Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering.

Tamás Fehér, Anne-Gaëlle Planson, Pablo Carbonell, Alfred Fernández-Castané, Ioana Grigoras, Ekaterina Dariy, Alain Perret, Jean-Loup Faulon

Validation of RetroPath, a Computer Aided Design Tool for Metabolic Pathway Engineering

Tamás Fehér*[, 1], Anne-Gaëlle Planson*[, 1, 2], Pablo Carbonell [1], Alfred Fernández-Castané [1], Ioana Grigoras [1], Ekaterina Dariy [3], Alain Perret [3], Jean-Loup Faulon [‡, 1]

* These authors contributed equally to this work

[1] Institute of Systems and Synthetic Biology, University of Evry-Val-d'Essonne, CNRS FRE3561, Genopole® Campus 1, Genavenir 6, 5 rue Henri Desbruères, F-91030 Evry Cedex, France

[2] Present address: INRA, AgroParisTech, UMR1319 Micalis, Jouy-en-Josas F-78350, France

[3] Genoscope, CEA, INSERM, CNRS UMR8030, 2 rue Gaston Crémieux, F-91057 Evry Cedex, France

[‡] Corresponding author. Tel.: +33(0)1 69 47 44 47 email: jean-loup.faulon@issb.genopole.fr address: ISSB, 5 rue Henri Desbruères, Genavenir 6, F-91030 Evry Cedex, France

Keywords

Synthetic biology; Computer-aided design; Rational pathway engineering; Pathway prediction; Flavonoid production

Abstract

Metabolic engineering has succeeded in biosynthesis of numerous commodity or high value compounds. However, the choice of pathways and enzymes used for such processes was many times made *ad hoc*, or required expert knowledge of the specific biochemical reactions. In order to rationalize this process we developed the computer-aided design (CAD) tool RetroPath that explores and enumerates metabolic pathways connecting the endogenous metabolites of a chassis cell to the target compound. To experimentally validate our tool, we constructed twelve top-ranked enzyme combinations producing the flavonoid pinocembrin, four of which displayed significant yields. Namely, our tool queried the enzymes found in metabolic databases based on their annotated and predicted activities. Next, it ranked pathways based on the predicted efficiency of the available enzymes, the toxicity of the intermediate metabolites and the calculated maximum product flux. To implement the top-ranking pathway, our procedure narrowed down a list of nine million possible enzyme combinations to twelve, a number easily assembled and tested. One round of metabolic network optimization based on RetroPath output further increased pinocembrin titers 17-fold. In total, 12 out of the 13 enzymes tested in this work displayed a performance that was in accordance with its predicted score. These results validate the ranking function of our CAD tool, and open the way to its utilization in the biosynthesis of novel compounds.

Practical applications

Our computer-aided design tool, RetroPath allows the engineering of overproducer strains without detailed knowledge on the natural biosynthesis routes of the compounds of interest. RetroPath explores and ranks the pathways connecting the host metabolome with the target compound and provides the highest-ranking enzymes available in public knowledge bases for each step. Besides finding the most efficient routes for bioproduction, it may allow

engineering breakdown- or biotransformation-pathways as well. Its greatest potential lies in the capability of exploring novel biochemical reactions yielding compounds not yet detected in living organisms.

Abbreviations

CAD: computer aided design; PAL: phenylalanine ammonia lyase; 4CL: coumaroyl-CoA ligase; CHS: chalcone synthase; CHI: chalcone isomerase ; FBA: flux balance analysis; DW: dry weight; OD: optical density; LIC: ligation-independent cloning.

1. Introduction

The field of metabolic engineering has undergone significant advances through the development of novel strategies combining genetic engineering and systems biology to allow the production of many organic compounds using a cellular host. The diversity of molecules biosynthesized in engineered hosts has recently been extended, the feat of further extension however, remains challenging [1]. The development of novel biosynthetic routes is a multi-step process not only implying the identification of biosynthetic enzymes and their assembly to create a metabolic pathway, but also characterizing the interconnection of the pathway with the host metabolism. Consequently, a number of computational tools as well as experimental techniques have been developed [2]. A large number of constraint–based frameworks have been proposed to optimize the production host for the synthesis of a molecule of interest, or to determine individual performances for the predicted pathways [3]. However, in the majority of the papers published today, enzymes are chosen *ad hoc*, relying on information available in the literature or on expertise accumulated by the workgroup [4-6]. Several factors, such as enzyme efficiency, byproduct-toxicity and competitive pathways may influence the

performance of an engineered strain. To take into account these constraints, we are using here a unified framework to design heterologous biosynthetic pathways. The method we developed, known as RetroPath, uses a retrosynthetic approach in the reaction signature space [7]. Reaction signatures are biochemical reactions coded into the molecular signature representation and are used to search and enumerate similar reactions [8]. Our tool allows one to prioritize the engineering of the most promising routes due to its ranking function [7], which quantifies the enzyme performance and compatibility between the host and the candidate exogenous genes screened from a metabolic database like KEGG [9] or MetaCyc [10], the estimation of steady-state fluxes from the *in silico* reconstructed model of the engineered strain [11], and the estimation of metabolite toxicity [12]. This approach, hereafter named retrosynthetic biology [12], is used here as a streamlined manufacturing pipeline for the production of compounds with therapeutic interest. RetroPath provides for this purpose a methodology that is general enough to be applied to the production of virtually any compound through metabolic engineering. Our tool does not primarily seek to bypass any of the necessary optimization steps that would at later stages be required for scaling up the engineered strains to industrial production, but rather to provide the design techniques that would univocally identify, assemble and optimize constructs that eventually would lead to the streamlined cell factories.

In the present paper, we illustrate the use of Retropath for the production of flavonoids, which are compounds naturally produced in plants. Plant-derived natural products hold much promise for the development of new drugs and nutraceuticals [13, 14]. Modification of such compounds for drug development however, is still under-explored due to their low bioavailability and the difficulty of their chemical synthesis. Among these compounds with therapeutic interest are the flavonoids, which are plant secondary metabolites functioning as flower pigments, UV-protectants, insect repellants or initiators of symbiosis [15].

Transferring the flavonoid synthesis pathway into recombinant microorganisms has been successfully carried out to produce various members of these drug candidates [6, 16-20]. It is worth noting that none of the enzymes used in the aforementioned works were found fully annotated in the metabolic databases, the authors, experts in the field of flavonoid production, most probably selected them based on personal experience [16, 17]. Conversely, the present work aims to assemble biosynthetic pathways for production of a compound of interest in *Escherichia coli* by systematically applying the RetroPath algorithm on an enzyme set obtained from a metabolic knowledge database.

Among flavonoids, pinocembrin, extracted from propolis was reported to possess numerous biological activities beneficial to health such as neuroprotective, antioxidant or antibacterial effects [21], [22]. Moreover, the flavanone pinocembrin is a starting point for the synthesis of over 8,000 different biologically active molecules through the action of a myriad of enzymes, making it a key compound [23]. In this study, we use RetroPath to provide the most feasible route for microbial synthesis of pinocembrin and validate its productivity by experimentally implementing the top-ranked constructs.

2. Materials and methods

2.1. Materials

Escherichia coli strain DH5α (Life Technologies, Darmstadt, Germany) was used for cloning, enzyme expression and metabolite production. Plasmids pQlinkN [24] and pRSFDuet were obtained from Addgene and Merck-Novagen (Darmstadt, Germany), respectively. All chemicals were obtained from SIGMA (St. Louis, MO, USA), unless specified otherwise. Antibiotics were used at the following concentrations: ampicillin (Ap): 50 μg/ml, chloramphenicol (Cm): 25 μg/ml, kanamycin (Km): 30 μg/ml.

