is the flux for pathway r producing compound *c*, as described in **Section 3.6**, $K(S(r))$ is the cost associated with sequence *S* of the enzyme catalyzing the reaction *r* in the pathway r, and $T(p)$ is the toxicity ($-\log_{10}(IC_{50})$) associated with the metabolite *p* product of reaction *r*, as defined in **Section 3.7**.

2. The cost for the sequence $K(S(r))$ has to take into account the fact of whether reaction is found annotated in databases for the given sequence or it is found based on a prediction as described in **Section 3.3**. In the case of a putative reaction, a penalty is added to the cost as follows:

$$K(S(r)) = \Gamma_{pred}(r) + \Delta(S) \qquad (3)$$

where $\Delta(S)$ is the compatibility for sequence *S* as defined in **Section 3.5**, and $\Gamma_{pred}(r)$ is defined in order to assign an additional cost to those enzyme sequences *S* where the

corresponding reaction *r* is either catalyzed as a promiscuous or side reaction or its assignment is only putative. Therefore:

$$\Gamma_{pred}(r) = \begin{cases} \Gamma_{penalty} & \text{promiscuous/predicted} \\ 0 & \text{annotated} \end{cases} \quad (4)$$

with penalty constant $G_{penalty}$ arbitrarily set to a value that is a upper bound for the score of sequence compatibility: $\Delta(S) \leq \Gamma_{penalty}$.

3. The choice of values for parameters (l_{flux}, l_{path}, l_{tox}, $G_{penalty}$) depends on each experimental set up as well as on the preferences set up by expert designers. A first approach is to set the values so that the cost function assigns less cost to that pathways that contain only enzymes annotated in databases (no putative enzymes). In ***(15)***, parameters were fitted in this way to (0.025,1.0,0.398,5.0).

Example of ranking resveratrol pathways. For the resveratrol pathways, gene costs have been computed from the RetroPath server ***(15)*** and their values are shown in **Table 4**. The total cost associated with each pathway, according to **Equation 2** and to the weighting parameters is shown in **Table 5**. From these results, Pathway 2 appears finally as the best candidate pathway to engineer in *E. coli* for the production of resveratrol, a result that is due both to the fact that the pathway contains less putative enzymatic steps and to the higher expected yield. Pathway 1 appears as the second ranked pathway because even if it involves only three enzymes in comparison with the four enzymes from the other pathways, it contains two predicted or putative reactions (cinnamic acid production from PAL and again STS). Pathway 3, which is similar to Pathway 2 except for the first step (2-enoate reductase), predicted as a promiscuous reaction, appears next in the ranking. Finally, Pathways 5 and 4, containing three out of four predicted reactions (promiscuous activity predicted for 4CL and CH4, and STS), appear at the end of the ranking.

3.9. Pathway implementation

Validation of the retrosynthetic design is performed through pathway implementation of the top ranked heterologous pathways. The process consists typically of the following steps (**Fig. 1**):

1. Gene amplification: Genes can be amplified from genomic DNA, when available, or synthesized. The ability to synthesize genes in whole novel genetic pathways is now routinely used for metabolic engineering. Software and websites to facilitate the execution of oligonucleotide assembly into long custom sequences are available ***(39)***. Chemical synthesis allows synthesizing oligonucleotides of up to 120–150 nucleotides in length. Numerous methods have been developed to assemble relatively short synthetic oligonucleotides into longer gene sequences through ligation or PCR-mediated assembly. Also, codon usage varies by organism and has implications for heterologous expression of proteins. Codon optimization, which consists to render a nucleotide sequence with suitable codon usage for the expression host, might also help the gene expression but will not necessarily maximize the protein expression level.
2. Expression strain: Common expression strains are obtained from resources such as *E. coli* genetic Stock Center (New Haven, CT) or companies as Life TechnologiesTM (Paisley, UK) and New England Biolabs (Ipswich, MA, USA). *E. coli* strains specially developed for gene expression are chosen according to the expression system needed: tightly-controlled expression or expression level modulation, and depending on the type of the promoter used.
3. Promoter selection: Popular promoters are the lac-promoter, allowing gene expression modulation, and the promoters pBAD of the arabinose operon and pRHA of the rhamnose operon that offer a tight control of gene expression. Tight expression control prior to target protein induction can be crucial for expression of host-toxic proteins to

avoid deleterious events ending in mutations that may affect target protein function, or cell death.

