

# A retrosynthetic biology approach to metabolic pathway design for therapeutic production

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

▶ **To cite this version:**

Pablo Carbonell, Anne-Gaëlle Planson, Davide Fichera, Jean-Loup Faulon. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Systems Biology, 2011, 5 (122), 10.1186/1752-0509-5-122 . hal-02652758

**HAL Id: hal-02652758**
**https://hal.inrae.fr/hal-02652758**

Submitted on 29 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Carbonell et al. BMC Systems Biology 2011, 5:122
http://www.biomedcentral.com/1752-0509/5/122

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

# A retrosynthetic biology approach to metabolic pathway design for therapeutic production

Pablo Carbonell, Anne-Gaëlle Planson, Davide Fichera and Jean-Loup Faulon*

## Abstract

**Background:** Synthetic biology is used to develop cell factories for production of chemicals by constructively importing heterologous pathways into industrial microorganisms. In this work we present a retrosynthetic approach to the production of therapeutics with the goal of developing an *in situ* drug delivery device in host cells. Retrosynthesis, a concept originally proposed for synthetic chemistry, iteratively applies reversed chemical transformations (reversed enzyme-catalyzed reactions in the metabolic space) starting from a target product to reach precursors that are endogenous to the chassis. So far, a wider adoption of retrosynthesis into the manufacturing pipeline has been hindered by the complexity of enumerating all feasible biosynthetic pathways for a given compound.

**Results:** In our method, we efficiently address the complexity problem by coding substrates, products and reactions into molecular signatures. Metabolic maps are represented using hypergraphs and the complexity is controlled by varying the specificity of the molecular signature. Furthermore, our method enables candidate pathways to be ranked to determine which ones are best to engineer. The proposed ranking function can integrate data from different sources such as host compatibility for inserted genes, the estimation of steady-state fluxes from the genome-wide reconstruction of the organism's metabolism, or the estimation of metabolite toxicity from experimental assays. We use several machine-learning tools in order to estimate enzyme activity and reaction efficiency at each step of the identified pathways. Examples of production in bacteria and yeast for two antibiotics and for one antitumor agent, as well as for several essential metabolites are outlined.

**Conclusions:** We present here a unified framework that integrates diverse techniques involved in the design of heterologous biosynthetic pathways through a retrosynthetic approach in the reaction signature space. Our engineering methodology enables the flexible design of industrial microorganisms for the efficient on-demand production of chemical compounds with therapeutic applications.

## Background

Synthetic biology is being used for therapeutic production either to develop cell factories using industrial microorganisms [1,2] or to synthesize genetic circuits allowing *in situ* therapeutic delivery [3]. Recombinant DNA technology has already provided the ability to genetically engineer cell strains in order to import pathways from other organisms capable of producing small molecule chemicals into microbial chassis. Moreover, to estimate the efficiency of the overall process, metabolic engineering-based tools consider models of cell metabolism as a whole, allowing the identification and redesign of bottlenecks in the biosynthetic pathways. Therefore, the next challenge ahead remains the integration of all these design steps into a flexible and automated biosynthetic manufacturing pipeline of molecules.

In recent years, many successful examples of bioproduction of chemicals with therapeutic interest through metabolic engineering have been reported. Among others, plant secondary metabolites that are of medicinal value, such as the terpenoids artemisinic acid [4] and paclitaxel (taxol) [5], benzylisoquinoline alkaloids [6], and flavonoids [7,8] have been successfully produced by metabolically engineered microorganisms. Similarly, heterologous production of therapeutically important antibiotics such as aminoglycosides derivatives, which include ribostamycin [9], neomycin, gentamicin and

* Correspondence: jfaulon@gmail.com
iSSB, Institute of Systems and Synthetic Biology, University of Evry, Genopole Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 EVRY Cedex, France

BioMed Central

© 2011 Carbonell et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

kanamycin, as well as other natural products like polyketides [10,11] and nonribosomal peptides [12] have been reported. Flexible production of novel antibiotics is of special interest in order to fight against the increasing emergence of multidrug-resistant pathogens [13-15].

In an attempt to rationalize the biosynthetic design process, metabolic engineering models the metabolic network of the cell as a whole [16,17]. A suitable topological representation of the metabolic network can be achieved by using directed hypergraphs [18,19] where catalytic reactions are hyperedges connecting node substrates to products. Moreover, genome-wide reconstructions of an organism's metabolism with explicit reference to the stoichiometry of the reactions can be studied in order to estimate steady-state fluxes [20]. Sensitivity analysis of fluxes provides a systematic way to determine production bottlenecks, where gene overexpression or repression might enhance production for the target compound [21,22]. In addition, stochastic and deterministic system dynamics methods are used to simulate enzymatic reaction kinetics [23].

Through metabolic modeling, the repertoire of biochemical transformations in *de novo* biosynthetic pathways are extended beyond what is present in metabolic databases like KEGG [24] and MetaCyc [25]. *In silico* methods for *de novo* pathway prediction and optimization are mainly based on two approaches: homologies of chemical structure transformation patterns [26-29], and knowledge-based expert systems [30,31]. Retrosynthesis algorithms [32] perform a backward search for biosynthetic routes leading from a target compound to the host metabolites through iterative application of a defined set of biochemical transformation rules. One approach is BNICE [33], where molecules and reactions are represented by bond-electron matrices (BEM) [34]. BEM entries correspond to the covalent bond order between atoms, whereas the Dugundji-Ugi model for a metabolic reaction is implemented through the matrix difference between the BEM of products and substrates. With BNICE, reactions in the KEGG database [24] are represented through approximately 250 unique elementary transformations, which approximately correspond to the classification at the 3rd EC digits [35,36]. In the same fashion, the molecular signature descriptor [37] is an algorithm that returns for a given target compound fourth and third EC digits, respectively, of predicted enzymes capable of producing the structure. Similarly, other retrosynthetic framework has been proposed based on 50 reaction rules [38].

A retrosynthetic search in the metabolic hypergraph might lead to a combinatorial explosion. For instance, using only 50 reaction rules, 100,000 reaction routes were predicted for the production of isobutanol [38], far more than what could be realistically tested in the laboratory. Thus, in order to find a trade-off between the inherent complexity of *de novo* pathway design and the use of experimental information, we present here a tool based on the coding of compounds and reactions through molecular signatures [39]. The molecular signature is a canonical representation of the subgraph surrounding a particular atom in a molecular structure up to a predefined diameter or height *h*. A metabolic reaction signature is given by the difference between the signatures of products and substrates [40]. As further described in the Methods section, the signature coding system can be made more or less specific to compounds and reactions by selecting the height, low heights are less specific (as molecular signatures become more and more ambiguous) and high heights are specific (as molecular signatures become more and more precise), thus the numbers of *de novo* reactions and consequently *de novo* pathways can be controlled.

Once metabolic models for the heterologous biosynthesis of target compounds have been determined, individual performances for the predicted pathways need to be characterized in order to prioritize the engineering of the most promising routes into the chassis organism. Several computational frameworks have proposed different factors that might influence the performance of an engineered strain. PathMiner introduced a path cost associated with the number of heterologous enzymes measured through a chemical distance [41]. BNICE applied the group contribution method [42] for reactants and products in order to rank pathways based on the thermodynamic favorability [43]. Other aspects influencing the pathway performance are pathway length, organism specificity [38], heterologous expression, growth rate, and precursor supply [44]. In addition, many other factors might be considered, for instance, PathoLogic defined 123 pathway features that may be relevant to the pathway ranking problem [31]. Therefore, subsequent optimization of the heterologous engineered strain through genetic, metabolic and enzyme design approaches would be usually necessary in order to attain the desired final yields in the production of the target compound. Moreover, increasing efficiency levels for rate-limiting enzymatic reactions involved in the pathway is an additional strategy for the rational design of industrial strains [37,45,46]. We present here a unified framework that combines several techniques involved in the design of heterologous biosynthetic pathways through a retrosynthetic approach in the reaction signature space, enabling the flexible design of industrial microorganisms for the efficient on-demand production of chemical compounds of interest.

## Results and Discussion

### The extended metabolic reaction space (EMRS)

Our method starts by mapping the metabolic network into the signature space in order to build an extended representation of the metabolic reaction space (see definitions in Methods). Molecular signatures, which are a representation of the molecular graph, can be used in order to code reactions [46]. This process is illustrated in Figure 1. Canonical molecular signatures identify unique compounds when they are computed at the height $h$ of the maximum diameter of the molecular graph, whereas signatures at lower height $h$ provide a way to search for and enumerate similar chemical compounds. Likewise, reaction signatures, which are biochemical reactions coded into the molecular signature representation (Equation 7 in Methods), are used in order to search for and enumerate similar reactions. We define the extended metabolic reaction space (EMRS) as the set of all possible reactions that can be generated from signatures contained in the metabolic network. Therefore, the EMRS consists of both reactions in the metabolic network and additional putative reactions, which are assumed to be promiscuously catalyzed by enzymes present in the organism. Given a finite height $h$, novel reactions are discovered through this method by performing a search in the metabolite signature space of combinations of stoichiometric coefficients of metabolites having the same signature as either the substrates or the products. Figure 2 shows the metabolic reaction map of the 966 endogenous metabolites in *E. coli* (Figure 2A) and of the additional 2338 compounds that are reachable from *E. coli* after the generation of the EMRS (Figure 2B).

An illustrative example is the metabolite signature space of height $h = 0$. In this space, compounds are represented by their elemental formula. Similarly, the combination of the substrates and products in the reaction are represented by their total molecular formula. Thus, any combination of compounds satisfying the elemental formula is considered a putative set of reactants. In order to compute the EMRS of height $h = 0$, we need to solve Diophantine equations in the signature space of height $h = 0$ that generally lead to a set of solutions too large to be of use. In the case of $h = 1$, the number of newly created reactions is still significantly large (for a network like the one shown in Figure 2, it would be above $10^6$). This number, nevertheless, becomes tractable once we consider heights higher than $h = 2$, which corresponds to a 17.72% increase in the number of reactions with respect to nominal reactions, as it is shown in Table 1. Starting from the list of coded reactions, the iterative backward application of the biochemical transformations to compounds of interest allows the identification of enzymatic routes linking the desired compound to precursors that are endogenous to the chassis organism. Each of these routes constitutes an exogenous biosynthetic pathway for that compound. In the next sections, we present an approach for ranking the biosynthetic pathways of a given compound in order to select the best pathways to engineer in the chassis organism.

### Decision flowchart for selecting and ranking best pathways

The EMRS introduces putative novel reactions that share the same signature as their parent nominal reactions at the chosen height $h$ of molecular resolution (Equation 8 in Methods). Those putative reactions generated by our molecular signature-based algorithm need to be validated and ranked. Figure 3 shows the decision flowchart used in our approach in order to accept or reject putative reactions in a pathway as well as to score its overall performance once inserted into the chassis organism. Reactions are first tested for their thermodynamic feasibility. Next, if no known enzyme sequences are available in the database, the enzyme sequence space is searched in order to find candidate sequences that might catalyze the given reaction as a promiscuous activity. Gene compatibility, enzymatic performance, toxicity of products and steady state fluxes are finally estimated in order to score the pathway.