2.2. Pathway construction

Genes encoding the enzymes expressed in this study are listed in **Table 1**. Total RNA, extracted from fresh *Arabidopsis thaliana* ecotype Columbia leaves [25] was a kind gift of Dr. Bruno Gronenborn (ISV, Gif sur Yvette, France). Total RNA was reverse transcribed using RevertAid Premium First Strand cDNA Synthesis kit (Thermo Fisher Scientific Biosciences GmbH- Fermentas; Vilnius, Lithuania). Genomic DNA of *Streptomyces coelicolor* and *Bacillus subtilis* 168, kind gifts of Dr. Sylvie Lautru (IGM, Orsay, France) and Dr. Hamid Nouri (iSSB, Evry, France) respectively, were prepared using the GeneJet Genomic DNA Purification Kit (Thermo Fisher Scientific Biosciences GmbH - Fermentas; Vilnius, Lithuania). For the pinocembrin-producing pathway, PCR was used to amplify *hPAL, Pal2* and *hCHS* genes from the cDNA library of *A. thaliana*, as well as *h4CL* and *lCHS* from *S. coelicolor* and *B. subtilis* genomic DNA, respectively. The restriction enzymes used for gene cloning into pQLinkN can be inferred from the primer list (**Table S1**). *l4CLopt*, *m4CLopt*, *hCHIopt* and *lCHIopt* were codon-optimized for expression in *E. coli,* synthesized by GenScript (Piscataway, NJ, USA) and cloned into pQLinkN using BamHI and HindIII restriction/ligation. The genes were then combined into pQLinkN to make complete pinocembrin-producing pathways using ligation-independent cloning (LIC) as described earlier [24]. For the pathways producing malonyl-CoA, genes *mmsA,* and *cagg1256,* as well as the *atoDA* operon were codon-optimized and synthesized by GenScript, each with NdeI and XhoI sites flanking 5' and 3', respectively. Plasmid pACYC*matCmatB* [26] was a kind gift of Dr. Jingwen Zhou (LBBE, Jiangnan, China), and pETM6-M.*accABCD* , carrying the genes of the *E. coli* acetyl-CoA carboxylase complex in a monocystronic form [27] was a kind present from Prof. Mattheos Koffas (Rensselaer Polytechnic Institute, Troy, NY, USA). pRSF*matCmatB* was constructed by cloning the 3032 bp EcoRI-XhoI fragment of pACYC*matCmatB* into pRSFDuet. pRSFM.*accABCD* was constructed by ligating the 4824

bp SalI-AvrII fragment of pETM6-M.*accABCD* into SalI-AvrII-digested pRSFDuet. pRSF*mmsA* and pRSF*cagg1256* were made by cloning the respective NdeI-XhoI-restricted synthetic *mmsA* or *cagg1256* gene into the similarly digested pRSFDuet. pRSF*matCatoDA* was constructed by ligating the NdeI-XhoI-digested synthetic *atoDA* operon into the NdeI-XhoI cleaved pRSF*matCmatB*. Nucleotide sequences of the plasmids harboring the complete pathways were verified by sequencing. All DNA manipulation and culturing was carried out according to standard protocols [28].

2.3. Conditions of culture and protein overexpression

Bacterial cultures fully grown overnight in LB medium were diluted to an OD_{600} of 0.1 in Terrific Broth (TB) [28] supplemented with 0.5 M sorbitol, 5 mM betaine [29] and the adequate antibiotics, and were shaken at 37 ºC, 250 rpm. Protein expression was induced after three hours with 1 mM IPTG together with the addition of 3 mM phenylalanine, 2 mg/ml sodium malonate and when indicated, 20 μg/ml cerulenin. For strains harboring pRSF*mmsA* and pRSF*cagg1256*, malonate was replaced with 3 mM of β-alanine. Cultures were sampled 24 hours after induction.

2.4. Extraction of metabolites

Metabolites were extracted using a protocol adapted from Wu *et al.* [26]. Briefly, biomass was separated from the culture medium by centrifugation at 13,000 g and flavonoids were extracted from 1 mL of supernatant with an equivalent volume of pre-chilled ethyl acetate. Samples were vortexed and subsequently centrifuged for 15 min at 4 ºC, 13,000 g. The upper organic layer was then vacuum centrifuged for 2 hours at mid temperature. Dried pellets were stored at -80ºC until analysis by HPLC or LC-MS.

2.5. HPLC analysis

HPLC analysis was carried out using a Shimadzu Prominence LC20/SIL-20AC equipped with a Kinetex XB-C18 reversed phase column (250 x 4.5 mm, 5 μm) and a UV-Vis detector. Mobile phase was composed of 0.1 % formic acid in water (A) and 0.1 % formic acid in acetonitrile (B). A linear gradient elution using a binary pump was performed as follows: 0 - 10 min, 10 % B to 90 % B; 10 - 20 min, 90 % B to 10 % B; 20 – 30 min, 10 % B to 10 % B; 30 min, stop. Samples were thawed and resuspended in 200 μl of 80 % acetonitrile. The flow rate was 500 μl/min, the sample injection volume was 10 μl, and the column was thermostated at 40 °C. Pinocembrin and trans-cinnamate were monitored at 290 nm. Quantification of metabolites was done by interpolation of the integrated peak areas using a calibration curve prepared with standard samples.

2.6. LC-MS analysis

Formal identification of flavonoids was done using LC-MS. Samples were thawed and resuspended in 80 % acetonitrile and 20 % of 10 mM ammonium carbonate. Chromatographic separation was performed using a normal phase method on a SeQuant ZIC pHILIC HPLC column (150 x 4.6 mm, 5 μm polymeric beads PEEK) from Merck (Darmstadt, Germany). The flow rate was 500 μl/min, the sample injection volume was 10 μl, and the column was thermostated at 40 °C. The mobile phase A was 10 mM of ammonium carbonate in water, and the mobile phase B was acetonitrile. The starting conditions were 20 % of A and 80 % of B. The following gradient profile was used: 0 – 2 min 20 % A, 2 – 16 min 20 – 60 % A, 16 – 24 min 60 % A. The mobile phase was allowed to return to the starting conditions within 6 min, and the column was re-equilibrated for 15 min. The LTQ Orbitrap mass spectrometer was equipped with an ESI source operated in the negative ion mode. The ionization conditions were optimized for the detection of the compound of interest. The spray voltage was set to –

3.5 kV, the capillary voltage to – 47 V, tube lens offset to – 120 V, the sheath gas and auxiliary gas flow rates were 10 a.u. and 18 a.u., respectively.

2.7. RetroPath pathway design

Metabolic pathways were designed using the RetroPath tool, a design pipeline based on retrosynthesis [7, 8]. This software searches and enumerates all possible pathways producing the desired target compound in a host organism [30]. In the present study, the target compound is pinocembrin and the host cell is *E.coli*. Each pathway provided by RetroPath corresponds to a unique list of enzymatic reactions that potentially can lead to the production of the target compound starting from precursors available in the host organism or from added supplements in the medium. Pathways were ranked by RetroPath by defining a score function based on three criteria: enzyme performance, toxicity of intermediates, and pathway maximum yield [7]. The contribution of each term to the total score was computed as follows: (1) for each enumerated pathway P, the system provided a score $w(r)$ for the enzymes that can putatively catalyze each of the reaction steps r based on the prediction of enzyme performance through the tensor product technique [31, 32]. Such score was previously shown to parallel the ability of enzymes to catalyze multiple reactions, i.e. the prediction of enzyme's promiscuity [32], and it is therefore used here in order to evaluate the likeliness of an enzyme to catalyze a given reaction. (2) Toxicity score for each intermediate metabolite c produced by the reactions r in the pathway P was evaluated by using the EcoliTox server [33], with the associated score given by the sum of predicted log(IC50). (3) Flux balance analysis (FBA) was performed in order to compute the growth-target coupled flux, i.e. the product of maximum growth μ_{max} and maximum target compound flux v_{max} for each pathway P [8]. The associated score $S(P)$ of a given pathway P was defined by the weighted sum of the three previous terms:

$$S(P) = \lambda_{path} \sum_{r \in P} w(r) + \lambda_{tox} \sum_{r \in P} \sum_{c \in r} \log(IC50(c)) + \lambda_{flux}\, \mu_{\max} v_{\max}$$

where the weighting coefficients were set to (λ_{path} , λ_{tox} ,λ_{flux}) = (1.0, 0.4, 2.5) as in [8]. The way the weights were chosen was through an optimization algorithm that was previously described in [7].

2.8. Metabolic flux analysis

We modeled the effect of pathway insertion into metabolic fluxes by importing predicted RetroPath pathways into an *in silico* reconstructed model for *E. coli* [34] that was obtained from the BioModels database [35], which contains 2381 reactions and 1668 metabolites. More precisely, the pinocembrin and malonyl-CoA producing heterologous pathways were added to the model, as well as transport and exchange reactions for trans-cinnamate, pinocembrin, malonate, β-alanine and phenylalanine. FBA simulations were performed using the COBRApy package [36], where measured fluxes for key intermediates and final products as well as growth were fixed to their experimental values, as additional system constraints. The obtained fitted solutions were used in order to determine phenylalanine, acetyl-CoA, and malonate or β-alanine consumption, and the drains of trans-cinnamate and malonyl-CoA precursors going into biomass formation.