4. Plasmid construction: Constructing a multiple enzyme biosynthetic pathway implies to combine several genes into a single plasmid or to use compatible expression plasmids. Commercially available plasmids as pETDuet-1, pACYCDuet-1 and pCDFDuet-1 (Novagen, ***(40)***) are widely used in metabolic engineering. Other systems allowing the cloning of multiple genes into one single plasmid such as pQlink vectors are also available ***(41)*** from repository sources such as Addgene (Cambridge, MA, USA). Gene assembly methods are also an alternative to combine multiple genes into a plasmid that have their expression under the control of their own promoter ***(42,43)***. Nowadays, a wide selection of expression vectors is available, which differ in their origins of replication, promoters, translation initiation regions, antibiotic resistance markers and transcription terminators.
5. Bacterial culture: Bacterial culture is commonly carried out at 30°C or 37°C in rich medium, although optimization might be needed. It might be necessary, however, to address problems related to protein misfolding and solubility. For example, to limit protein aggregation the temperature can be decreased to 25°C, and the use of minimal medium can be more favorable for metabolite production ***(44)***. Optimization of growth temperature and induction conditions, chaperone-coexpression system, and fusions to solubilizing partners are among numerous solutions to increase product yields.
6. Verification of protein expression: Preparation of total cell protein samples is followed by the separation of protein samples by SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) and eventually western-blot if antibodies recognizing the protein target are available. For high-throughput screening for protein expression, fluorescent partners can be used. Fluorescent proteins can be used to

monitor the expression level of soluble or membrane-embedded proteins ***(45)*** and coupled with flow cytometry allow large fluorescence-based library screening.

7. Identification of the target compound: Once the protein expression has been successful, the target compound must be identified using analytical techniques. Chemical production can be determined using the intrinsic spectroscopic properties of the chemical. Metabolites contain a high chemical diversity and the two main analytical methods that can provide structural data are the nuclear magnetic resonance (NMR) and mass spectrometry (MS) with different ion sources and mass detectors ***(46)***. To detect the metabolite of interest, MS is a robust technique when coupled with chromatography. Gas chromatography (GC-MS) has the main advantage of providing high separation efficiency. However, a major drawback of GC-MS is that the compound must be volatile. Liquid-chromatography coupled to mass spectrometry (LC-MS) represents an attractive alternative to GC-MS because of its versatile separation technique (hydrophilic interaction liquid chromatography (HILIC) MS, reverse phase LC-MS) ***(46,47)***. LC/MS has emerged as a popular and powerful tool. Following this step, preparative methods such as HPLC coupled with spectrometry are usually used to quantify the metabolite production. For example, metabolite can be specifically separated on a C18 column with a determined acetonitrile/water gradient. A large number of purification methods exist and need to be optimized for each compound.

4. Notes

1. The *in silico* model should contain at least a reconstructed stoichiometric network of the organism substantially covering their main metabolic routes. Transcription regulation,

thermodynamics, kinetics as well as other information is increasingly becoming available in these models and will bring in the future the design to finer levels of detail.

2. There is a basic difference between the information that is required in the model in order to design heterologous metabolic pathways and to estimate steady-state fluxes. In the former case, the most essential information is the knowledge about the metabolites that are endogenous to the organism and therefore can be used as precursors in the heterologous pathway. In the latter case, the accuracy of the stoichiometric relationship between those reactions that directly influence the pathway is required, while partial knowledge about upstream reactions with low influence into the pathway can be tolerated.

3. In bond-electron matrices (BEM), each row and column correspond to one atom of the compound, and each entry is the order of the covalent bond between the atoms. The BEM of a reaction is defined as the difference between the end (right) and begin (left) BEMs.