We introduce the following function to quantify the cost of inserting an exogenous enzyme sequence $S_i$ processing the reaction $r^*$ in the pathway:

$$\begin{gathered}\mathcal{W}(r^*, S_i) = \\ = 1 - \omega_p \ \mathtt{promis}(S_i) + 1 - \omega_e \ \mathtt{perf}(r^*, S_i) + \mathtt{het}(S_i) \\ 0 \le \mathtt{promis}(S_i), \mathtt{perf}(r^*, S_i), \mathtt{het}(S_i) \le 1 \\ 0 \le \omega_p, \omega_e \le 1\end{gathered} \tag{1}$$

where $\mathtt{promis}(S_i)$ is the predicted enzyme promiscuity for the sequence $S_i$, $\mathtt{perf}(r^*, S_i)$ is the estimated catalytic performance of the given sequence $S_i$ for reaction $r^*$, and $\mathtt{het}(S_i)$ is the gene compatibility of the sequence $S_i$. $\omega_p$, $\omega_e$ are parameters used in order to weight the contribution of each term to the cost function. All three terms are normalized before entering the expression so that each score is always given by a value in the same range between 0 and 1. Therefore, the cost function $\mathcal{W}(r^*, S_i)$in Equation 1 is always defined positive and bounded. Promiscuity contributes negatively to the cost function because enzymes with higher level of promiscuity are considered better candidates for catalyzing the desired transformation $r^*$ as a side reaction. In the same fashion, enzyme performance contributes negatively since reactions with higher performance are considered

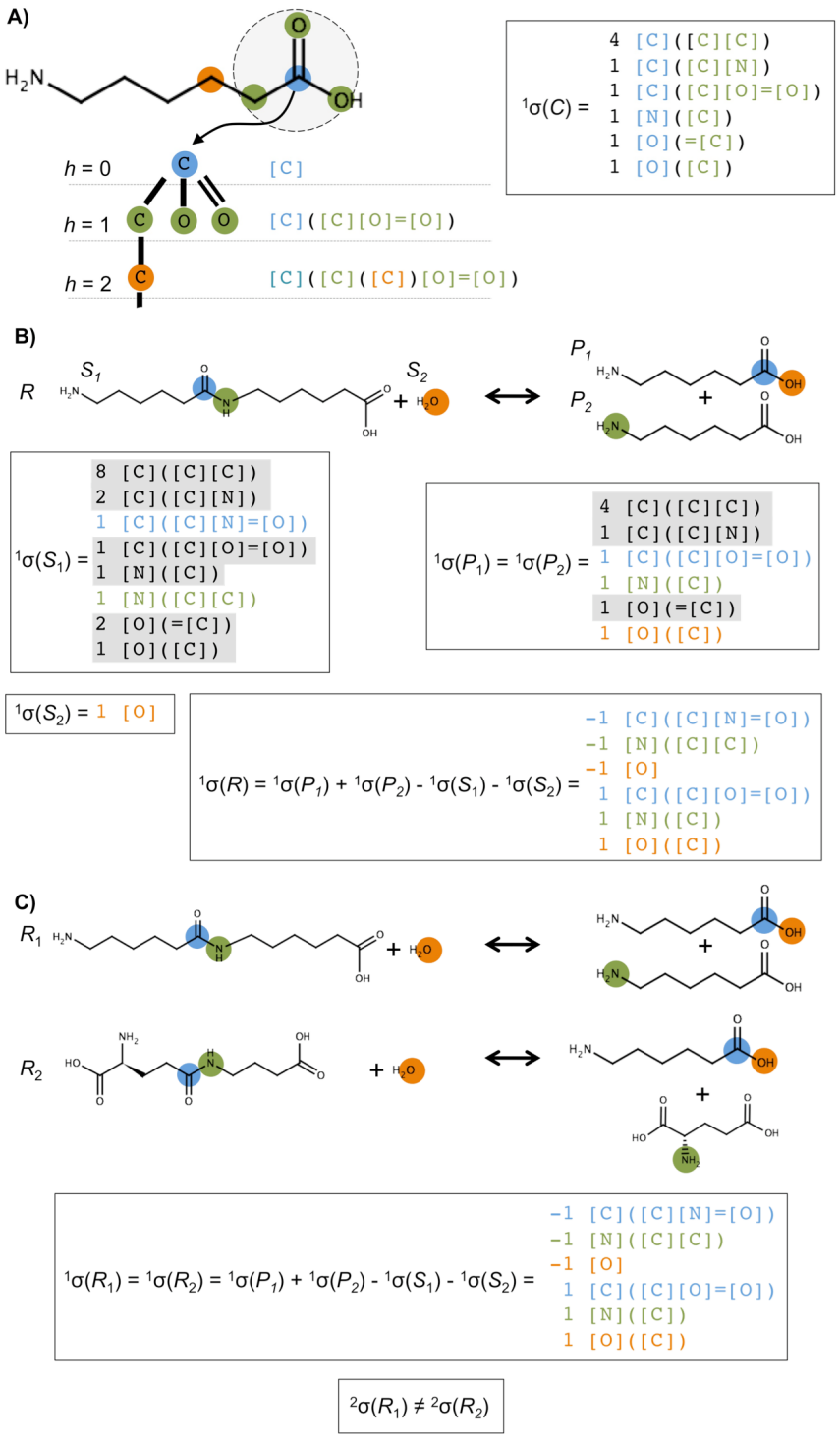



**Figure 1 The atomic, molecular and reaction signature coding**. A) The process for computing the molecular signature for a compound *C* is illustrated for 6-aminohexanate. The process starts by computing the atomic signature for each atom. In the given example, the atomic signature for the carbon in the carboxylic group is computed up to height $h = 2$. At height $h = 0$ (blue), the molecular graph rooted at the atom is given by the atom itself; at height $h = 1$ (green) a canonical representation of the root atom and its first atomic neighbors are given; the process continues similarly for heights $h = 2$ (orange) and higher until the diameter of the graph is reached. Atomic signatures are collected for all atoms and sorted in order to provide the molecular signature, for instance the molecular signature ${}^1\boldsymbol{\sigma}(C)$ of height $h = 1$ is given at the left; B) The coding of reactions signatures is illustrated for the 6-aminohexanoate hydrolase (EC 3.5.1.46). The reaction signature contains the net difference between the products and the substrates. In the figure, the reaction signature ${}^1\boldsymbol{\sigma}(R)$ was computed for height $h = 1$; C) Illustration of how signatures of reactions provide a way to measure their chemical similarity. For example, the previous reaction (EC 3.5.1.46) has the same signature at height $h = 1$ than 4-(γ-glutamylamino)butanoate amidohydrolase (EC 3.5.1.94). However, both signatures differ at height $h = 2$, having in this case a Tanimoto similarity of ${}^2s(R_1, R_2) = 0.81$ (see Equation 14 in Methods).

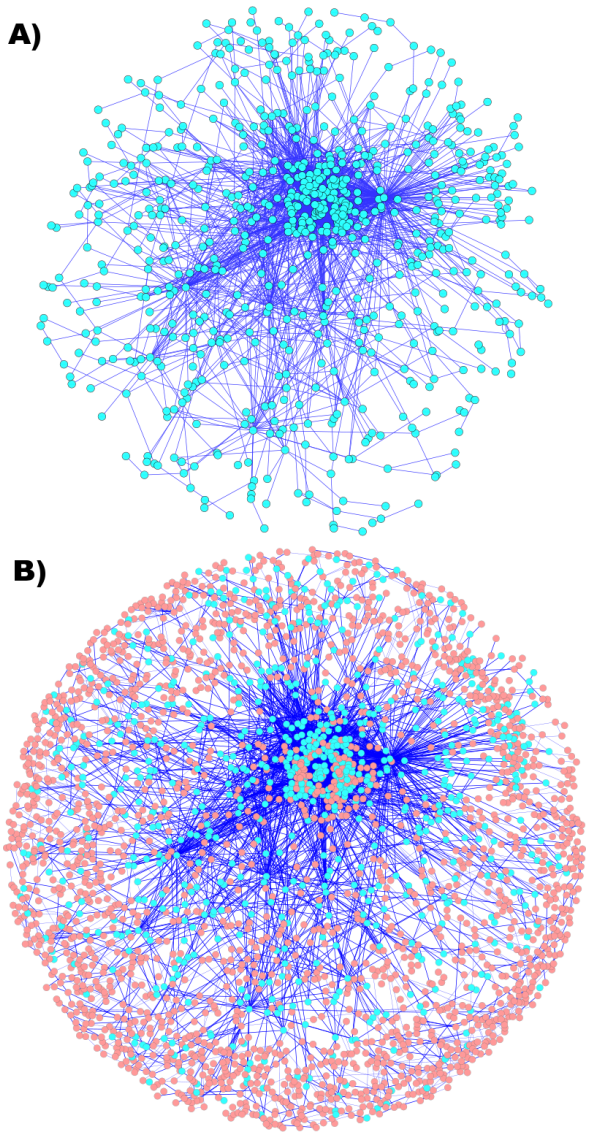



**Figure 2 Metabolic networks in the EMRS**. A) Metabolic reaction map of *E. coli*, where endogenous metabolites are depicted as light blue nodes connected by edges representing reactions; B) Retrosynthetic map containing reachable compounds in the *E. coli* EMRS through exogenous reactions, where exogenous metabolites are represented by pink nodes connected through reactions (thin edges) to the *E. coli* network. There are 966 endogenous and 2338 exogenous compounds, respectively, that can be reached through reactions in the EMRS. There are 4,344 edges connecting endogenous compounds and 8,931 edges leading to exogenous compounds.

less expensive in terms of the cost of insertion. Once all terms are defined for each reaction $r^*$ in the EMRS, this strategy enables the determination of those gene sequences $S^*$ that minimize the insertion cost:

$$S^*(r^*) = \arg\min_{S_i} \mathcal{W}(r^*, S_i) \tag{2}$$

**Table 1 Reactions in the EMRS**

| height *h* | reactions | % increase from canonical |
|---|---|---|
| 2 | 9083 | 17.72% |
| 3 | 7882 | 2.15% |
| 4 | 7800 | 1.09% |
| 5 | 7752 | 0.47% |
| 6 | 7725 | 0.12% |
| canonical | 7716 | 0% |

Number of novel generated putative reactions in the EMRS for different heights *h*.

Furthermore, several additional adverse effects may be hindering the successful expression and performance of inserted enzymes, while the toxicity of intermediate metabolites might impede cell survival and growth. These effects need to be taken into account in order to rank the pathways. For instance, an estimate of toxicity values (IC50 or half minimal inhibitory concentration) for the intermediates $p$ in the chassis organism, may be obtained either from experimental databases [47] or from structure-activity relationship models [48]. In addition, compound yields from inserted pathways are rarely additive, since other routes may be competing with the target pathway and inhibiting the production of the desired compound [5]. Here we use a multi-criteria approach in order to score the cost of pathway insertion with respect to the general goal of producing a target molecule $c$, with a cost function defined as follows:

$$W(c,\rho) = -\lambda_{\mathtt{flux}} v_c(\rho) + \\ + \lambda_{\mathtt{path}} \sum_{r^* \in \rho}^{N} \mathcal{W}(r^*, S^*(r^*)) + \lambda_{\mathtt{tox}} \sum_{r^* \in \rho}^{N} \sum_{p \in r^*} \mathtt{tox}(p) \tag{3}$$

where $W(c, \rho)$ considers the following effects: $v_c(\rho)$, nominal yield of compound $c$ in pathway $\rho$; $\mathcal{W}(r^*, S^*(r^*))$, minimum cost of insertion of each enzyme in the pathway as given by Equation 1; and tox $(p)$, the inverse of the IC50 value. Parameters ($\lambda_{\mathtt{flux}}$, $\lambda_{\mathtt{path}}$, $\lambda_{\mathtt{tox}}$) need to be adjusted in function of the desired weight given to the costs of pathway insertion and metabolite toxicity. In our method, these parameters were optimized so that pathways that are fully annotated in the reference database, for instance KEGG, are ranked first with respect to predicted pathways (see details in Methods).