Growth was estimated from OD600 measurements based on the approximate conversion between biomass and OD600: biomass [gDW/L] ≅ 0.44 × OD (where DW stands for dry weight)[37]. We estimated growth rate (μ) as the difference between the measured OD and the OD_0 at the time of induction:

$$\mu = \frac{\ln(2)}{t_d} = \frac{\ln(OD/OD_0)}{\Delta t} \quad \left[hr^{-1}\right]$$

where Δt is the time elapsed after induction, and t_d is the doubling time.

Flux units were expressed in mmol gDW^{-1} hr^{-1}. Fluxes were estimated from measured concentrations [c] as follows:

$$v_c = \mu[hr^{-1}] \times \frac{[c]-[c]_0[mg/l]}{M(c)[g/mol^{-1}] \times [OD-OD_0] \times 0.44[gDW/l]} \quad [mmol\, gDW^{-1}\, hr^{-1}]$$

where $M(c)$ is the molar mass of the chemical compound, and $[c]_0$ is the initial concentration.

The inhibition effect on growth from trans-cinnamate was estimated using the EcoliTox server [33].

3. Results and discussion

3.1 Predicted pathways

Here, we use the term *pathway* to designate a series of enzymatic steps (defined by their EC numbers) connecting designated source metabolites to target compounds. Since usually more than one enzyme is available for each step, a pathway can be assembled using multiple combinations of enzymes, referred to as *constructs*. Under such definitions, we applied RetroPath to design and optimize the pathway and the associated constructs leading to the heterologous production of the flavonoid pinocembrin in *E. coli*. In total, RetroPath found and ranked eleven heterologous pathways connecting the endogenous metabolites of *E. coli* to pinocembrin. **Figure 1** depicts the map of these pathways, together with their connection to the metabolome of the chassis cell. The contribution of each score term (enzyme efficiency, metabolite toxicity, product flux) to the total pathway score is provided in the inset of **Figure 1**. The top ranked pathway (No. 1.1) uses phenylalanine as main precursor, which is converted into trans-cinnamate by phenylalanine/tyrosine ammonia lyase (PAL/TAL). The trans-cinnamate is subsequently transformed into cinnamoyl-CoA by 4-coumaroyl-CoA ligase

(4CL), and then into pinocembrin chalcone by a chalcone synthase (CHS) and further into pinocembrin by chalcone isomerase (CHI). This is the natural pathway of flavonoid synthesis, and has already been demonstrated to function appropriately in recombinant *E. coli* cells [5, 19]. However, within the natural pathway, there is still an impressive number of possible enzymes combinations (i.e. constructs) depending on which species the enzyme originated from. The choice of enzymes proposed by RetroPath is a critical feature that reduces the number of constructs one needs to implement to only those that have the best predicted performances.

For the top ranked pathway (No. 1.1), RetroPath compiled the list of genes available for each enzymatic step. To test all gene-combinations supporting this pathway (**Table S2A**), one would need to assemble more than 8.8 million constructs (110 x 112 x 45 x 16). We narrowed this list by using RetroPath (see Section 2.7), which ranked the genes for each reaction according to the expected performance of the encoded enzymes. The score of the enzymes taken into consideration for each step is shown in **Table S2B**. We privileged using the wild type alleles recovered from genomic DNA or cDNA libraries. If any of these failed to give a band on a Coomassie-stained polyacrylamide gel upon induction, it was codon-optimized and commercially synthesized for expression in *E. coli*, and was marked by an "opt" suffix in its tag. By default, we tested two alternative enzymes for each step. For PAL, we were in the fortunate situation that its product (trans-cinnamate) is directly measurable by HPLC. The two highest-scoring candidates *Arabidopsis thaliana*'s *Pal1* and *Pal2* produced 87.98 mg/L and 0.22 mg/L trans-cinnamate, respectively, when expressed individually. Therefore we decided to include only the prior in the final constructs, since it is the better trans-cinnamate producer. In addition, plant-derived 4CL enzymes have been described to have low cinnamoyl-CoA ligase activity [38]. We therefore included h4CL as a positive control, since it had already proven to be effective in this pathway [18]. h4CL is not present in KEGG, our initial enzyme

source, but exhibited a significantly high score when included in the list of queried enzymes. Altogether, genes for one enzyme displaying PAL activity, three displaying 4CL, two displaying CHS and two exhibiting CHI activities were cloned to make a full pathway (**Table 1**). To simplify their nomenclature, we differentiate them with the letters "h", "m" or "l" referring to their scores being "high", "medium" or "low". These enzymes can be assembled in twelve combinations, depicted in **Table S3**.

3.2 Assembly and analysis of pinocembrin-producing pathways

Based on the provided scores of the individually cloned genes, we set out to implement the highest-ranked pathway (No. 1.1) in the *E. coli* chassis. Genes were assembled into twelve possible constructs, i.e. into twelve possible combinations of enzymes corresponding to steps in the highest-ranked pathway, (**Table S3**) using the rapid method of LIC [24].

The performance of each construct was assessed experimentally by measuring the concentration of pinocembrin in the culture medium 24 h after induction, as described in the Methods section. For the top-ranked constructs HHHH and HHHL we confirmed the effective production of pinocembrin at low titers using LC-MS analysis (**Figure S1**, **Table S4**). However, we observed that trans-cinnamate, the first intermediate in the pathway, was produced by both strains in significant amounts (38.2 and 53.6 mg/L for HHHH and HHHL, respectively) compared to the negative control (strain with induced empty plasmid, DH5α + pQlinkN in **Table S4**). Therefore, the trans-cinnamate produced from phenylalanine by overexpression of *hPAL*, was not efficiently being consumed to produce pinocembrin and was accumulated in the medium.

3.3 Network-level optimization of pinocembrin production

Flux balance analysis (FBA) provided a possible explanation for the low production of pinocembrin obtained in the initial tests, and guided us towards pathway optimization using RetroPath. The analysis was performed on an *E. coli in silico* model constrained with experimental measurements for trans-cinnamate, pinocembrin and biomass. **Figure S2** shows the main fluxes involved in the pathway. The model predicted that under the observed conditions of low pinocembrin yield, most of trans-cinnamate was consumed for growth, while the rest was accumulated in the medium because the excess of trans-cinnamate was not compensated by the availability of malonyl-CoA, mostly consumed in the competitive pathways of fatty acids synthesis. Therefore, in order to boost the production of pinocembrin, the available pool of malonyl-CoA needed to be increased. We thus returned to the design task of our metabolic engineering pipeline and used Retropath to enumerate all pathways enabling the production of malonyl-CoA. RetroPath ranked four heterologous pathways producing malonyl-CoA (**Figure 2**), which are as follows:

2.1. Malonyl-CoA production from malonate through malonyl-CoA synthase (EC 6.2.1.-) using CoA and ATP as cofactors (pathway score 13.01). This reaction is not annotated for *E. coli*, but the product of the *matB* gene [39], originating from *Rhizobium trifolii* has been applied successfully for this purpose [4, 26].

2.2. Acetyl-CoA carboxylase (EC 6.4.1.2, pathway score 12.40), producing malonyl-CoA from acetyl-CoA and bicarbonate. In *E. coli*, a four-enzyme complex comprising carboxyltransferase α, biotin-carboxyl-carrier-protein, biotin carboxylase and carboxyltransferase β is responsible for this step. The endogenous complex [40], a similar complex of *Photorabdus luminescens* [16], as well as a two-enzyme complex from *Corynebacterium glutamicum* [5] have all been successfully expressed in *E. coli*.

2.3. Malonate-CoA transferase (EC 2.8.3.3). This enzyme can also produce malonyl-CoA from malonate but requires acetyl-CoA as a co-substrate (pathway score 12.77). This reaction is not annotated for *E. coli*.

2.4. Malonyl-CoA reductase (EC 1.2.1.75, pathway score 12.97). It requires 3-oxopropanoate, CoA and $NADP^+$ as substrates. This reaction is not annotated for *E. coli*, either.