4. The reason why we need to enumerate all pathways instead of searching for the shortest one is because not always the shortest is the best in terms of the cost associated to the pathway, as described in **Section 3.8**.

5. We are only interested in enumerating minimal hyperpaths, loosely meaning cycle-free hyperpaths (see ***(10)*** for proper definitions).

6. The basic limitation of the elementary modes approach is its computational complexity, which can make slow the computation in case of large heterologous networks.

7. The hypergraph representation is used in order to require all substrates to be present for the reaction to be activated.

8. The hypergraph approach can lead to solutions that are not stoichiometrically balanced, since stoichiometry is not taken into account. These solutions need to be filtered out from the output of the algorithm.

9. A detailed description of such type of predictor can be found in ***(15)*** as well as in the patented method from DNA 2.0 ***(30)***.

10. As *in silico* models become more detailed, higher attention needs to be paid in order to describe accurately the process inside the cell. Cell compartmentalization, for instance, might imply the need for enzyme co-localization in order for the pathway to proceed. This aspect is especially relevant for plant metabolism. Similarly, exporting the metabolite out of the cell might involve several transport processes through different cell compartments ***(11)***.

11. The flux balance might require for metabolites to be taken outside of the extracellular medium in order to make the net flux zero, avoiding accumulation.

12. Several clustering methods can be applied, depending first on the type of molecular descriptors used to define molecular similarity and on the clustering algorithm. Hierarchical agglomerative clustering should be preferred, since it allows building libraries of variable size.

Acknowledgements

This work was funded by Genopole® (ATIGE grant) and Agence Nationale de la Recherche (ANR Chaire d'excellence).

References

1. Oberhardt M.A., Palsson B.O., Papin J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol Sys Biol* **5**, 320.
2. Yadav V.G., De Mey M., Lim C.G. et al. (2012) The future of metabolic engineering and synthetic biology: Towards a systematic practice. *Metab Eng* **14**, 233-241.

3. Curran K.A., Crook N.C., Alper H.S. (2012) Using flux balance analysis to guide microbial metabolic engineering. *Methods Mol Biol* **834**, 197-216.
4. Planson A.G., Carbonell P., Grigoras I. et al. (2012) A retrosynthetic biology approach to therapeutics: from conception to delivery. *Curr Opin Biotechnol.* doi:10.1016/j.copbio.2012.03.009
5. Caspi R., Altman T., Dreher K. et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **40**, D742-D753.
6. Kanehisa M., Goto S., Sato Y. et al. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-D114.
7. Chang A., Scheer M., Grote A. et al. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* **37**, D588-92.
8. Schellenberger J., Park J., Conrad T. et al. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213.
9. Li C., Donizelli M., Rodriguez N. et al. (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* **4**, 92.
10. Carbonell P., Fichera D., Pandit S.B. et al. (2012) Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst Biol* **6**, 10.
11. Schellenberger J., Que R., Fleming R.M.T. et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290-1307.

12. Rocha I., Maia P., Evangelista P. et al. (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst Biol* **4**, 45.
13. Hoops S., Sahle S., Gauges R. et al. (2006) COPASI - a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067-3074.
14. Planson A.G., Carbonell P., Paillard E. et al. (2012) Compound toxicity screening and structure-activity relationship modeling in *Escherichia coli*. *Biotechnol Bioeng* **109**, 846-850.
15. Carbonell P., Planson A.G., Fichera D. et al. (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst Biol* **5**, 122.
16. Feist A.M., Herrgard M.J., Thiele I. et al. (2008) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* **7**, 129-143.
17. Menzella H., Reeves C. (2007) Combinatorial biosynthesis for drug development. *Curr Opin Microbiol* **10**, 238-245.
18. Ajikumar P.K., Xiao W.H., Tyo K.E.J. et al. (2010) Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. *Science* **330**, 70-74.
19. Orth J.D., Conrad T.M., Na J. et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Mol Syst Biol* **7**, 535.
20. Krivoruchko A., Siewers V., Nielsen J. (2011) Opportunities for yeast metabolic engineering: Lessons from synthetic biology. *Biotechnol J* **6**, 262-276.
21. Boghigian B.A., Seth G., Kiss R. et al. (2010) Metabolic flux analysis and pharmaceutical production. *Metab Eng* **12**, 81-95.
22. Halls C., Yu O. (2008) Potential for metabolic engineering of resveratrol biosynthesis. *Trends Biotechnol* **26**, 77-81.
23. Beekwilder J., Wolswinkel R., Jonker H. et al. (2006) Production of Resveratrol in Recombinant Microorganisms. *Appl Environ Microbiol* **72**, 5670-5672.