The minimum of this cost function $W(c, \rho^*(c))$ at the optimal pathway:

$$\rho^*(c) = \arg_{\rho} \min W(c,\rho) \tag{4}$$

provides a trade-off between the simultaneous goals of obtaining the maximum nominal yield while keeping the

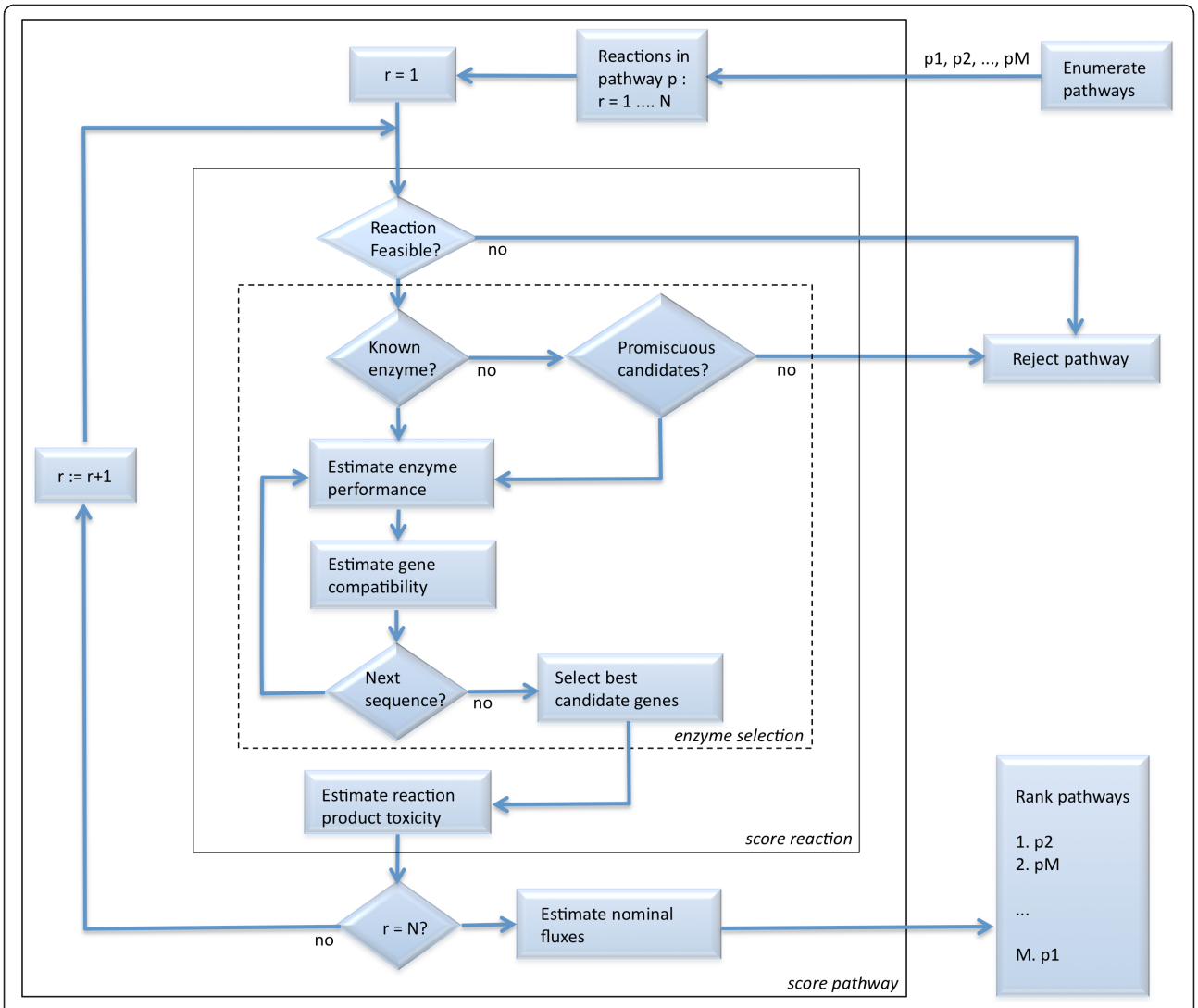



**Figure 3 Flowchart for ranking pathways**. In order to rank pathways, each reaction $r = 1 \ldots N$ in the enumerated pathways $p_1 \ldots p_M$ is first tested for thermodynamical feasibility; enzyme candidates are subsequently tested for performance and homogeneity so that the one with the lowest cost is selected; the cost of toxicity of each reaction product is then added; finally, the nominal flux is estimated for the overall pathway.

overall process efficient and side effects attenuated. In the following sections, we present our approach in order to quantify each term in the pathway cost function $W(c, \rho)$ in Equation 3.

### Predicting enzyme activity in promiscuous putative reactions

Our method provides for each reaction in the EMRS a ranked list of candidate sequences, as given by the score in Equation 1, along with their predicted catalytic efficiencies. When there is no sequence information in databases about enzymes catalyzing the desired reaction $r^*$, we must rely on the prediction of enzymes as putative candidates to process other substrates (multispecificity) or to catalyze a promiscuous reaction other than their native ones [49]. Furthermore the thermodynamic feasibility of that reactions as well as the performance of the predicted promiscuous enzymes need to be evaluated. These preliminary evaluations, which are described next, are carried out in order to implement an early rejection of false hits, as shown in the flowchart of Figure 3.

#### *Thermodynamic feasibility*

Putative reactions need to be validated for their directionality or thermodynamic feasibility. We performed this validation assuming that the metabolites' concentration are spatially invariant and that temperature and pressure are constant. Under these assumptions, standard Gibbs free energies of reactions can be estimated using a group contribution approach [43,50]. Only

reactions estimated to be thermodynamically feasible are added to the EMRS.

#### *Enzyme promiscuity*

As part of our methodology for biosynthetic pathway design, candidate enzyme sequences catalyzing feasible reactions in the EMRS have to be identified from the set of known enzymes. Namely, our procedure for extending the metabolic network relied on the underlying assumption that reactions with the same signature are likely to be processed by similar enzyme sequences [37], which in turn implies that the ability to catalyze the putative reaction is already present in the enzyme in the form of latent promiscuous activity. We have shown in a previous study [46] that enzyme multispecificity and promiscuity are properties that can be characterized by using a molecular signature representation. Thus, those multiple reactions generated in the EMRS from a nominal parent reaction can be interpreted as promiscuous activities predicted to be present in the set of known enzymes. Therefore, as shown in the decision flowchart in Figure 3, in order to consider a given sequence as a potential candidate that processes the putative reactions, a preliminary requirement has been introduced in the decision chart so that the enzyme should exhibit promiscuous activity based on the estimations performed by a molecular signature-based predictor (see Methods).

#### *Tensor product*

The next step consists of searching for candidate enzymes to process a given reaction in the EMRS. In case that no known enzyme sequences were available for the reaction, candidate enzymes were determined by a signature-based enzyme-reaction predictor by following the procedure known as the tensor product [51]. We assumed that best candidate enzyme sequences for a putative reaction were more likely to belong to the list of sequences known to catalyze reactions that are chemically similar to the given reaction.

Therefore, reactions generated by the enumeration algorithm in the EMRS were first clustered into groups of similar reactions by a distance metrics, which was defined as the Tanimoto similarity of reaction signatures [52]. The tensor product procedure (see Methods) was then used in order to locate best enzyme sequence candidates within the reaction cluster.

#### *Enzyme performance*

In addition, performance of exogenous enzymes needs to be evaluated. We have developed a decision tree learning method to estimate enzyme activity [53] using kinetics information from the BRENDA database [54] (turnover rates, Michaelis constant $K_M$ , and inhibition constant $K_i$). As shown in Figure 3, predictions of enzyme performance perf($r^*$, $S_i$) for the list of candidate enzymes entered into our decision flowchart in order to score the sequences in Equation 1.

### Quantifying the compatibility between the host and heterologous genes

Another aspect to be addressed when considering the overall enzyme cost defined in Equation 1, is the effect of inserting heterologous genes, since the diversity of base-pair content is organism-specific. By minimizing this difference, expression levels can be maximized [55]. In order to quantify the compatibility between the host and heterologous genes, we have implemented a machine-learning approach based on several descriptors: gene sequence descriptors (sequence length, GC content); organism specificity (phylogenetic distance between source organism and chassis); probability of protein expression as inclusion bodies; protein descriptors (percentage of hydrophobic and charged amino acids); and secondary structure distribution. These descriptors were computed for the entire KEGG database of non-redundant enzyme sequences and then used in order to train support vector machine-based predictors for the chassis organisms of interest (see Methods).

Furthermore, the successful expression of a heterologous gene depends on several additional sequence-independent factors, such as an adequate selection of promoters, RBS, and codons [56]. In some cases, we should also consider the need for some other type of specific modifications depending strictly on the type of compound to be synthesized, such as protein engineering of P450s [57] or modular design for the complex assembly machinery involved in the production of secondary metabolites like polyketides [58] and nonribosomal peptides [59,60], which need to be evaluated on a case-by-case basis in order to rank and select the best genes to engineer.

### Predicting compound toxicity

Exogenous enzymes inserted in the chassis might catalyze reactions synthesizing new products in the organism. As a side effect, however, intermediate metabolites involved in the exogenous pathways as well as any other side product of the new reactions may induce undesired toxic responses in the cell. Therefore, it is necessary to consider toxicity effects of the compounds. For this purpose, we have developed a structure-activity relationship model based on a library of 150 tested compounds covering a wide range of toxicity levels [61]. The model was built by using several molecular descriptors including molecular signatures as input descriptors, achieving a performance of $Q^2$ = 0.68. For any given reaction in the EMRS, toxicity was given by the sum of the predicted toxicity for each product, allowing us to identify pathways involving highly toxic metabolites in order to rank them with lower score.

Special consideration when predicting compound toxicity should be given to those cases when genes encoding

for resistance to the compound are going to be engineered as part of the biosynthetic gene cluster. For instance, when producing an antibiotic in *E. coli* such as penicillin, it is necessary to introduce genes that code for $\beta$-lactam resistance in the organism in order to make bacteria immune to that antibiotic. Therefore, if resistance to some product is going to be inserted into the strain, the attenuation effects in toxicity for that family of compounds has to be updated into the model.

### Estimation of nominal fluxes

The insertion of new reactions into the chassis organism can perturb its metabolic network and therefore the equilibrium of the steady-state fluxes might be altered [20]. By using a constraints-based flux analysis on a genome-wide reconstructed metabolic model of the engineered strain, we obtain the solution space within cell's capacity. Our objective is to maximize the production of the desired compound while keeping cell growth. For each engineered strain, we obtained an estimate of expected net yield of product, which is not further metabolized, at the given controlled media conditions. In addition, flux balance analysis is a flexible analytical technique that can also be applied in other ways in our design framework for biosynthetic pathways. For instance, it can be used in a systematic way in order to perform a sensitivity analysis to determine production bottlenecks, where overexpression and gene knockouts might enhance production for the target compound [21,22].