These pathways, integrated with pathway 1.1 are summarized in **Table 2**. verify the pathway-predicting function of RetroPath, we implemented all four malonyl-CoA producer pathways using the top-scoring genes available for each (**Table S5**). For reaction 1.2.1.75, besides testing *mmsA*, we also implemented *cagg1256*, which was discovered by running the query after a recent update of the KEGG-database. The transmembrane dicarboxylate carrier protein, encoded by *matC* was co-expressed with *matB* and *atoDA* to grant sufficient uptake of malonate by the cells. In order to provide the substrate for *mmsA* and *cagg1256,* cultures were supplemented with β-alanine, which is converted to 3-oxopropanoate (malonate semialdehyde) by the cell's endogenous 4-aminobutyrate aminotransferase. The absolute pinocembrin titers achieved with HHHH together with these supplementary pathways are shown on **Figure S3**. It is apparent that pathway 2.1, which has the highest score yielded the greatest titer (4.22 mg/L). The relation of pinocembrin production efficiencies to the predicted scores (**Figure 3D**) further supports our predictions.

In order to divert further the malonyl-CoA excess into the flavonoid pathway, cerulenin was added into the medium to inhibit fatty acid formation [4, 41]. The inhibitory effect of cerulenin on the production of fatty acids was estimated again by flux analysis (**Figure S2**) leading to a lower consumption of malonyl-CoA for fatty acids biosynthesis, which in turn led to an increase in the flux for production of pinocembrin and a larger accumulation of trans-cinnamate. By combining the best performing malonyl-CoA producer (*matCmatB* cassette)

with the cerulenin inhibition, measurements were carried out for all twelve pinocembrin production constructs, obtaining in this case higher pinocembrin titers (**Table S7**), which were routinely detectable thereafter using HPLC (**Figure S4**). Estimated flux values are provided in **Table S6**.

3.4. Validation of the predictions and evaluation of results

The accuracy of enzyme-score predictions was verified using the concentrations of the intermediate metabolite trans-cinnamate, and the target compound pinocembrin. In the case of using trans-cinnamate levels for validation, the PAL score is taken as it is, since it is a contributor, while the 4CL score is subtracted, since it is a consumer of trans-cinnamate. In the two cases where pinocembrin was produced, we also subtracted the scores for CHS and CHI as these enzymes also decrease the levels of trans-cinnamate. We obtain a good agreement between the RetroPath scores combined this way and the experimentally measured trans-cinnamate levels (**Figure 3A)**. Specific trans-cinnamate levels for each construct are given in **Table S7**. The constructs containing the *l4CLopt* gene from *S. maritimus* (H**L**HH, H**L**HL, H**L**LH, H**L**LL), which had a lower score correspond to a higher accumulation of trans-cinnamate, which might indicate that less trans-cinnamate was transformed. The constructs that contained the 4CL gene with the highest score, *h4CL* from *S. coelicolor* (H**H**HH, H**H**HL, H**H**LH, H**H**LL) led to less trans-cinnamate accumulation indicating its more efficient transformation, although they showed more variability.

Among the genes chosen for the pinocembrin producing constructs, only two, *l4CLopt* and *lCHS* were found to lack activity, narrowing the number of successful constructs to four (HHHH, HHHL, HMHH, HMHL). For proper comparison of these constructs, we normalized the pinocembrin production titers with the biomass and the fermentation time (**Figure 3B**) to obtain a parameter we refer to as efficiency. Our analysis revealed HHHH to be the most

efficient producer, supporting the construct-scores predicted by RetroPath. In addition, we found that for each enzymatic step, inserting an enzyme with a higher score led to a construct with a more efficient performance (**Figure 3C**). It is apparent from **Figure 3C**, that the impact of 4CL score on construct efficiency is higher than that of CHS or CHI. Additionally, the scores of the CHI enzymes did not provide further improvement in predicting pinocembrin production, since the choice of the gene for this last step did not significantly alter construct efficiency (**Table S7**). This is due most likely to the fact that only 16 enzyme sequences were available in databases in order to build the training set (accuracy and specificity < 70 %). This issue of low performance for intramolecular lyase activity (EC 5.5) prediction has been previously reported by several groups using different learning techniques [31, 42, 43].

From the industrial point of view, the maximal production titer may in certain cases be of greater importance than the efficiency normalized by the biomass. In this aspect, the HHHL, and not the HHHH construct, producing extracellular pinocembrin levels of 24.14 mg/L (±5.47 mg/L) proved to be the best (**Table S7**), and was comparable to other metabolic engineering projects targeting flavanones (**Table S9)**. The reason for this discrepancy among our top-producers is the fact that a uniform protocol of culturing and induction was used for all of our strains. The strains carrying the HHHH construct, due to their slower growth, would most probably need longer fermentation times to generate a proper biomass prior to induction to surpass HHHL in absolute pinocembrin titers. Such an optimization is not included in RetroPath in its current version, nor is the optimization of expression levels of the inserted heterologous enzymes.

In the current work Retropath was nevertheless used in a round of network-level optimization to select the most efficient pathway to boost malonyl-CoA production, and eliminate the bottleneck of the CHS step. Implementing the four malonyl-CoA producer pathways, and

comparing their effect on pinocembrin production efficiency provided a good validation of the pathway-ranking function (**Figure 3D**). This result highlights the ability of the predicted scores to rank pathways even in cases where predicted scores are relatively close in terms of magnitude (score differences of approximately 2%). Based on the validation results from Figures 3C and 3D, we would recommend considering pathway scores as significantly different if they differ in at least 2% of the averaged score per gene, which is computed by dividing the total score by the number of enzymatic steps. Only pathway 2.8.3.3 turned out to be an outlier, with *atoDA* performing much weaker than expected. Since this was the only putative pathway among the four alternatives, it may indicate the need of further optimization on the pathway scoring function by introducing a penalty for putative steps. Malonyl-CoA, the compound targeted in the optimization process, was chosen based on its position on the metabolic map. In the future, this choice could also be automated by incorporating existing methods such as OptKnock [44] or Redirector [45] into our pipeline.

4. Concluding remarks

Our results underline several factors that future users of RetroPath must be aware of, when planning to experimentally investigate their predictions. First, due to the nature of automated pipelines, false positive hits or score differences lying below the threshold of significance for proper ranking are probably inevitable in some cases. Second, a certain percentage of the predicted candidates are "lost" due to problematic cloning or expression. The continuously decreasing price of gene synthesis will aid the evasion of the latter caveat, multiple candidates for each enzymatic step should nevertheless be considered. Third, several enzymes proven to be effective in this pathway were missing or not annotated in metabolic knowledge bases. Besides emphasizing the need to update these databases, this information highlights the necessity of integrating the data from multiple resources. And fourth, classical metabolic

engineering experiments, i.e. optimizing gene expression, culture conditions and extraction protocols will still be required to fully exploit the capacities of the top-ranked pathways.

Despite these difficulties, the goal of assembling a pathway to produce a target compound using retrosynthetic biology was nevertheless successfully reached. RetroPath predicted and ranked eleven pathways linking endogenous *E. coli* metabolites to pinocembrin. For the four steps in the top-ranked pathway, our CAD application guided us in order to select eight enzymes, six of which exhibited the required activities, therefore four out of the twelve resulting constructs produced significant amounts of pinocembrin. We obtained the best production efficiency with the construct consisting of the highest-scoring enzymes, and found a good correlation between enzyme scores and their contribution to construct performance. These results validated the enzyme ranking function of RetroPath, and allowed a small number of constructs to be tested to find target-producing hits, despite the initial pool of 9 million enzyme combinations. The top ranking constructs proved to be the best producers of pinocembrin. The pathway ranking function was also validated by implementing four pathways for malonyl-CoA production, three of which performed according to our predictions. In total 12 out of the 13 enzymes (2 PAL, 3 4CL, 2 CHS, 2CHI, and 4 for malonyl-CoA synthesis) tested in this work displayed a performance that was in accordance with its predicted score (Figures 3ACD). Overall, choosing the highest-ranking pathway and the top-scoring enzyme available for each step would have led to the best performing strain in the first place, which in our opinion is a strong indicator of the value of our predictions for the metabolic engineering community.

The support that our tool provided in pathway and construct ranking significantly shortened the design phase and alleviated the need for either expert knowledge concerning the engineered pathway or numerous trial-and-error experiments. We believe therefore that its adoption would substantially accelerate projects targeting the biosynthesis of other compounds beyond our tested flavonoid pathways. Furthermore, a remarkable feature of our method is that it considers and screens for promiscuous enzymatic activities when predicting novel pathways, a capability that notably allows its use to engineer pathways that did not exist before in nature; opening in that way the possibility of producing non-natural compounds as a result [46, 47]. In addition, RetroPath could be used to select the chassis cell itself for more effective production of natural or non-natural compounds. To that end, our forthcoming aim is to validate and explore further the advanced capabilities of RetroPath, building upon the promising results showcased in the present work, in order to extend its applicability into an ever-wider range of metabolic engineering and synthetic biology projects.