24. Lim C.G.G., Fowler Z.L., Hueller T. et al. (2011) High-yield resveratrol production in engineered *Escherichia coli*. *Appl Environ Microbiol* **77**, 3451-3460.
25. Brunk E., Neri M., Tavernelli I. et al. (2012) Integrating computational methods to retrofit enzymes to synthetic pathways. *Biotechnol Bioeng* **109**, 572-582.
26. Cho A., Yun H., Park J.H.H. et al. (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst Biol* **4**, 35.
27. Limem I., Guedon E., Hehn A. et al. (2008) Production of phenylpropanoid compounds by recombinant microorganisms expressing plant-specific biosynthesis genes. *Process Biochem* **43**, 463-479.
28. Kamp A., Schuster S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* **22**, 1930-1931.
29. Gerlee P., Lizana L., Sneppen K. (2009) Pathway identification by network pruning in the metabolic network of Escherichia coli. *Bioinformatics* **25**, 3282-3288.
30. Welch M., Villalobos A., Gustafsson C. et al. (2011) Designing genes for successful protein expression. *Methods Enzymol* **498**, 43-66.
31. Tamura K., Peterson D., Peterson N. et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28**, 2731-2739.
32. Kingsford C., Salzberg S.L. (2008) What are decision trees? *Nat Biotechnol* **26**, 1011-1013.
33. Yamaguchi T., Kurosaki F., Suh D. et al. (1999) Cross-reaction of chalcone synthase and stilbene synthase overexpressed in *Escherichia coli*. *FEBS Lett* **460**, 457-461.
34. Patil K., Rocha I., Forster J. et al. (2005) Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6**, 308.

35. Ma S.M., Garcia D.E., Redding-Johanson A.M. et al. (2011) Optimization of a heterologous mevalonate pathway through the use of variant HMG-CoA reductases. *Metab Eng* **13**, 588-597.

36. Pitera D.J., Paddon C.J., Newman J.D. et al. (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metab Eng* **9**, 193-207.

37. Tropsha A., Golbraikh A. (2010) Predictive quantitative structure-activity relationship modeling. Development and validation of QSAR models. In: Faulon J.L and Benders A. (eds) Handbook of Chemoinformatics Algorithms, Chapman and Hall/CRC, pp. 211-232.

38. Mevik B.H., Wehrens R. (2007) The pls package: principal component and partial least squares regression in R. *J Stat Softw* **18**, 1.

39. Hughes R.A., Miklos A.E., Ellington A.D. (2011) Gene synthesis: methods and applications. *Methods Enzymol* **498**, 277-309.

40. Tolia N.H. Joshua-Tor L. (2006) Strategies for protein coexpression in *Escherichia coli*. *Nat Methods* **3**, 55-64.

41. Scheich C., Kummel D., Soumailakakis D. et al. (2007) Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res* **35**, e43.

42. Tsvetanova B., Peng L., Liang X. et al. (2011) Genetic assembly tools for synthetic biology. *Methods Enzymol* **498**, 327-348.

43. Gibson D.G. (2011) Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* **498**, 349-361.

44. Santos C.N.S., Koffas M., Stephanopoulos G. (2011) Optimization of a heterologous pathway for the production of flavonoids from glucose. *Metab Eng* **13**, 392-400.

45. Makino T., Skretas G., Georgiou G. (2011) Strain engineering for improved expression of recombinant proteins in bacteria. *Microb Cell Fact* **10**, 32.