### Pathway enumeration and optimal search in the EMRS

Given a biosynthetic pathway $\rho(c)$ that produces a compound $c$, we have shown in the previous sections how to estimate the individual contributions to the cost function (Equation 3). By using the cost function, thus, biosynthetic pathways $\rho(c)$ can be ranked. However, in order to rank all viable biosynthetic pathways $\rho(c)$ for a compound $c$ of interest, the problem of pathway enumeration needs to be addressed. For this purpose, modeling of the metabolic network in the EMRS was done by using directed hypergraphs, where reactions are represented by hyperedges that connect sets of vertices (the substrates) to disjoint sets of vertices (the products) [62]. Directed hypergraph formalism, though more complex than simple-graph models, provides a complete representation of all compounds involved in biochemical transformations. By using the hypergraph formalism, we implemented a retrosynthetic algorithm that enumerates all pathways starting from target compounds of interest. One main advantage of the pathway enumeration in the EMRS is that complexity can be controlled by tuning the atomic height $h$ of the molecular signatures. Higher values of $h$ imply that the number of pathways between two metabolites is approximately the same as the number of possible pathways in metabolic networks of annotated databases such as KEGG [24] or MetaCyc [25], while lower $h$ values generate more novel reactions and therefore more possible pathways are formed between those two metabolites. This result is illustrated for the case of pathway enumeration between chorismate and tyrosine in *E. coli* for different heights $h$ using a reaction representation at the level of the 3rd digit in the EC number classification (Figure 4). In general, the possible number of pathways that can be formed between these two metabolites increases exponentially with the number of reaction steps. When values of the molecular signature height are high ($h \geq 6$), new reactions are unlikely to be generated and therefore the number of pathways becomes the same as the number of pathways available in KEGG (the reference metabolic database); whereas as the height $h$ decreases, the number of new reactions and, thus, the number of pathways starts growing while getting closer to those results that were obtained in BNICE by using BEM matrices [33].

In a general setting, the problem of finding the optimal pathway $\rho^*(c)$ in Equation 4 that produces the target compound $c$ is equivalent to finding the shortest hyperpath in a weighted hypergraph. This problem is known to be an NP-hard problem [62], although it can be reduced to a polynomially solvable problem if the cost function in Equation 3 is reformulated as an additive objective function [62]. In the EMRS approach,

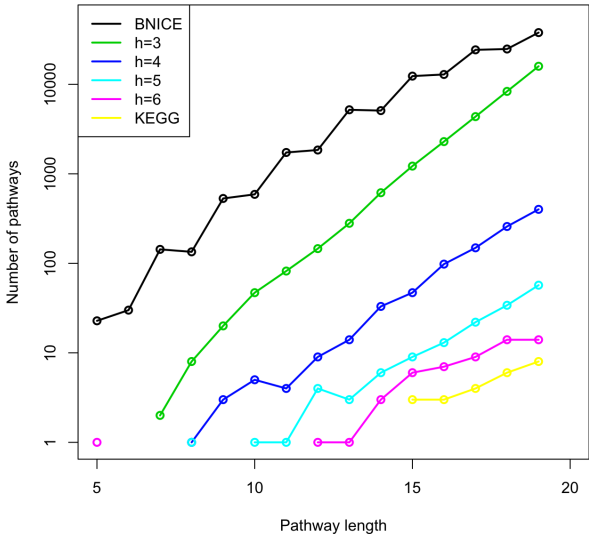



**Figure 4 Controlling the complexity of the pathway enumeration problem through molecular signatures.** Comparison between pathway length distributions between tyrosine and chorismate for novel reactions generated by the BEM representation (BNICE) [33], the EMRS of heights $h$ = 3 to 6, and the original reactions in the KEGG metabolic database.

complexity can be controlled by varying the specificity of the molecular signature. This flexibility allows us to follow the strategy of enumerating all biosynthetic pathways $\rho(c)$ and computing their associated costs $W(c, \rho)$ in Equation 3.

### Validation test for auxotrophic production in *E. coli*

A validation test for the ranking score was carried out by testing its ability to identify native biosynthetic pathways for several essential metabolites (the 20 amino acids, citrate, ATP, ADP, GTP and GDP) in auxotrophic strains of *E. coli*. These strains were rendered unable to synthesize those essential compounds by inactivation of the enzymes that natively produce them. We assumed that native pathways, which have been selectively conserved under evolutionary pressure, are efficient ways to produce the compounds while keeping cell growth. The validation, thus, consisted of checking if the pathways that were ranked at the top of the list correspond to native pathways. In order to find all possible ways to synthesize the amino acids, we ran the retrosynthetic search in these auxotrophic strains. The output results of the search, which are given in Table 2, provided both native and alternative pathways connecting the auxotroph to the compounds. The results of this test showed that in 98% of the cases native pathways were found within the top 10 ranked pathways for each amino acid.

An additional validation test was performed in order to evaluate the accuracy of the gene compatibility predictor. The result of this test, summarized in Table 2, showed that genes from the full list in the database that were predicted to be the best candidates to be inserted in the auxotroph strains corresponded significantly ($p$-value <0.05) to native genes of *E. coli*. These results are significant since the sequences under test were not part of the training set used for building the gene compatibility predictor. In summary, this test showed that the proposed ranking function can potentially identify heterologous biosynthetic pathways to insert in an organism to produce a desired compound while selecting the ones that are close to native pathways in the chassis.

**Table 2 Native pathway identification in auxotrophic *E. coli***

| Compound | Total | Native pathways | % of natives in best 10 | *p*-value |
|---|---|---|---|---|
| Alanine | 19 | 7 | 60.00% | 8.59e-08 |
| Arginine | 1 | 1 | 100.00% | 4.65e-02 |
| Asparagine | 3 | 3 | 100.00% | 8.42e-05 |
| Aspartic acid | 7 | 7 | 100.00% | 6.91e-09 |
| Cysteine | 11 | 5 | 100.00% | 1.32e-07 |
| Glutamic acid | 74 | 53 | 100.00% | 6.41e-52 |
| Glutamine | 7 | 5 | 100.00% | 1.41e-05 |
| Glycine | 31 | 16 | 80.00% | 3.16e-08 |
| Histidine | 4 | 4 | 100.00% | 4.85e-04 |
| Isoleucine | 1 | 1 | 100.00% | 4.55e-02 |
| Leucine | 1 | 1 | 100.00% | 4.54e-02 |
| Lysine | 1 | 1 | 100.00% | 6.38e-02 |
| Methionine | 107 | 106 | 100.00% | 1.08e-259 |
| Phenylalanine | 4 | 1 | 100.00% | 6.38e-02 |
| Proline | 1 | 1 | 100.00% | 2.17e-02 |
| Serine | 2 | 2 | 100.00% | 1.11e-03 |
| Threonine | 1 | 1 | 100.00% | 4.35e-02 |
| Tryptophan | 2 | 2 | 100.00% | 1.89e-03 |
| Tyrosine | 2 | 1 | 100.00% | 1.92e-02 |
| Valine | 1 | 1 | 100.00% | 4.55e-02 |
| Citrate | 4 | 3 | 100.00% | 5.71e-05 |
| ATP | 12 | 6 | 100.00% | 6.73e-06 |
| GTP | 2 | 2 | 100.00% | 2.08-03 |
| ADP | 316 | 204 | 100.00% | 2.89e-169 |
| GDP | 283 | 171 | 100.00% | <1.0e-324 |

Biosynthetic pathways for the 20 amino acids, citrate and ATP/ADP and GTP/GDP enumerated for *E. coli* strains where enzymes producing the compounds have been deleted. For each test, columns correspond to: total number of enumerated pathways; number of native pathways in the wild-type strain; percentage of native pathways in the top 10 best ranked pathways; and $p$-values for the number of genes in *E. coli* strains top ranked with respect to the total enzyme genes in the database.

### The RetroPath webserver

As shown in previous sections, the procedure of pathway selection by retrosynthesis is a complex task that implies the adoption of several design decisions, some of them on a case-by-case basis. In order to help on the decision-making process, we have developed an online tool: the RetroPath webserver that guides the designer through the retrosynthesis process. The design starts by choosing the target compound, which can be given as an SDF molecular file. Additionally, the user decides the level of molecular resolution to use in the molecular signature representation. For instance, we have analyzed the set of molecular structures in DrugBank [63] as initial target compounds. For this set, we found that more than 50% of reactions producing these compounds belong to the *E. coli* EMRS of height $h$ = 6. Furthermore, the distribution of the number of alternative biosynthetic pathways in the compound set follows a power law, as it is shown in Figure 5, which means that in some cases there might be thousands of alternative pathways that have to be ranked according to the ranking function in Equation 3 to search for the optimal pathway.

In order to illustrate the design process, we next present three examples of heterologous pathway design using the RetroPath webserver, the production of two $\beta$-

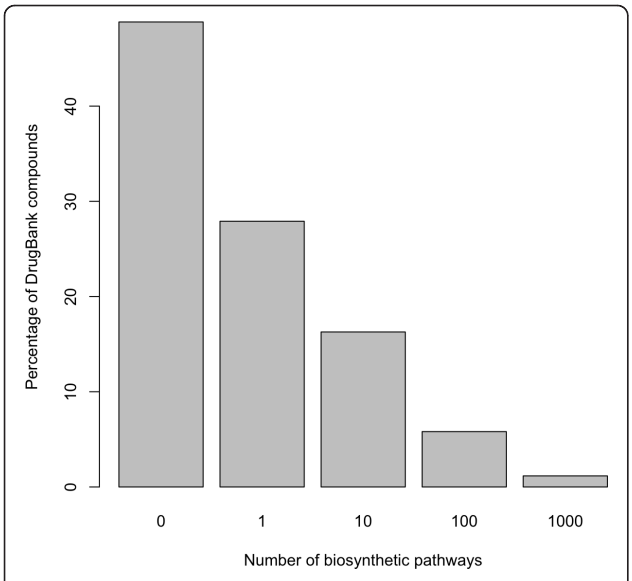



**Figure 5 Retrosynthetic pathways for production of DrugBank compounds in *E. coli***. Percentage of compounds in the DrugBank database with alternative biosynthetic pathways in *E. coli*.

lactams antibiotics in *E. coli* (penicillin G and cephalosporin), and of one antitumor drug (taxol) in yeast.

## Design examples of retrosynthetic pathways

### *Production of β-lactams in E. coli*

The industrial production of penicillin G occurs via fermentation using the filamentous fungus *Penicillium chrysogenum*. A recent study has opened up the possibility of producing penicillin G in an organism that is used as a producer of pharmaceuticals; the yeast *Hansenula polymorpha* [64]. Interestingly, the biosynthetic pathways of penicillin G are shared by another $\beta$-lactam antibiotic, cephalosporin, which is produced in the fungus *Acremonium chrysogenum* and synthetised from isopenicillin N, the penultimate precursor for penicillin production [65].

Using the retrosynthetic method that we have developed, retrosynthetic graphs were generated for $\beta$-lactam antibiotics, in particular penicillin and cephalosporin (Figures 6A,B). The chosen chassis organism was *E. coli*. Four different pathways were found at a signature reaction height of $h$ = 4 for penicillin N production and in particular one involving the nonribosomal peptide synthetase $\delta$-(L-$\alpha$-aminoadipyl)-L-cysteinyl-D-valine synthetase (EC 6.3.2.26) and isopenicillin N synthetase (EC 1.21.3.1). These pathways are the same as those that were implemented in the aforementioned studies in yeast and fungi to produce the isopenicillin N. In the cephalosporin biosynthesis pathway the isopenicillin N is converted into penicillin N, itself transformed into deacetoxycephalosporin. The retrosynthetic maps of height $h$ = 4 for heterologous production of penicillin N and deacetoxicephalosporin in *E. coli* are shown in Figure 6. In the retrosynthetic graph, the enzyme deacetoxycephalosporin-C synthase (EC 1.14.20.1) is the one responsible of the deacetoxycephalosporin formation. Table 3 ranks the 6 pathways in the map leading to penicillin N according to the cost function in Equation 3. Toxicity values for intermediates were predicted by using our model built from an experimental library of toxicity values in *E. coli* while fluxes were estimated from a reconstructed metabolic model of *E. coli*, as described in the Methods section. The optimal pathway involves five exogenous enzymatic steps, while the alternative pathways involves up to seven steps. In Table 3, the alternative routes for the production of penicillin N are generated by the synthesis step of the precursor L-2-aminoadipate-6-semialdehyde, where the retrosynthetic search identified several enzymatic routes that can be connected to precursors in *E. coli*.