Acknowledgements

We thank Elodie Paillard and Fred Green (iSSB, Evry, France) for technical assistance, as well as Cyrille Pauthenier (iSSB, Evry, France) for helpful discussions. This project was financed by the ATIGE and the ANR Chair of Excellence grants.

Tables

Table 1. Top scoring genes used for pathway No. 1.1 construction and their associated *p*-values.

Gene tag	Encoded enzymatic activity	Source	Accession number	RetroPath Score
hPAL	Phe ammonia-lyase	*Arabidopsis thaliana*	TAIR: AT2G37040	2.39
h4CL	4-coumarate:CoA ligase	*Streptomyces coelicolor*	EMBL: CAB95894.1	1.39
l4CLopt	4-coumarate:CoA ligase	*Streptomyces maritimus*	UniProt: Q9KHL1	1.07
m4CLopt	4-coumarate:CoA ligase	*Arabidopsis thaliana*	TAIR: AT1G65060.1	1.08
hCHS	chalcone synthase	*Arabidopsis thaliana*	TAIR: AT5G13930	0.96
lCHS	chalcone synthase	*Bacillus subtilis*	SubtiList: BSU22050	0.75
hCHIopt	chalcone isomerase	*Arabidopsis thaliana*	TAIR: AT3G55120	0.69
lCHIopt	chalcone isomerase	*Arabidopsis thaliana*	TAIR: AT5G66220	-0.07

Table 2. Ranking of the pinocembrin-producing pathways derived from pathway 1.1 (Table 1) with an additional enzyme to increase production of the malonyl-CoA precursor.

Rank	No.	Pathway	Enzyme score	Toxicity score log(IC50)	Coupled flux ($g/gDWh^2$)	Total score

1	2.1	Malonate + CoA +ATP→ Malonyl-CoA L-Phenylalanine → trans-Cinnamate → Cinnamoyl-CoA → Pinocembrin chalcone → Pinocembrin	1.46	4.98	2.03	**13.01**
2	2.2	3-Oxopropanoate + CoA + NADP+ → Malonyl-CoA L-Phenylalanine → trans-Cinnamate → Cinnamoyl-CoA → Pinocembrin chalcone → Pinocembrin	1.19	4.98	2.12	**12.97**
3	2.3	Malonate + Acetyl-CoA → Malonyl-CoA L-Phenylalanine → trans-Cinnamate → Cinnamoyl-CoA → Pinocembrin chalcone → Pinocembrin	1.22	4.98	2.03	**12.77**
4	2.4	CCCP + Acetyl-CoA → Malonyl-CoA L-Phenylalanine → trans-Cinnamate → Cinnamoyl-CoA → Pinocembrin chalcone → Pinocembrin	1.31	4.98	1.84	**12.40**

Figure legends

Figure 1. Pathways enumerated and ranked by RetroPath leading to production of pinocembrin in *E. coli*. Solid circles correspond to enzymes with documented activity, while dotted circles correspond to enzymes with putative activity predicted by RetroPath. Endogenous *E. coli* metabolites are displayed in grey boxes. The inset table shows the ranking of pinocembrin pathways based on a score function consisting of three terms: enzyme efficiency, metabolite toxicity, and calculated growth-target coupled flux.

Figure 2. Pathways enumerated by RetroPath for the production of malonyl-CoA. Four 1-step pathways are possible: EC 6.2.1.- and EC 2.8.3.3, which consume malonate added as a supplement; as well as EC 6.4.1.2 and EC 1.2.1.75. Solid circles correspond to enzymes with documented activity, while dotted circles correspond to enzymes with putative activity, predicted by RetroPath. Endogenous and heterologous *E. coli* metabolites are displayed in grey and yellow boxes respectively.

Figure 3. (A) The accumulation of trans-cinnamate in the medium versus the calculated RetroPath scores. A higher score means higher predicted trans-cinnamate levels, as described

in the text. Cerulenin was administered in all cases. **(B)** The efficiency of pinocembrin production for the four strains carrying the successful constructs. The *matCmatB* cassette was present in each strain, and cerulenin was administered in all cases. **(C)** Effect of individual enzyme scores on construct efficiencies. Variation of average efficiencies and average enzyme scores associated with enzyme change at various positions of the pathway. The letter preceding the gene names is "l" for low, "m" for medium and "h" for high predicted activities, as described in Section 3.1. The *matCmatB* cassette was present in each strain, and cerulenin was administered in all cases. (**D**) Validation of the pathway ranking function. Measured pinocembrin production efficiencies are shown *vs.* the predicted scores of pathways listed in Table 2. In each case, the HHHH construct was present in the cell, with the appropriate pathway for malonyl-CoA production present on a second, pRSFDuet-based plasmid. The scores, as well as the efficiencies for *mmsA* and *cagg1256* were averaged, since both correspond to the same pathway. No cerulenin was added in this experiment.

5. References

[1] Curran, K. A., Alper, H. S., Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metab. Eng.* 2012, *14*, 289-297.
[2] Shin, J. H., Kim, H. U., Kim, D. I., Lee, S. Y., Production of bulk chemicals via novel metabolic pathways in microorganisms. *Biotechnol. Adv.* 2012, *31*, 925-935.
[3] Copeland, W. B., Bartley, B. A., Chandran, D., Galdzicki, M.*, et al.*, Computational tools for metabolic engineering. *Metab. Eng.* 2012, *14*, 270-280.
[4] Leonard, E., Yan, Y., Fowler, Z. L., Li, Z.*, et al.*, Strain improvement of recombinant *Escherichia coli* for efficient production of plant flavonoids. *Mol. Pharm.* 2008, *5*, 257-265.
[5] Miyahisa, I., Kaneko, M., Funa, N., Kawasaki, H.*, et al.*, Efficient production of (2*S*)-flavanones by *Escherichia coli* containing an artificial biosynthetic gene cluster. *Appl. Microbiol. Biotechnol.* 2005, *68*, 498-504.
[6] Santos, C. N., Koffas, M., Stephanopoulos, G., Optimization of a heterologous pathway for the production of flavonoids from glucose. *Metab. Eng.* 2011, *13*, 392-400.
[7] Carbonell, P., Planson, A. G., Fichera, D., Faulon, J. L., A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.* 2011, *5*, 122.
[8] Carbonell, P., Planson, A. G., Faulon, J. L., Retrosynthetic design of heterologous pathways. *Methods Mol. Biol.* 2013, *985*, 149-173.

[9] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012, *40*, D109-114.
[10] Caspi, R., Altman, T., Dreher, K., Fulcher, C. A.*, et al.*, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2012, *40*, D742-753.
[11] Curran, K. A., Crook, N. C., Alper, H. S., Using flux balance analysis to guide microbial metabolic engineering. *Methods Mol. Biol.* 2012, *834*, 197-216.
[12] Planson, A. G., Carbonell, P., Grigoras, I., Faulon, J. L., A retrosynthetic biology approach to therapeutics: from conception to delivery. *Curr. Opin. Biotechnol.* 2012, *23*, 948-956.
[13] Chin, Y. W., Balunas, M. J., Chai, H. B., Kinghorn, A. D., Drug discovery from natural sources. *AAPS J.* 2006, *8*, E239-253.
[14] Cragg, G. M., Newman, D. J., Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* 2013, *1830*, 3670-3695.
[15] Dixon, R. A., Lamb, C. J., Masoud, S., Sewalt, V. J., Paiva, N. L., Metabolic engineering: prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses--a review. *Gene* 1996, *179*, 61-71.
[16] Leonard, E., Lim, K. H., Saw, P. N., Koffas, M. A., Engineering central metabolic pathways for high-level flavonoid production in *Escherichia coli*. *Appl. Environ. Microbiol.* 2007, *73*, 3877-3886.
[17] Miyahisa, I., Funa, N., Ohnishi, Y., Martens, S.*, et al.*, Combinatorial biosynthesis of flavones and flavonols in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* 2006, *71*, 53-58.
[18] Hwang, E. I., Kaneko, M., Ohnishi, Y., Horinouchi, S., Production of plant-specific flavanones by *Escherichia coli* containing an artificial gene cluster. *Appl. Environ. Microbiol.* 2003, *69*, 2699-2706.
[19] Kaneko, M., Hwang, E. I., Ohnishi, Y., Horinouchi, S., Heterologous production of flavanones in *Escherichia coli*: potential for combinatorial biosynthesis of flavonoids in bacteria. *J. Ind. Microbiol. Biotechnol.* 2003, *30*, 456-461.
[20] Watts, K. T., Lee, P. C., Schmidt-Dannert, C., Exploring recombinant flavonoid biosynthesis in metabolically engineered *Escherichia coli*. *ChemBioChem* 2004, *5*, 500-507.
[21] Liu, R., Wu, C. X., Zhou, D., Yang, F.*, et al.*, Pinocembrin protects against beta-amyloid-induced toxicity in neurons through inhibiting receptor for advanced glycation end products (RAGE)-independent signaling pathways and regulating mitochondrion-mediated apoptosis. *BMC Med.* 2012, *10*, 105.
[22] Soromou, L. W., Zhang, Y., Cui, Y., Wei, M.*, et al.*, Subinhibitory concentrations of pinocembrin exert anti-*Staphylococcus aureus* activity by reducing alpha-toxin expression. *J. Appl. Microbiol.* 2013, *115*, 41-49.
[23] Fowler, Z. L., Koffas, M. A., Biosynthesis and biotechnological production of flavanones: current state and perspectives. *Appl. Microbiol. Biotechnol.* 2009, *83*, 799-808.
[24] Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U., Bussow, K., Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.* 2007, *35*, e43.
[25] Grigoras, I., Timchenko, T., Gronenborn, B., Transcripts encoding the nanovirus master replication initiator proteins are terminally redundant. *J. Gen. Virol.* 2008, *89*, 583-593.
[26] Wu, J., Du, G., Zhou, J., Chen, J., Metabolic engineering of *Escherichia coli* for (2S)-pinocembrin production from glucose by a modular metabolic strategy. *Metab. Eng.* 2013, *16*, 48-55.