46. Roux A., Lison D., Junot C. et al. (2011) Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clin Biochem* **44**, 119-135.

47. Garcia D.E., Baidoo E.E., Benke P.I. et al. (2008) Separation and mass spectrometry in microbial metabolomics. *Curr Opin Microbiol* **11**, 233-239.

Figure Captions

Fig. 1. Flowchart of the retrosynthetic methodology for heterologous pathway design. The process consists of nine steps: it starts by choosing the chassis organism and the target compound, followed by a selection of an *in silico* model of the chassis, and the definition of the metabolic space. The process continues with pathway enumeration, for each reaction in the pathway, genes are selected based on their compatibility to the host and a penalty is added if the reaction appears as a putative promiscuous reaction. Next, steady state yields are estimated, as well as toxicity of reaction products. These values are combined in order to score and select the best pathway(s), which are finally implemented for verification.

Fig. 2. Outline of phenylpropanoid alternative pathways producing resveratrol. In the center, the backbone of the native pathway is shown, consisting of phenyalanine ammonia lyase (PAL), cinnamate-4 hydroxylaxe (C4H), 4-coumarate-CoA ligase (4CL) and stilbene synthase (STS), the last step. 4-coumarate can be produced by an alternative pathway with the tyrosine ammonia lyase (TAL), a promiscuous reaction of PAL. Other predicted alternative routes are 2-enoate reductase (2ER) producing cinnamic acid from phenylpropanoic acid, and 4CL and C4H using alternative substrates cinnamic acid and cinnamoyl-CoA, respectively.

Fig. 3. Example of two promiscuous reactions generated by the same reaction rule defined by the 4CL enzyme (EC 6.2.1.12). In the first case, substrate 4-coumaric acid is converted into 4-coumaroyl-CoA, while in the second case cinnamic is converted into cinnamoyl-CoA. Both, substrates and products, differ in the 4-hydroxy group, but the net balance of bonds that are formed or broken is identical for both reactions.

Fig. 4. Five alternative resveratrol biosynthetic pathways computed by the Metahype server. 1) Pathway 1 contains three enzymes leading to the target. The L-phenylalanine/tyrosine ammonia lyase (PAL/TAL, EC 4.3.1.25) produces the precursor 4-coumarate from L-

tyrosine, the 4CL (EC 6.2.1.12) produces 4-coumaroyl-CoA that is converted into resveratrol by STS (EC 2.3.1.95). 2) In Pathway 2, which contains 4 enzymes, PAL produces cinnamic acid from L-phenylalanine. Cinnamic acid is then converted by 4CH (EC 1.14.13.11) into the precursor 4-coumarate, which is further processed into resveratrol as in Pathway 1. 3) Pathway 3 differs from Pathway 2 only in the first step producing cinnamic acid , which in this case in accomplished by a promiscuous reaction of found in 2-enoate reductase (EC 1.3.1.31) that produces cinnamic acid from phenylpropanoic acid. 4) In Pathway 4, cinnamic acid produced as in Pathway 2 from PAL is used to synthetize cinnamoyl-CoA, which is further transformed into 4-coumaroyl-CoA through promiscuous reactions from 4CL and C4H, as discussed in Section 3.2.1. 5) Pathway 5 uses 2-enoate reductase to produce the precursor cinnamic acid, which is processed downstream through cinnamoyl-CoA into resveratrol in the same way as in Pathway 4.

Table Captions

Table 1. Sequence features and organism compatibility for top CHS genes according to RetroPath for gene insertion in *E. coli*. Input features consisted of sequence length, GC content, probability to be expressed as inclusion bodies, isoelectric point (pI), hydrophobicity, secondary structure distribution, and distance to prokaryotes.

Table 2. Optimal steady state fluxes for the five pathways in **Fig. 3** maximizing biomass (first column) or the production of resveratrol (second column).

Table 3. Predicted toxicity in *E. coli* of metabolite intermediates of the resveratrol pathways in **Fig. 3**.