### *Production of taxol (paclitaxel) in yeast*

Taxol (paclitaxel) is an anticancer drug first isolated from the Pacific yew tree *Taxus brevifolia*. Today, taxol derives largely by semisynthesis from the advanced taxoid 10-deacetylbaccatin III obtained from the European yew tree *Taxus baccata* [66]. Currently its production has a limiting rate as it depends on plant cell processes as well as chemical and biotechnological semisynthesis processes. For the past few years, a number of studies have been contributing to the elucidation of the biosynthetic mechanism of taxol and efforts have been made in order to attain cost-effective production through heterologous biosynthesis of taxol and its analogues [5]. The retrosynthetic graph for yeast (*Saccharomyces cerevisiae*) which was computed for a signature height $h$ = 4, (Figure 7), goes from the isopentenyl to taxol and contains 2 different pathways with 8 and 9 steps, respectively, that share most of the intermediates and only differ at two steps:

1. From isopentenyl (IPP) to taxadien-5$\alpha$-ol: The isopentenyl, native to yeast, undergoes 3 reaction steps to form the taxadien-5$\alpha$-ol. Those are catalyzed by the geranylgeranyl-diphosphate (GGPP) synthase (EC 2.5.1.29) that forms the GGPP (C00353) [67], by the taxadiene synthase (EC 4.2.3.17) that forms taxa-4(5),11(12)diene (C11894) [68] and finally by the taxadiene 5$\alpha$ hydrolase (EC 1.14.99.37) that forms the taxadien-5$\alpha$-ol [69]. The first two reactions have been reportedly implemented in *E. coli* [5,70] and, furthermore, were subject to engineering in order to optimize the taxadiene pathway production [5].
2. From the taxadien-5-$\alpha$-ol to taxol: From the taxa-4(20),11(12)-dien-5$\alpha$-ol two pathways are possible,

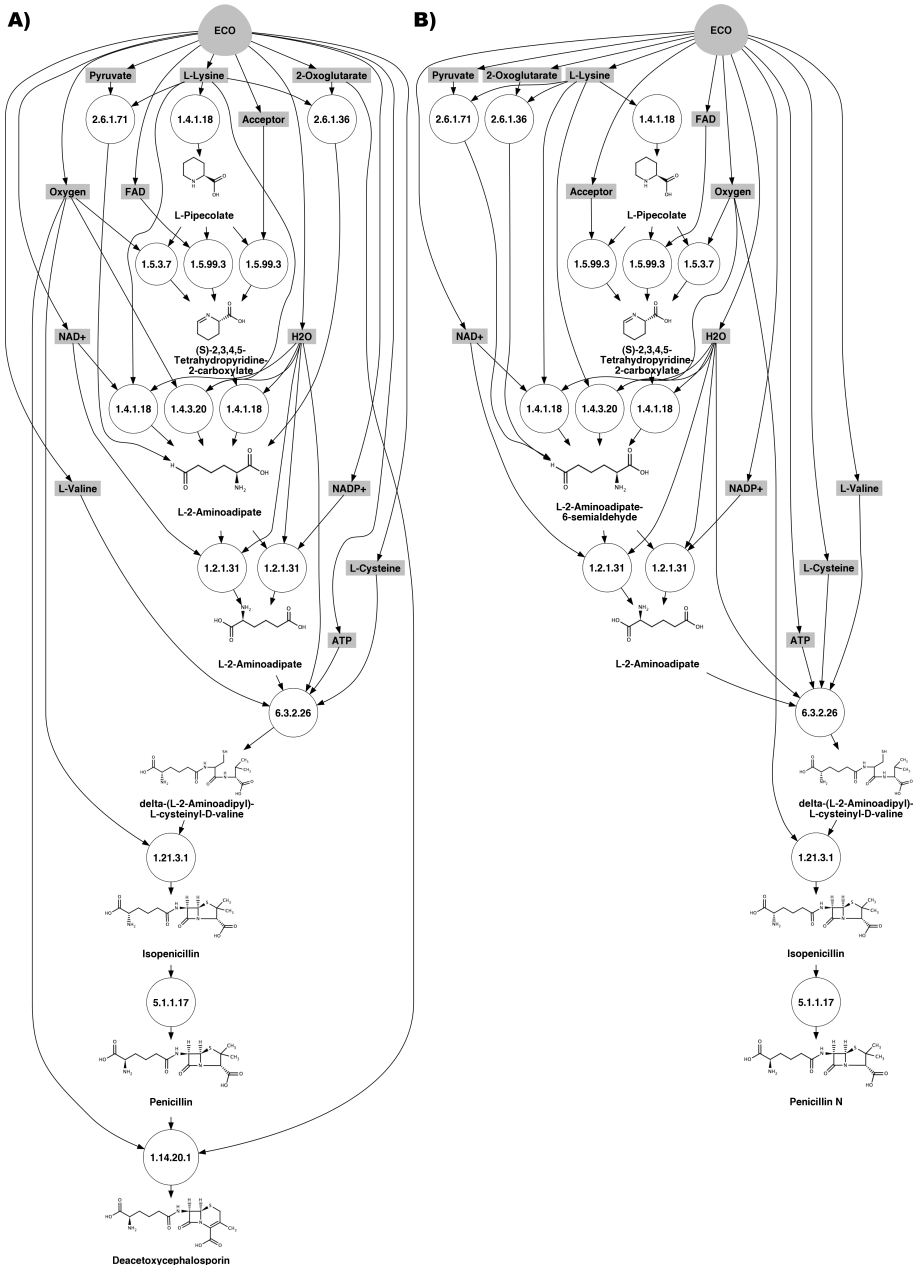



**Figure 6 Retrosynthetic maps for the production in *E. coli* of A) penicillin G and B) cephalosporin**. Compounds in gray are endogenous to the chassis organism (*E. coli*); enzymatic reactions are represented as circles; branching compounds, which can be produced by more than one biochemical transformation, are (S)-2,3,4,5-tetrahydropyridine-2-carboxylate, L-2-aminoadipate-6-semialdehyde; the target compound is at the bottom of the plot.

producing either taxa-4(20),11(12)-dien-5$\alpha$,13$\alpha$-diol by the taxane 13$\alpha$ hydrolase (EC 1.14.13.77) or the taxa-4(20),11(12)-dien-5$\alpha$-yl acetate by the taxadien-5$\alpha$-O-acetyl transferase [71]. Taxane 10$\beta$ hydroxylase (EC 1.14.13.76) is producing the taxa-4(20),11(12)-dien5$\alpha$,10$\beta$-diol 5 acetate [72]. This part of the pathway was implemented in the yeast *Saccharomyces cerevisiae* [73]. The next steps described successively the formation of 10-deacetyl-2-debenzoylbaccatin III (C11899), the 10-deacetylbaccatin III (C11700) catalyzed by taxane 2$\alpha$-O-benzoyltransferase (EC 2.3.1.166), and the Baccatin III catalyzed by taxane 2$\alpha$-O-benzoyltransferase [74] to finally form the taxol.

Table 4 ranks the 2 retrosynthetic pathways in yeast leading to taxol production according to the cost

**Table 3 Ranked pathways for biosynthesis of Penicillin N in *E. coli***

| EC number | product | cost (toxicity) | $\rho_1$ | $\rho_2$ | $\rho_3$ | $\rho_4$ | $\rho_5$ | $\rho_6$ |
|---|---|---|---|---|---|---|---|---|
| 5.1.1.17 | C06564 | 1.17 (2.63) | X | X | X | X | X | X |
| 1.21.3.1 | C05557 | 0.81 (2.94) | X | X | X | X | X | X |
| 6.3.2.26 | C05556 | 2.05 (0.87) | X | X | X | X | X | X |
| 1.2.1.31 | C00956 | 1.09 (0.13) | X | X | X | X | X | X |
| 2.6.1.36 | C04076 | 0.74 (0.72) | X | - | - | - | - | - |
| 1.4.3.20 | C04076 | 1.13 (0.72) | - | - | - | X | - | - |
| 2.6.1.71 | C04076 | 1.34 (0.72) | - | - | X | - | - | - |
| 1.4.1.18 | C04076 | 1.54 (0.72) | - | X | - | - | X | X |
| 1.5.3.7 | C00450 | 4.94 (0.91) | - | - | - | - | - | X |
| 1.5.99.3 | C00450 | 5.30 (0.91) | - | - | - | - | X | - |
| 1.4.1.18* | C00408 | 1.43 (0.93) | - | - | - | - | X | X |
| | | $v_c(\rho)$ | 6.56 | 6.53 | 6.55 | 5.55 | 5.56 | 5.55 |
| | | $W(\rho)$ | 9.33 | 9.67 | 9.93 | 9.74 | 12.52 | 16.52 |

Each row corresponds to the insertion of one enzyme in the pathway in order to produce the given intermediate product in the second column. The estimated cost and product toxicity are given in the third column. The last two rows provide the estimate of maximum flux $v_c(\rho)$ for the pathway, and the total cost $W(\rho)$ according to Equation 3. C06564: Penicillin N; C05557: Isopenicillin; C05556: $\delta$-(L-$\alpha$-Aminoadipyl)-L-cysteinyl-D-valine; C00956: L-2-Aminoadipate; C04076: L-2-Aminoadipate-6-semialdehyde; C00450: (S)-2,3,4,5-Tetrahydropyridine-2-carboxylate; C00408: L-Pipecolate. Starred EC numbers correspond to putative enzymes.

function in Equation 3. In this example, toxicity of intermediates was not considered in the ranking ($\lambda_{tox}$ = 0) and therefore the optimal weighting terms in Equation 3 were taken accordingly (see details in Methods). The two alternative routes are given by the way the precursor 10-deacetyl-2-debenzoylbaccatin is produced. The optimal pathway involves 8 exogenous enzymes, while the alternative one involves 9 enzymes.

## Conclusions

We presented in this work an automated protocol to assist synthetic biologists and metabolic engineers in the design and insertion of efficient heterologous biosynthetic metabolic pathways in chassis organisms. Our method is based on the retrosynthesis algorithm, an idea borrowed from the allied field of synthetic chemistry. In order to perform this analysis in metabolic networks, we used the molecular signature descriptor, a 2D representation of molecular graphs that provides a characterization of compounds and reactions in the network. Molecular signatures provide a homogeneous way to represent through hypergraphs the set of chemical species and transformations present in cell's metabolism. The representation in the molecular signature space is also an efficient way to measure chemical similarity

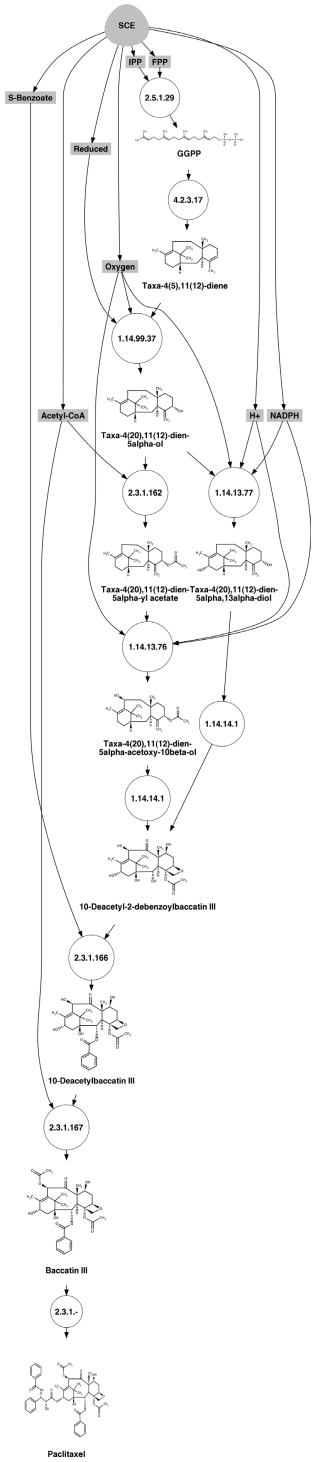



**Figure 7 Retrosynthetic map for the production of paclitaxel (taxol) in yeast**. Compounds in gray are endogenous to the chassis organism (*S. cerevisiae*); enzymatic reactions are represented as circles; 10-deactyl-2-debenzoylbaccatin III appears as a branching compound; the target compound is at the bottom of the plot.