[27] Xu, P., Vansiri, A., Bhan, N., Koffas, M. A., ePathBrick: a synthetic biology platform for engineering metabolic pathways in E. coli. *ACS Synth. Biol.* 2014, *1*, 256-266.
[28] Sambrook, J., Fritch, E. F., Maniatis, T., *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York 1987.
[29] Blackwell, J. R., Horgan, R., A novel strategy for production of a highly expressed recombinant protein in an active form. *FEBS Lett.* 1991, *295*, 10-12.
[30] Carbonell, P., Fichera, D., Pandit, S. B., Faulon, J. L., Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst. Biol.* 2012, *6*, 10.
[31] Faulon, J. L., 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*, 225-233.
[32] Carbonell, P., Faulon, J. L., Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* 2010, *26*, 2012-2019.
[33] Planson, A. G., Carbonell, P., Paillard, E., Pollet, N., Faulon, J. L., Compound toxicity screening and structure-activity relationship modeling in *Escherichia coli*. *Biotechnol. Bioeng.* 2012, *109*, 846-850.
[34] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M.*, et al.*, A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 2007, *3*, 121.
[35] Li, C., Donizelli, M., Rodriguez, N., Dharuri, H.*, et al.*, BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* 2010, *4*, 92.
[36] Ebrahim, A., Lerman, J. A., Palsson, B. O., Hyduke, D. R., COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* 2013, *7*, 74.
[37] Conrad, T. M., Frazier, M., Joyce, A. R., Cho, B. K.*, et al.*, RNA polymerase mutants found through adaptive evolution reprogram *Escherichia coli* for optimal growth in minimal media. *Proc. Natl. Acad. Sci. U S A* 2010, *107*, 20500-20505.
[38] Kaneko, M., Ohnishi, Y., Horinouchi, S., Cinnamate:coenzyme A ligase from the filamentous bacterium *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 2003, *185*, 20-27.
[39] An, J. H., Lee, G. Y., Jung, J. W., Lee, W., Kim, Y. S., Identification of residues essential for a two-step reaction by malonyl-CoA synthetase from *Rhizobium trifolii*. *Biochem. J.* 1999, *344 Pt 1*, 159-166.
[40] Davis, M. S., Solbiati, J., Cronan, J. E., Jr., Overproduction of acetyl-CoA carboxylase activity increases the rate of fatty acid biosynthesis in *Escherichia coli*. *J. Biol. Chem.* 2000, *275*, 28593-28598.
[41] Heath, R. J., Rock, C. O., Regulation of malonyl-CoA metabolism by acyl-acyl carrier protein and beta-ketoacyl-acyl carrier protein synthases in *Escherichia coli*. *J. Biol. Chem.* 1995, *270*, 15531-15538.
[42] Shen, H. B., Chou, K. C., EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.* 2007, *364*, 53-59.
[43] Matsuta, Y., Ito, M., Tohsato, Y., ECOH: an enzyme commission number predictor using mutual information and a support vector machine. *Bioinformatics* 2013, *29*, 365-372.
[44] Burgard, A. P., Pharkya, P., Maranas, C. D., Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 2003, *84*, 647-657.
[45] Rockwell, G., Guido, N. J., Church, G. M., Redirector: designing cell factories by reconstructing the metabolic objective. *PLoS Comput Biol* 2013, *9*, e1002882.

[46] Carbonell, P., Parutto, P., Baudier, C., Junot, C., Faulon, J. L., Retropath: Automated Pipeline for Embedded Metabolic Circuits. *ACS Synth. Biol.* 2013, *10.1021/sb4001273 [doi]*.
[47] Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B., Faulon, J. L., XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.* 2014, *DOI:10.1093/nar/gku362*.

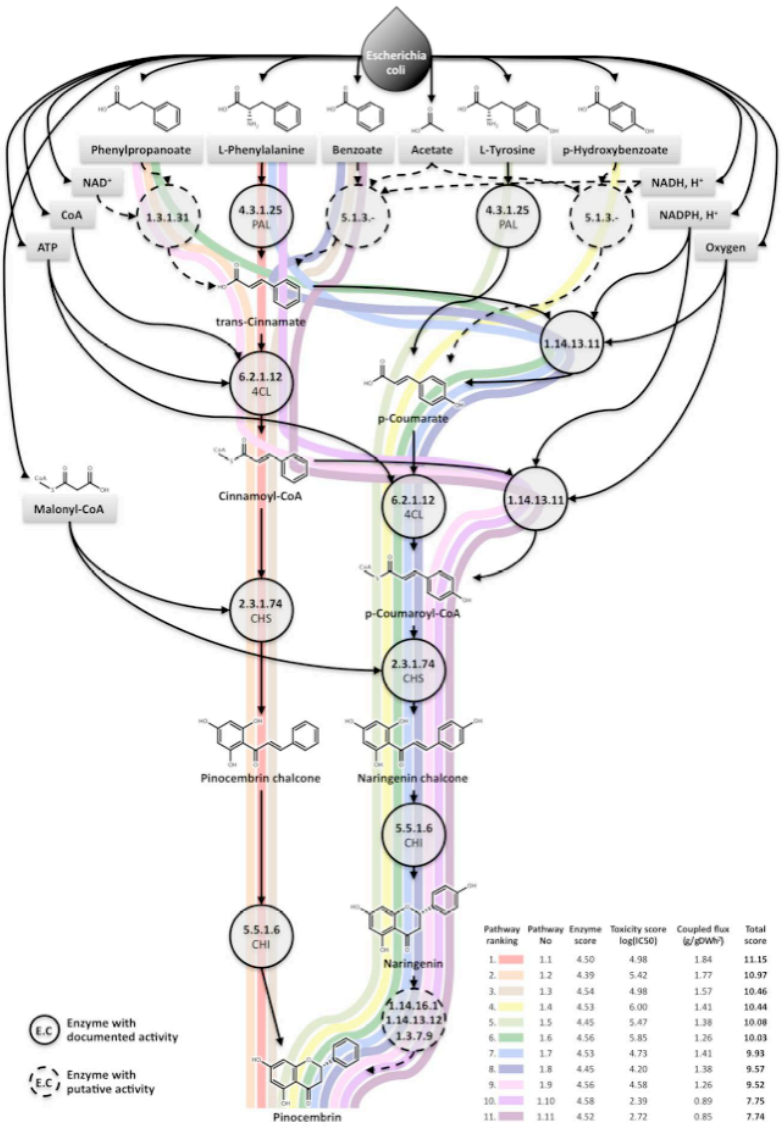

Figure 1. Pathways enumerated and ranked by RetroPath leading to production of pinocembrin in E. coli. Solid circles correspond to enzymes with documented activity, while dotted circles correspond to enzymes with putative activity predicted by RetroPath. Endogenous E. coli metabolites are displayed in grey boxes. The inset table shows the ranking of pinocembrin pathways based on a score function consisting of three terms: enzyme efficiency, metabolite toxicity, and calculated growth-target coupled flux.
549x793mm (72 x 72 DPI)