Table 4. Gene costs $K(S(r))$ associated with each gene in the pathway in **Fig. 3**.

Table 5. Cost of each of the five resveratrol pathways according to the cost function in **Equation 2**.

gene_id	Cost	L	GC	ibody	pI	hyd	helix	sheet	turns	coil	d	Organism
Bind_3897	0.99	406	63.79	0.54	6.4	-86.9	41.5	22.6	17.2	22.8	0	*Beijerinckia indica*
Bind_2602	1.02	354	68.46	0.6	45.4	10.2	32.0	28.1	16.0	28.7	0	*Beijerinckia indica*
Ping_0256	1.29	362	63.4	40.52	6.6	-10.4	56.9	15.9	13.9	17.9	0	*Psychromonas ingrahamii*
Gbem_1028	1.47	349	68.86	0.66	6.5	54.5	42.0	30.0	15.6	17.1	0	*Geobacter bemidjiensis*
Psyc_0421	1.66	362	63.44	0.56	6.3	-5.1	63.3	14.5	12.7	14.2	0	*Psychrobacter arcticum*
Mrad2831 _4712	1.94	357	69.47	0.58	7.5	19.5	46.9	24.9	13.8	19.1	0	*Methylobacterium radiotolerans*
Msil_3391	2.27	360	70.46	0.57	6.5	78.6	44.8	27.9	12.8	19.2	0	*Methylocella silvestris*
sce2182	2.87	371	71.16	0.63	6.8	95.2	46.2	25.1	11.0	22.3	0	*Sornagium cellulosum*
RB8853	3.27	367	69.6	60.5	34.6	38.8	31.1	29.9	20.2	23.4	1	*Rhodopirellula baltica*

Table 1

Strain	Biomass (a.u.)	Resveratrol (a.u)
WT	0.737	-
Path 1	0.737	1.591
Path 2	1.111	3.589
Path 3	0.798	1.907
Path 4	1.110	1.959
Path 5	0.830	3.951

Table 2

Pathway	Compound	Predicted toxicity (IC_{50})
1,2,3	4-Coumarate	0.69 g/l
4,5	4-Coumaroyl-CoA	0.25 g/l
1,2,3,4,5	Resveratrol	0.42 g/l
2,3,4,5	Cinnamic acid	0.18 g/l
4,5	Cinnamoyl-CoA	0.16 g/l

Table 3

Pathway	EC	Gene	Organism	Substrate	Product	Penalty	Cost
1	4.3.1.25	RSP_3574	*Rhodobacter sphaeroides*	tyrosine	4-coumarate	5.0	5.20
1,2,3	6.2.1.12	RPA4421	*Rhodopseudomonas palustris CGA009*	4-coumarate	4-coumaroyl-CoA	0.0	0.99
2,3	1.14.13.11	4338409	*Oryza sativa japonica*	cinnamic acid	4-coumarate	0.0	4.77
2,4	4.3.1.25	4336415	*Oryza sativa japonica*	phenylalanine	cinnamic acid	0.0	4.64
3,5	1.3.1.31	CKL_1689	*Clostridium kluyveri DSM 55*	phenylpropanoate	cinnamic acid	5.0	3.84
4,5	6.2.1.12	RPA4421	*Rhodopseudomonas palustris CGA009*	cinnamic acid	cinnamoyl-CoA	5.0	0.99
4,5	1.14.13.11	4336415	*Oryza sativa japonica*	cinnamoyl-CoA	4-coumaroyl-CoA	5.0	4.65
1,2, 3,4,5	2.3.1.95	Bind_3897	*Beijerinckia indica*	4-coumaroyl-CoA	resveratrol	5.0	1.00

Table 4

	Sequence compatibility	Expected yield	Metabolites toxicity	Total cost
Pathway 1	7.19	1.640	1.110	17.603
Pathway 2	11.40	2.350	1.290	16.858
Pathway 3	10.60	1.353	1.290	21.080
Pathway 4	11.28	1.535	1.010	26.644
Pathway 5	10.48	2.391	1.010	25.822

Table 5