**Table 4 Ranked pathways for biosynthesis of taxol in *S. cerevisiae***

| EC number | product | cost | $\rho_1$ | $\rho_2$ |
|---|---|---|---|---|
| 2.3.1.- * | C07394 | 1.34 | X | X |
| 2.3.1.167 | C11900 | 1.17 | X | X |
| 2.3.1.166 | C11700 | 1.07 | X | X |
| 1.14.14.1* | C11899 | 1.78 | X | X |
| 1.14.13.76 | C11898 | 5.64 | - | X |
| 1.14.13.77 | C11897 | 5.62 | X | - |
| 2.3.1.162 | C11896 | 1.06 | - | X |
| 1.14.99.37 | C11895 | 5.62 | X | X |
| 4.2.3.17 | C11894 | 1.28 | X | X |
| 2.5.1.29 | C00353 | 0.69 | X | X |
| | | $v_c(\rho)$ | 0.094 | 0.093 |
| | | $W(\rho)$ | 18.64 | 19.72 |

Each row corresponds to the insertion of one enzyme in the pathway in order to produce the given intermediate product in the second column. The estimated cost is given in the third column 3. The last two rows provide the estimate of maximum flux $v_c(\rho)$ for the pathway, and the total cost $W(\rho)$ according to Equation 3. C07394: Paclitaxel; C11900: Baccatin; C11700: 10-Deacetylbaccatin; C11899: 10-Deacetyl-2-debenzoylbaccatin; C11898: Taxa-4(20),11(12)-dien-5$\alpha$-acetoxy-10$\beta$-ol; C11897: Taxa-4(20),11(12)-dien-5$\alpha$,13$\alpha$-diol; C11896: Taxa-4(20),11(12)-dien-5$\alpha$-yl; C11895: Taxa-4(20),11(12)-dien-5$\alpha$-ol; C11894: Taxa-4(5),11(12)-diene; C00353: Geranylgeranyl. Starred EC numbers correspond to putative enzymes.

between compounds and reactions. We use this framework to develop an algorithm to generate putative novel reactions between compounds in the network, leading to the extended metabolic reaction space (EMRS). The algorithm consists of searching for all combinations of compounds with the same signature as known metabolic reactions.

In order to explore new biosynthetic pathways for compounds, we implemented a retrosynthetic algorithm in the EMRS to build the map of all reachable compounds from the chassis organism through biochemical transformations. The complexity of the retrosynthetic search, a problem that has been limiting so far the adoption of the retrosynthetic approach into the manufacturing pipeline, has been efficiently addressed in our method through the atomic height $h$ of the molecular signature. As $h$ increases, the number of possible pathways converges to the annotated pathways in metabolic databases. Lowering $h$, in turn, leads progressively to a combinatorial explosion with multiple alternative pathways containing putative reactions, as the ones obtained in the bond-electron and atom tracking models of the metabolic graph.

The retrosynthetic map contains both annotated and putative reactions catalyzed by identified exogenous enzymes, providing several alternative pathways leading to the target compound. We associated a cost of insertion to each pathway based on several criteria such as gene insertion cost, expression levels, enzyme efficiency and nominal fluxes. Furthermore, an algorithm similar to the shortest pathway search has been implemented in order to rank all possible pathways. We showed that the distribution of alternative biosynthetic pathways for *E. coli* in the list of compounds of medicinal interest in DrugBank follows a power law, being in some cases in the order of thousands. Therefore, it is necessary to implement an efficient ranking function as the one presented here in order to select the best heterologous pathways to insert in the chassis organism. For instance, we applied the retrosynthetic algorithm in order to search for heterologous biosynthetic pathways for two compounds in DrugBank: penicillin N and taxol. In both cases, several alternative pathways for bioproduction were found. The identified pathways contained both known biochemical transformations previously reported as well as other alternative pathways. In order to select the best combinations to engineer, pathways were ranked according to several cost factors such as number of inserted enzymes, gene compatibility, toxicity, and nominal fluxes. The individual contribution of these factors to the ranking function was optimally adjusted so that native pathways were ranked first with respect to predicted pathways. Our multi-criteria approach can be easily tuned depending on the data available for organisms, as it was illustrated by the optimal adjustment of the weighting parameters for different combinations of factors. The ability of the ranking function to identify native pathways was tested and validated in the case of biosynthetic pathways of several essential metabolites (amino acids, citrate, ATP/ADP, GTP/GDP) in auxotrophic strains of *E. coli*. In this test, native enzymes were correctly ranked by means of our methodology at the top of the enumerated biosynthetic pathways.

Even though our methodology searches for enzymes providing the best performance for the overall process, we might be interested in some cases in increasing the efficiency levels for some of the promiscuous reactions involved in the pathway due to their poor performance, a rate-limiting factor in the production of the target compound. In that case, it would be necessary to introduce mutations in order to re-engineer enzyme variants with detectable levels of the desired catalytic activity. Using protein molecular signatures and kernel methods, we already proposed a methodology to search for promiscuity hot-spot residues in the enzyme sequence and outlined a method to find variants with enhanced promiscuity levels [46], which might be applicable in this case. Therefore, this work provides a full biosynthetic automated pipeline for the design and production of therapeutics and other compounds in flexible on-demand cell factories. Going beyond classical metabolic engineering, our synthetic biology approach meets the

expected requirements of reusability and modularity that will become integral part of next generation biosynthetic devices.

## Methods

### Definitions

#### *Atomic signature*

Let $G = (V, E)$ be a molecular graph, where vertices $V$ correspond to atoms, and edges $E$ to bonds. An atomic signature is a canonical representation of the subgraph of $G$ surrounding a particular atom $x \in V$. This subgraph includes all atoms and bonds up to a predefined distance from the given atom, the signature height $h$.

#### *Molecular signature*

The molecular signature is a vector whose components are represented in the space defined by a basis formed by atomic signatures. Initially developed for chemicals [39], the signature molecular descriptor was later extended to protein sequences [37,75]. Each component of a molecular signature counts the number of occurrences of a particular atomic signature in the molecule. If $G = (V, E)$ is a molecular graph, where vertices $V$ correspond to atoms, and edges $E$ to bonds, then the molecular signature of $G$ is given by:

$$^{h}\sigma(G) = \sum_{x \in V} {}^{h}\sigma(x_i) \tag{5}$$

where $^{h}\sigma(x)$ is the atomic signature of $G$ rooted at atom $x_i$ of height $h$.

#### *Reaction signature*

We assume that enzymatic reactions take the general form: $r : s_1S_1 + ... + s_nS_n \rightarrow p_1P_1 + ... + p_mP_m$, where $s_i$ and $p_j$ are the stoichiometric coefficients of substrates $S_i$ and products $P_j$. The signature of reaction $R$ of height $h$ is defined by the vector:

$$^{h}\sigma(r) = \left( \sum_{P_i \in r} p_i {}^{h}\sigma(P_i) - \sum_{S_i \in r} s_i {}^{h}\sigma(S_i) \right) \tag{6}$$

#### *Metabolic reaction space*

In a metabolic network, the reaction space $R$ is formed by the set of reactions $r$ in the network, which are defined as ordered pairs of substrates $s \in C$ and products $p \in C$ belonging to the metabolite space $C$:

$$r = \left( \{s\}, \{p\} \right) \tag{7}$$

#### *Metabolic reaction signature space*

The signature reaction space $^{h}\sigma(R)$ of height $h$ is given by mapping of the set of metabolic reactions $r \in R$ into signature reactions $^{h}\sigma(r)$ according to Equation 7. **Extended metabolic reaction space (EMRS)**. The extended metabolic reaction space generated by signatures of height $h$, $^{h}\sigma^{-1}(R)$, corresponds to the inverse mapping from the signature space into the reaction space. Since the projection of $R$ into the signature space $^{h}\sigma(R)$ involves some degeneracy, the regeneration of the metabolic map creates new putative reactions consisting of combinations of substrates and products that verify the reaction signatures in addition to the nominal ones:

$$R \rightarrow {}^{h}\sigma(R) \rightarrow {}^{h}\sigma^{-1}(R) \tag{8}$$

#### *Reaction chemical similarity*

We define a measure of chemical similarity in the signature space between reaction signatures of height $h$ by using the Tanimoto similarity coefficient:

$$^{h}s(r_i, r_j) = \frac{|{}^{h}\sigma(r_i) \cdot {}^{h}\sigma(r_j)|}{|{}^{h}\sigma(r_i)|^2 + |{}^{h}\sigma(r_j)|^2 - |{}^{h}\sigma(r_i) \cdot {}^{h}\sigma(r_j)|} \tag{9}$$

where operations are applied into the vector space determined by the net difference of the signatures of products and substrates (Equation 7).

This similarity measure focus on the reaction centers that define the chemical transformation rather than on the full atomic structure. As the height $h$ is increased up to the maximum diameter of the graph or canonical signature, the similarity measure extends further up to the rest of the molecular structure. By definition, two reactions $r$ and $r^*$ that share the same signature up to some height $h$ possess identical signatures up to that molecular resolution $h$:

$$^{h}s(r, r^*) = 1, \quad h = 1 \ldots \arg\max_{h} {}^{h}\sigma(r) = {}^{h}\sigma(r^*) \tag{10}$$

#### *Exogenous biosynthetic pathway*

An exogenous biosynthetic pathway $\rho \in \rho(c)$ for a target compound $c \in C$ is defined as a collection of reactions $\{r_1, r_2, ..., r_n\}$ in the EMRS that connects metabolites in the chassis organism to the product $c$ through biochemical transformations.

### Ranking terms

**Reaction thermodynamic feasibility** was computed through the estimation of Gibbs energy of the reaction by using a group contribution approach [43,50]. This method considers the Gibbs energy of each metabolite species as the sum of the contributions of their constituents structural subgroups, estimated by linear regression from experimental data. We used the dataset given in [43] in order to compute metabolite Gibbs energy, whereas the reaction Gibbs energy is computed as the energetic balance between its products and substrates:

$$\Delta G_r = \sum_{i \in r} n_i \Delta G_i \quad (11)$$

where $n_i$ is the stoichiometric coefficient of each species, and $\Delta G_i$ their estimated Gibbs energy.

**Enzyme promiscuity** was estimated by a support vector machine that was trained from the string or *k*-mer spectra [76] of enzyme sequences ${}^k\sigma(S)$ in KEGG [24] as inputs by defining the following kernel function:

$${}^kK(S_i, S_j) = {}^k\sigma(S_i)^k\sigma(S_j) \quad (12)$$

Enzyme promiscuity in the training set was defined by comparing the chemical similarity of reactions catalyzed by the enzyme sequence *S*, as in [46]:

$$\begin{cases} {}^hs(r_i, r_j) = 0 & \texttt{non - promiscuous} \\ {}^hs(r_i, r_j) > 0 & \texttt{promiscuous} \end{cases} \quad i, j \in S \quad (13)$$

**Reaction clustering** of the reaction signature space was performed by a hierarchical agglomerative algorithm using as distance metrics the chemical dissimilarity between reactions $d(r_i, r_j)$:

$${}^hd(r_i, r_j) = 1 - {}^hs(r_i, r_j) \quad (14)$$

The optimal partition of the reaction signature space into $C_i$, $i = 1...n$ clusters was determined by the maximum average silhouette [77].