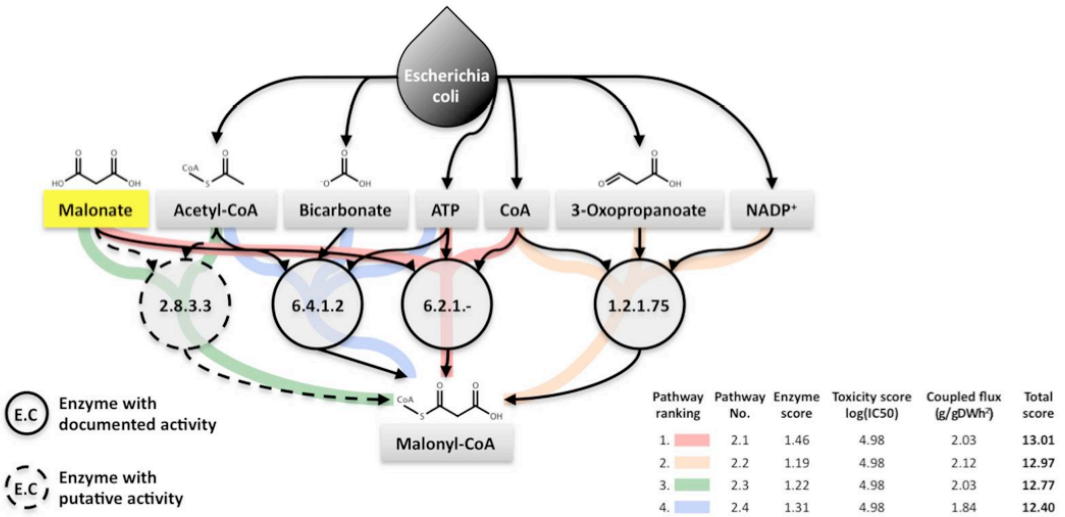

Figure 2. Pathways enumerated by RetroPath for the production of malonyl-CoA. Four 1-step pathways are possible: EC 6.2.1.- and EC 2.8.3.3, which consume malonate added as a supplement; as well as EC 6.4.1.2 and EC 1.2.1.75. Solid circles correspond to enzymes with documented activity, while dotted circles correspond to enzymes with putative activity, predicted by RetroPath. Endogenous and heterologous E. coli metabolites are displayed in grey and yellow boxes respectively.
522x255mm (72 x 72 DPI)

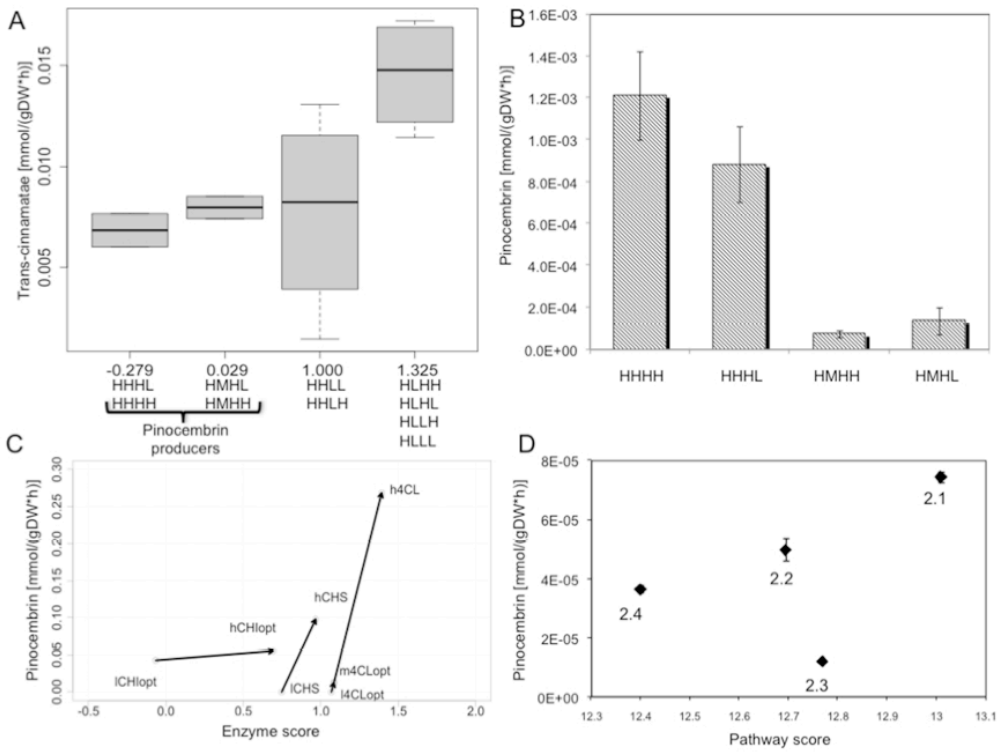

Figure 3. (A) The accumulation of trans-cinnamate in the medium versus the calculated RetroPath scores. A higher score means higher predicted trans-cinnamate levels, as described in the text. Cerulenin was administered in all cases. (B) The efficiency of pinocembrin production for the four strains carrying the successful constructs. The matCmatB cassette was present in each strain, and cerulenin was administered in all cases. (C) Effect of individual enzyme scores on construct efficiencies. Variation of average efficiencies and average enzyme scores associated with enzyme change at various positions of the pathway. The letter preceding the gene names is "l" for low, "m" for medium and "h" for high predicted activities, as described in Section 3.1. The *matCmatB* cassette was present in each strain, and cerulenin was administered in all cases. (D) Validation of the pathway ranking function. Measured pinocembrin production efficiencies are shown vs. the predicted scores of pathways listed in Table 2. In each case, the HHHH construct was present in the cell, with the appropriate pathway for malonyl-CoA production present on a second, pRSFDuet-based plasmid. The scores, as well as the efficiencies for *mmsA* and *cagg1256* were averaged, since both correspond to pathway 2.2. No cerulenin was added in this experiment. Pathway numbering corresponds to that of Table 2.

352x264mm (72 x 72 DPI)

Supplementary Material

Validation of RetroPath, a Computer Aided Design Tool for Metabolic Pathway Engineering

Tamás Fehér, Anne-Gaëlle Planson, Pablo Carbonell, Alfred Fernández-Castané, Ioana Grigoras, Ekaterina Dariy, Alain Perret, Jean-Loup Faulon

Table S1. Primers used for PCR amplification and cloning

Primer name	*Sequence*
P1-EcoRI-AT2G37040 (hPAL)	CGCGAATTCATGGAGATTAACGGGGCACACAAG
P2-NotI-AT2G37040 (hPAL)	AGAGCGGCCGCTTAACATATTGGAATGGGAGCTCC
P1-EcoRI-AT3G10340 (Pal2)	CGCGAATTCATGGAGCTATGCAATCAAAACAATCACATCACCGCCG
P2-NotI-AT3G10340 (Pal2)	ATAGCGGCCGCTCAACAGATTGAAACCGGAGCTCCGTTCC
P1-BamHI-ScCCL (h4CL)	TCTGGATCCATGTTCCGCAGCGAGTACGCAGACGTCC
P2-HindIII-ScCCL (h4CL)	GAGAAGCTTTTATCGCGGCTCCCTGAGCTGTCGGCG
P1-BamHI-AT5G13930 (hCHS)	CGCGGATCCATGGTGATGGCTGGTGCTTCTTCTTTGG
P2-NotI-AT5G13930 (hCHS)	AGAGCGGCCGCTTAGAGAGGAACGCTGTGCAAGACGAC
P1-EcoRI-BSU22050 (lCHS)	CGCGAATTCATGGCGTTTATTTTATCCATTGG
P2-BamHI-BSU22050 (lCHS)	CGCGGATCCTCAGGCCCCCTTTTCCCAGCTGA

Table S2A. Number of genes available in metabolic knowledge databases for various steps of pinocembrin production

Enzymatic step	*KEGG Database hits*	*MetaCyc Database hits*
PAL	110	3
4CL	112	5
CHS	45	7
CHI	16	2
Total combinations:	**8870400**	**210**

Table S2B. Genes we attempted to clone and express for pinocembrin production (black), and genes used for pinocembrin production in the literature (green)