**Enzyme-metabolite interaction** prediction was computed within a given signature reaction space cluster by using a kernel approach known as tensor product [51]. For each reaction cluster, a training set was built consisting of pairs of known enzyme sequences and reactions annotated in KEGG. This dataset was used in order to train a support vector machine defined by the following kernel function:

$${}^{k,h}K((S_i, r_i), (S_j, r_j)) = {}^kK(S_i, S_j) \cdot {}^hK(r_i, r_j) \quad (15)$$

where ${}^kK(S_i, S_j)$ is the sequence string kernel defined in Equation 12 and ${}^hk(r_i, r_j)$ is given by the reaction similarity matrix computed by using the reaction chemical similarity $s(r_i, r_j)$ defined in Equation 14. **Enzyme performance** was estimated through a decision tree algorithm implemented for each reaction cluster, as in [53]. Performance data were based on the experimental kinetic constants $k_{cat}$ and $K_M$ provided by the BRENDA database [54]. Input features consisted of chemical descriptors of substrates and products in reactions and protein sequence descriptors.

#### *Gene compatibility*

Sequence descriptors were computed from the EMBOSS package of sequence analysis [78]. Phylogenetic distance between the source organism and chassis organism was computed from KEGG taxonomy. These descriptors were computed for the entire KEGG database of non-redundant enzyme sequences and then used to train support vector machine-based predictors for the chassis organisms of interest (*E. coli* and yeast). The training set consisted of a balanced positive set, formed by the list of sequences in the chassis strains, and a negative set formed by sequences selected randomly from the list of organisms other than the chassis. In the model of *E. coli*, we found that the average score for positive hits had a *z*-score = 6.12 (*p*-value = 9.56e-10) for a positive set of 24,894 sequences in E. coli strains among the total set of 681,518 sequences. We used this predictor in order to rank the annotated genes for a given enzyme class, where a *p*-value was associated to each predicted gene by computing the probability of ranking that gene in the given percentile if it were picked at random from the list of genes.

**Compound toxicity** in the chassis organism *E. coli* was estimated through a partial least squares structure-activity relationship model implemented from an in-house database for *E. coli* of 150 compounds with experimentally determined IC50 (half maximal inhibitory concentration). Input descriptors of the model are given by the following molecular descriptors: molecular weight; solubility; average bond length; partition coefficient; molecular surface; and molecular signatures of the compounds.

**Nominal fluxes v** were predicted by using a reconstructed metabolic model of *E. coli* [79] and yeast [80] in the COBRA toolbox [81]. For a given target compound $c \in C$ and a putative exogenous biosynthetic pathway $\rho$, an augmented model of the metabolic phenotype of the engineered strain was built from the reference model. The nominal flux of the desired compound $v_c(\rho)$ in the augmented model was obtained through linear programming optimization of the stoichiometric mass balance subject to the following constraints:

$$\begin{array}{ll} \texttt{Maximize} & f(v_c(\rho), Z) \\ \texttt{Subject to} & \mathbf{S} \cdot \mathbf{v} = \mathbf{0} \\ & \alpha \le v_i \le \beta \end{array} \quad (16)$$

where **S** is the stoichiometric matrix, *Z* is the objective function of maximizing the biomass formation (growth) rate, $f(v_c(\rho), Z)$ is a definite positive function that monotonically increases with both $v_c(\rho)$ and *Z*, and $\alpha$, $\beta$ are the model flux constraints [79]. The chosen objective function in Equation 16 was $f(v_c(\rho), Z) = v_c(\rho) \cdot Z$, although in general other objectives might be possible as well (see for instance in [82]).

#### *Parameter optimization*

Weighting parameters ($\lambda_{\texttt{flux}}$, $\lambda_{\texttt{path}}$, $\lambda_{\texttt{tox}}$) in the cost function given by Equation 3 are adjusted by optimization. The chosen criteria is that pathways that are fully

annotated in KEGG should be ranked first with respect to other pathways based solely on predictions. Our approach is similar to the one proposed in [38], although the main difference here is that we optimize the three parameters simultaneously for all metabolites in KEGG by using an estimate of ranking accuracy for each pathway within the full set of enumerated pathways in the EMRS. For a pathway $\rho \in \rho(c)$ producing a given compound $c$, we define its ranking accuracy as:

$$\gamma_\rho = \frac{n_{TP}(\rho) + n_{TN}(\rho)}{|\rho(c)|} \quad (17)$$

where $n_{TP}$ are the number of pathways of $\rho(c)$ in KEGG that are ranked at the same or higher score than $\rho$, $n_{TN}$ are the number of pathways of $\rho(c)$ not in KEGG which are ranked below $\rho$, and $|\rho(c)|$ are the total number of pathways producing $c$.

The objective is to maximize the overall aggregate sum of $\gamma_\rho$ extended to the list of enumerated pathways in the EMRS:

$$\begin{array}{cc} \texttt{Maximize} & \sum_{c\in C}\sum_{\rho\in\rho(c)}\gamma_\rho\left(\lambda_{\texttt{path}},\lambda_{\texttt{tox}},\lambda_{\texttt{flux}}\right) \\ \lambda_{\texttt{path}},\lambda_{\texttt{tox}},\lambda_{\texttt{flux}} & \\ \texttt{Subject to} & \lambda_{\texttt{path}},\lambda_{\texttt{tox}},\lambda_{\texttt{flux}} > 0 \end{array} \quad (18)$$

For the ranking optimization problem, parameters ($\lambda_{\texttt{path}}$, $\lambda_{\texttt{tox}}$, $\lambda_{\texttt{flux}}$) are not independent since a simultaneous increase of the same magnitude in the three parameters would leave the ranking unchanged. Therefore, in order to solve the problem, we need to fix at least the value of one of the parameters, for instance taking $\lambda_{\texttt{path}} = 1.0$ and $\omega_p = \omega_e = 1$. We give three possible solutions depending whether both toxicity and fluxes estimates are available or only one of them (see parameter variations in Additional file 1 Figures S1, S2 and S3): without considering fluxes ($\lambda_{\texttt{flux}} = 0.0$), the optimal value for the toxicity parameter was $\lambda^*_{\texttt{tox}} = 0.575$; without considering toxicity ($\lambda^*_{\texttt{tox}} = 0.0$), the optimal value for the flux parameter was $\lambda^*_{\texttt{flux}} = 0.800$; finally considering both toxicity and fluxes, the optimal values were obtained at $\lambda^*_{\texttt{tox}} = 0.398$, $\lambda^*_{\texttt{tox}} = 0.398$.

## Additional material

**Additional file 1: Pathway ranking accuracies for different values of parameters ($\lambda_{tox}$, $\lambda_{flux}$)**. Figure S1 plots pathway ranking accuracy for different values of parameter $\lambda_{\texttt{tox}}$ without considering fluxes ($\lambda_{\texttt{flux}} = 0$); optimal value is obtained for $\lambda^*_{\texttt{tox}} = 0.575$. Figure S2 plots pathway ranking accuracy for different values of parameter $\lambda_{\texttt{flux}}$ without considering toxicity ($\lambda_{\texttt{tox}} = 0$); optimal value is obtained for $\lambda^*_{\texttt{flux}} = 0.800$. Figure S3 plots pathway ranking accuracy for different values of parameters ($\lambda_{\texttt{tox}}$, $\lambda_{\texttt{flux}}$); optimal values are ($\lambda^*_{\texttt{flux}} = 0.025$, $\lambda^*_{\texttt{tox}} = 0.398$).

**Acknowledgements**
The authors want to acknowledge the assistance of Fred Green in the computation of the flux balance analysis for the given examples and proof-reading the manuscript. The authors want to thank Ioana Grigoras who reviewed an early version of the manuscript. Funding: Genopole® through an ATIGE grant; ANR chair of excellence. Conflict of Interest: none declared.

**Authors' contributions**
JLF designed the retrosynthesis overall process. PC designed the experiments and pathway ranking strategy, collected the results, and wrote parts of the software. AGP designed and carried out the toxicity experiments, participated in the design of the pathway ranking strategy, and analyzed the retrosynthetic examples. DF developed the algorithms. PC, AGP, DF, and JLF wrote the manuscript. All authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

Received: 18 April 2011 Accepted: 5 August 2011
Published: 5 August 2011

**References**

1. Elowitz M, Lim WA: **Build life to understand it.** *Nature* 2010, **468(7326)**:889-890.
2. Keasling JD: **Manufacturing Molecules Through Metabolic Engineering.** *Science* 2010, **330(6009)**:1355-1358.
3. Khalil AS, Collins JJ: **Synthetic biology: applications come of age.** *Nat Rev Genet* 2010, **11(5)**:367-379.
4. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J, Chang MCY, Withers ST, Shiba Y, Sarpong R, Keasling JD: **Production of the antimalarial drug precursor artemisinic acid in engineered yeast.** *Nature* 2006, **440(7086)**:940-943.
5. Ajikumar PK, Xiao WH, Tyo KEJ, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G: **Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli.** *Science* 2010, **330(6000)**:70-74.
6. Minami H, Kim JS, Ikezawa N, Takemura T, Katayama T, Kumagai H, Sato F: **Microbial production of plant benzylisoquinoline alkaloids.** *Proc Natl Acad Sci USA* 2008, **105(21)**:7393-7398.
7. Fowler ZL, Gikandi WW, Koffas MAG: **Increased Malonyl Coenzyme A Biosynthesis by Tuning the Escherichia coli Metabolic Network and Its Application to Flavanone Production.** *Appl Environ Microbiol* 2009, **75(18)**:5831-5839.
8. Hwang EI, Kaneko M, Ohnishi Y, Horinouchi S: **Production of Plant-Specific Flavanones by Escherichia coli Containing an Artificial Gene Cluster.** *Appl Environ Microbiol* 2003, **69(5)**:2699-2706.
9. Kurumbang NP, Park JW, Yoon YJ, Liou K, Sohng JK: **Heterologous production of ribostamycin derivatives in engineered Escherichia coli.** *Res Microbiol* 2010, **161(7)**:526-533.
10. Watanabe K, Rude MA, Walsh CT, Khosla C: **Engineered biosynthesis of an ansamycin polyketide precursor in Escherichia coli.** *Proc Natl Acad Sci USA* 2003, **100(17)**:9774-9778.
11. Peiru S, Menzella HG, Rodriguez E, Carney J, Gramajo H: **Production of the Potent Antibacterial Polyketide Erythromycin C in Escherichia coli.** *Appl Environ Microbiol* 2005, **71(5)**:2539-2547.
12. Watanabe K, Oguri H, Oikawa H: **Diversification of echinomycin molecular structure by way of chemoenzymatic synthesis and heterologous expression of the engineered echinomycin biosynthetic pathway.** *Curr Opin Chem Biol* 2009, **13(2)**:189-196.
13. Soo VWC, Hanson-Manful P, Patrick WM: **Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli.** *Proc Natl Acad Sci USA* 2011, **108(4)**:1484-1489.
14. Menzella H, Reeves C: **Combinatorial biosynthesis for drug development.** *Curr Opin Microbiol* 2007, **10(3)**:238-245.
15. Menzella HG, Reid R, Carney JR, Chandran SS, Reisinger SJ, Patel KG, Hopwood DA, Santi DV: **Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes.** *Nat Biotechnol* 2005, **23(9)**:1171-1176.

16. Kim HU, Kim TY, Lee SY: **Metabolic flux analysis and metabolic engineering of microorganisms.** *Mol BioSyst* 2008, **4(2)**:113-120.
17. Lacroix V, Cottret L, Thébault P, Sagot MFF: **An introduction to metabolic networks and their structural analysis.** *IEEE/ACM Trans Comp Biol Bioinfo* 2008, **5(4)**:594-617.
18. Cottret L, Vieira Milreu P, Acuña V, Marchetti-Spaccamela A, Viduani Martinez F, Sagot MF, Stougie L: **Enumerating Precursor Sets of Target Metabolites in a Metabolic Network.** In *Algorithms in Bioinformatics, Volume 5251 of Lecture Notes in Computer Science*. Edited by: Crandall K, Lagergren J. Berlin, Heidelberg: Springer Berlin/Heidelberg; 2008:233-244.
19. Mithani A, Preston GM, Hein J: **Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison.** *Bioinformatics* 2009, **25(14)**:1831-1832.
20. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions.** *Mol Syst Biol* 2009, **5**.
21. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks.** *Proc Natl Acad Sci USA* 2002, **99(23)**:15112-15117.
22. Burgard AP, Pharkya P, Maranas CD: **Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnol Bioeng* 2003, **84(6)**:647-657.
23. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U: **COPASI-a COmplex PAthway SImulator.** *Bioinformatics* 2006, **22(24)**:3067-3074.
24. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, , **36 Database:** D480-484.
25. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2010, **38(suppl 1)**: D473-D479.
26. Arita M: **Metabolic reconstruction using shortest paths.** *Simulation Practice and Theory* 2000, **8(1-2)**:109-125.
27. Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D: **Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC).** *Bioinformatics* 2005, **21(7)**:1189-1193.
28. Blum T, Kohlbacher O: **MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization.** *Bioinformatics* 2008, **24(18)**:2108-2109.
29. Heath AP, Bennett GN, Kavraki LE: **Finding metabolic pathways using atom tracking.** *Bioinformatics* 2010, **26(12)**:1548-1555.
30. Paley SM, Karp PD: **Evaluation of computational metabolic-pathway predictions for Helicobacter pylori.** *Bioinformatics* 2002, **18(5)**:715-724.
31. Dale J, Popescu L, Karp P: **Machine learning methods for metabolic pathway prediction.** *BMC Bioinformatics* 2010, **11**:15+.
32. Law J, Zsoldos Z, Simon A, Reid D, Liu Y, Khew SY, Johnson AP, Major S, Wade RA, Ando HY: **Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation.** *J Chem Inf Model* 2009, **49(3)**:593-602.
33. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ: **Exploring the diversity of complex metabolic networks.** *Bioinformatics* 2005, **21(8)**:1603-1609.
34. Ugi I, Bauer J, Brandt J, Friederich J, Gasteiger J, Jochum C, Schubert W: **New applications of computers in chemistry.** *Angewandte Chemie* 1979, **18(2)**:111-123.
35. Tipton K, Boyce S: **History of the enzyme nomenclature system.** *Bioinformatics* 2000, **16**:34-40.
36. Leber M, Egelhofer V, Schomburg I, Schomburg D: **Automatic assignment of reaction operators to enzymatic reactions.** *Bioinformatics* 2009, **25(23)**:3135-3142.
37. Faulon JL, Misra M, Martin S, Sale K, Sapra R: **Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor.** *Bioinformatics* 2008, **24(2)**:225-233.
38. Cho A, Yun H, Park J, Lee S, Park S: **Prediction of novel synthetic pathways for the production of desired chemicals.** *BMC Syst Biol* 2010, **4**:35+.
39. Faulon JLL, Collins MJ, Carr RD: **The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences.** *J Chem Infor Comp Sci* 2004, **44(2)**:427-436.
40. Faulon JL, Carbonell P: **Reaction Network Generation.** In *Handbook of Chemoinformatics Algorithms*.. 1 edition. Edited by: Faulon JL, Bender A. Boca Raton, FL, USA: Chapman and Hall/CRC; 2010:317-342.
41. McShan D, Shah I: **Heurstic search for metabolic engineering: de novo synthesis of vanillin.** *Comp & Chem Eng* 2005, **29(3)**:499-507.
42. Mavrovouniotis ML: **Estimation of standard Gibbs energy changes of biotransformations.** *J Biol Chem* 1991, **266(22)**:14440-14445.
43. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V: **Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks.** *Biophys J* 2008, **95(3)**:1487-1499.
44. Rodrigo G, Carrera J, Prather KJ, Jaramillo A: **DESHARKY: automatic design of metabolic pathways for optimal cell growth.** *Bioinformatics* 2008, **24(21)**:2554-2556.
45. Chen Z, Wilmanns M, Zeng AP: **Structural synthetic biotechnology: from molecular structure to predictable design for industrial strain development.** *Trends Biotechnol* 2010, **28(10)**:534-542.
46. Carbonell P, Faulon JLL: **Molecular signatures-based prediction of enzyme promiscuity.** *Bioinformatics* 2010, **26(16)**:2012-2019.
47. Wexler P: **The U.S. National Library of Medicine's Toxicology and Environmental Health Information Program.** *Toxicology* 2004, **198(1-3)**:161-168.
48. Harder A, Escher BI, Schwarzenbach RP: **Applicability and Limitation of QSARs for the Toxicity of Electrophilic Chemicals.** *Environ SciTech* 2003, **37(21)**:4955-4961.
49. Khersonsky O, Tawfik DS: **Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective.** *Annu Rev Biochem* 2010, **79**:471-505.
50. Fleming RMT, Thiele I, Nasheuer HP: **Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to Escherichia coli.** *Biophys Chem* 2009, **145(2-3)**:47-56.
51. Martin S, Brown MM, Faulon JLL: **Using product kernels to predict protein interactions.** *Adv Biochem Eng Biotechnol* 2008, **110**:215-245.
52. Gasteiger J, Engel T, (Eds): *Chemoinformatics: A Textbook*. 1 edition. Wiley-VCH; 2003.
53. Sarnowski C, Carbonell P, Elati M, Faulon JL: **Prediction of catalytic efficiency to discover new enzymatic activities.** *Proc. of the Fourth International Workshop on Machine Learning in Systems Biology* 2010, 153-156.
54. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.** *Nucleic Acids Res* 2009, , **37 Database:** gkn820+.
55. Welch M, Villalobos A, Gustafsson C, Minshull J: **Designing genes for successful protein expression.** *Methods Enzymol* 2011, **498**:43-66.
56. Alper H, Fischer C, Nevoigt E, Stephanopoulos G: **Tuning genetic control through promoter engineering.** *Proc Natl Acad Sci USA* 2005, **102(36)**:12678-12683.
57. Chemler J, Koffas M: **Metabolic engineering for plant natural product biosynthesis in microbes.** *Curr Opin Biotechnol* 2008, **19(6)**:597-605.
58. Boghigian B, Pfeifer B: **Current status, strategies, and potential for the metabolic engineering of heterologous polyketides in Escherichia coli.** *Biotechnology Letters* 2008, **30(8)**:1323-1330.
59. Watanabe K, Hotta K, Praseuth AP, Koketsu K, Migita A, Boddy CN, Wang CCC, Oguri H, Oikawa H: **Total biosynthesis of antitumor nonribosomal peptides in Escherichia coli.** *Nat Chem Biol* 2006, **2(8)**:423-428.
60. Nguyen KT, Ritz D, Gu JQQ, Alexander D, Chu M, Miao V, Brian P, Baltz RH: **Combinatorial biosynthesis of novel antibiotics related to daptomycin.** *Proc Natl Acad Sci USA* 2006, **103(46)**:17462-17467.
61. Planson AG, Carbonell P, Paillard E, Pollet N, Faulon JL: **Compound toxicity screening and structure-activity relationship modeling in Escherichia coli.** 2011.
62. Klamt S, Haus UU, Theis F: **Hypergraphs and Cellular Networks.** *PLoS Comput Biol* 2009, **5(5)**:e1000385+.
63. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo ACC, Wishart DS: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, , **39 Database:** D1035-D1041.

64. Gidijala L, Kiel JAKW, Douma RD, Seifar RM, van Gulik WM, Bovenberg RAL, Veenhuis M, van der Klei IJ: **An Engineered Yeast Efficiently Secreting Penicillin.** *PLoS ONE* 2009, **4(12)**:e8317+.
65. Ullán RV, Casqueiro J, Bañuelos O, Fernández FJ, Gutiérrez S, Martin JF: **A Novel Epimerization System in Fungal Secondary Metabolism Involved in the Conversion of Isopenicillin N into Penicillin N inAcremonium chrysogenum.** *J Biol Chem* 2002, **277(48)**:46216-46225.
66. Frense D: **Taxanes: perspectives for biotechnological production.** *Appl Microbiol Biotechnol* 2007, **73(6)**:1233-1240.
67. Hefner J, Ketchum RE, Croteau R: **Cloning and functional expression of a cDNA encoding geranylgeranyl diphosphate synthase from Taxus canadensis and assessment of the role of this prenyltransferase in cells induced for taxol production.** *Arch Biochem Biophys* 1998, **360**:62-74.
68. Wildung MR, Croteau R: **A cDNA Clone for Taxadiene Synthase, the Diterpene Cyclase That Catalyzes the Committed Step of Taxol Biosynthesis.** *J Biol Chem* 1996, **271(16)**:9201-9204.
69. Hefner J, Rubenstein SM, Ketchum RE, Gibson DM, Williams RM, Croteau R: **Cytochrome P450-catalyzed hydroxylation of taxa-4(5),11(12)-diene to taxa-4(20),11(12)-dien-5alpha-ol: the first oxygenation step in taxol biosynthesis.** *Chem Biol* 1996, **3(6)**:479-489.
70. Huang KX, Huang QL, Wildung MR, Croteau R, Scott AI: **Overproduction, in Escherichia coli, of soluble taxadiene synthase, a key enzyme in the Taxol biosynthetic pathway.** *Protein Expression and Purification* 1998, **13**:90-96.
71. Jennewein S, Rithner CD, Williams RM, Croteau RB: **Taxol biosynthesis: taxane 13 alpha-hydroxylase is a cytochrome P450-dependent monooxygenase.** *Proc Natl Acad Sci USA* 2001, **98(24)**:13595-13600.
72. Schoendorf A, Rithner CD, Williams RM, Croteau RB: **Molecular cloning of a cytochrome P450 taxane 10 beta-hydroxylase cDNA from Taxus and functional expression in yeast.** *Proc Natl Acad Sci USA* 2001, **98(4)**:1501-1506.
73. Dejong JM, Liu Y, Bollon AP, Long RM, Jennewein S, Williams D, Croteau RB: **Genetic engineering of taxol biosynthetic genes in Saccharomyces cerevisiae.** *Biotechnol Bioeng* 2006, **93(2)**:212-224.
74. Walker K, Croteau R: **Molecular cloning of a 10-deacetylbaccatin III-10-O-acetyl transferase cDNA from Taxus and functional expression in Escherichia coli.** *Proc Natl Acad Sci USA* 2000, **97(2)**:583-587.
75. Martin S, Roe D, Faulon JL: **Predicting protein-protein interactions using signature products.** *Bioinformatics* 2005, **21(2)**:218-226.
76. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS: **Mismatch string kernels for discriminative protein classification.** *Bioinformatics* 2004, **20(4)**:467-476.
77. Rousseeuw P: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.** *J Comput Appl Math* 1987, **20**:53-65.
78. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6)**:276-277.
79. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Bernhard : **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol* 2007, **3.**
80. Duarte NC, Herrgård MJ, Palsson B: **Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model.** *Genome Res* 2004, **14(7)**:1298-1309.
81. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nat Protoc* 2007, **2(3)**:727-738.
82. Yousofshahi M, Lee K, Hassoun S: **Probabilistic pathway construction.** *Metab Eng* 2011, **13(4)**:435-444.

doi:10.1186/1752-0509-5-122
**Cite this article as:** Carbonell *et al.*: A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Systems Biology* 2011 **5**:122.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central