Gene tag	*Source*	*Gene name*	*Accession number*	*Experience or source publication*	*Score*
hPAL	*Arabidopsis thaliana*	Pal1	TAIR: AT2G37040	low expression	2.392528819
	Arabidopsis thaliana	Pal4	TAIR: AT3G10340	no expression	2.268511
	Arabidopsis thaliana	Pal2	TAIR: AT3G53260	high expression	2.189715194
	Arabidopsis thaliana	Pal3	TAIR: AT5G04230	low expression	2.047882289
	Rhodotorula glutinis	Pal	UniProt: Q2VMT1	(Wu et al. 2013)	2.80016536
	Rhodotorula mucilaginosa	Pal	UniProt: P10248	(Kaneko et al. 2003a)	2.48978283
	Arabidopsis	4CL5	TAIR: AT1G62940	no cloning	1.05617895

	thaliana				
	Arabidopsis thaliana	4CL2	TAIR: AT1G20480	low expression	1.14922729
	Arabidopsis thaliana	4CL4	TAIR: AT3G21230	no expression	1.05409046
m4CL (opt)	*Arabidopsis thaliana*	4CL3	TAIR: AT1G65060	without codon optimization for *E. coli*: no expression with codon optimization for *E. coli*: expression	1.08424057
h4CL	*Streptomyces coelicolor*	4CL-2	GenBank: CAB95894	expression	1.39236807
l4CL	*Streptomyces maritimus*	encH	GenBank: AAF81723	codon optimization for *E. coli:* expression	1.06787189
	Petroselinum crispum	4CL1	UniProt: P14912	(Wu et al. 2013, Leonard et al. 2007)	1.12678499
lCHS	*Bacillus subtilis*	bcsA	SubtiList: BSU22050	high expression	0.747079642
hCHS	*Arabidopsis thaliana*	TT4	TAIR: AT5G13930	high expression	0.96642208
	Petunia hybrida	chs	UniProt: Q9M5B2	(Wu et al. 2013, Leonard et al. 2007)	1.45589774
	Glycyrrhiza inflata	chs	UniProt: C4MJ52	(Kaneko et al. 2003a)	0.87049822
lCHI (opt)	*Arabidopsis thaliana*	CHI2	TAIR: AT5G66220	without codon optimization for *E. coli*: no cloning with codon optimization for *E. coli*: expression	-0.065805884
hCHI (opt)	*Arabidopsis thaliana*	CHI1	TAIR: AT3G55120	without codon optimization for *E. coli*: no expression with codon optimization for *E. coli*: expression	0.690604899
	Medicago sativa	CHI2	UniProt: P28013	(Wu et al. 2013)	-0.3248541
	Petunia hybrida	CHI-A	UniProt: P11650	(Leonard et al. 2007)	-0.21657359

Table S3. Plasmids carrying the assembled constructs. Construct tags are given by the initial letters of each enzyme.

Construct tag	Plasmid
HHHH	pQlinkN_*hPAL_h4CL_hCHS_hCHIopt*
HHHL	pQlinkN_*hPAL_h4CL_hCHS_lCHIopt*
HHLH	pQlinkN_*hPAL_h4CL_lCHS_hCHIopt*
HHLL	pQlinkN_*hPAL_h4CL_lCHS_lCHIopt*
HLHH	pQlinkN_*hPAL_l4CLopt_hCHS_hCHIopt*
HLHL	pQlinkN_*hPAL_l4CLopt_hCHS_lCHIopt*
HLLH	pQlinkN_*hPAL_l4CLopt_lCHS_hCHIopt*
HLLL	pQlinkN_*hPAL_l4CLopt_lCHS_lCHIopt*
HMHH	pQlinkN_*hPAL_m4CLopt_hCHS_hCHIopt*
HMHL	pQlinkN_*hPAL_m4CLopt_hCHS_lCHIopt*
HMLH	pQlinkN_*hPAL_m4CLopt_lCHS_hCHIopt*
HMLL	pQlinkN_*hPAL_m4CLopt_lCHS_lCHIopt*

Table S4. Metabolite levels of cultures expressing the top-ranked constructs

Construct	*Trans-cinnamate (mg/L)*	*Pinocembrin (mg/L)*
DH5α	0.05	0
DH5α + pQlinkN	0.03	0
DH5α + HHHL	53.61	1
DH5α + HHHH	38.19	1.39

Table S5. Top scoring genes used for malonyl-CoA synthesis

Gene tag	*Encoded activity*	*EC number*	*Source*	*Accession number*	*RetroPath score*
matB	malonyl-CoA synthase	6.2.1.-	*Rhizobium trifolii*	YP_766603.1	1.46
matC	transmembrane dicarboxylate carrier protein		*Rhizobium trifolii*	YP_766604.1	
atoD	acetyl-CoA:acetoacetyl-CoA transferase a	2.8.3.3	*Bradyrhizobium sp. ORS 278*	YP_001203366.1	1.22
atoA	acetyl-CoA:acetoacetyl-CoA transferase b	2.8.3.3	*Bradyrhizobium sp. ORS 278*	YP_001203367.1	1.22
mmsA	malonate-semialdehyde dehydrogenase (acetylating)	1.2.1.75	*Coxiella burnetii*	NP_819940.1	1.19
cagg1256	short-chain dehydrogenase/reductase SDR	1.2.1.75	*Chloroflexus aggregans*	YP_002462600.1	0.64
accA	carboxyltransferase a	6.4.1.2	*E. coli*	944895	1.31
accB	biotin-carboxyl-carrier-protein	6.4.1.2	*E. coli*	947758	1.31
accC	biotin carboxylase	6.4.1.2	*E. coli*	947761	1.31
accD	carboxyltransferase b	6.4.1.2	*E. coli*	946796	1.31

Table S6. Calculated flux values for *E. coli* cells carrying the induced HHHH construct alone (A), with addition of malonyl-CoA producing pathway (B), and with addition of malonyl-CoA producing pathway and Cerulenin (C).

Flux	*A*	*B*	*C*	*Units*
Malonate	0.00E+00	-4.56E-03	-2.75E-02	mmol/gDW/h
Acetyl-CoA	-8.84E-02	-1.16E-01	-6.83E-02	mmol/gDW/h
L-Phenylalanine	-1.47E-01	-9.13E-02	-5.49E-01	mmol/gDW/h

Fatty Acids	8.81E-01	1.20E-01	8.94E-01	mmol/gDW/h
Growth	3.65E-02	4.97E-02	3.71E-02	1/h
Trans-Cinnamate	6.91E-03	6.18E-03	1.82E-02	mmol/gDW/h
Pinocembrin	1.04E-04	1.30E-04	2.13E-03	mmol/gDW/h

Table S7. Pinocembrin production of the twelve constructs

Pathway tag	Flux *mmol/(gDW*h)*	SD *mmol/(gDW*h)*	Titer *mg/L*	SD *mg/L*
HHHH	**0.310**	**0.054**	5.078	1.363
HHHL	0.226	0.047	**24.139**	**5.471**
HHLH	0.000	0.000	0.000	0.000
HHLL	0.000	0.000	0.000	0.000
HLHH	0.000	0.000	0.000	0.000
HLHL	0.000	0.000	0.000	0.000
HLLH	0.000	0.000	0.000	0.000
HLLL	0.000	0.000	0.000	0.000
HMHH	0.018	0.004	2.230	2.303
HMHL	0.034	0.017	4.234	3.485
HMLH	0.000	0.000	0.000	0.000
HMLL	0.000	0.000	0.000	0.000

Table S8. Trans-cinnamate production of the twelve constructs

Pathway tag	Efficiency *mmol/(gDW*h)*	SD *mmol/(gDW*h)*	Titer *mg/L*	SD *mg/L*
HHHH	2.614e-3	1.769e-4	40.047	2.710
HHHL	3.325e-3	4.180e-4	50.448	6.342
HHLH	6.343e-4	8.320e-6	19.370	0.254
HHLL	5.667e-3	8.278e-4	62.530	9.134
HLHH	7.475e-3	8.654e-5	46.054	0.533
HLHL	7.211e-3	2.027e-5	49.635	0.140
HLLH	5.621e-3	1.575e-5	75.829	0.212
HLLL	4.965e-3	7.446e-4	57.322	8.597
HMHH	5.985e-3	5.051e-3	71.228	8.542
HMHL	4.328e-3	1.550e-3	91.433	14.298
HMLH	1.484e-2	1.655e-3	72.600	8.100
HMLL	1.609e-2	7.211e-4	78.700	3.530

Table S9. Titers of pinocembrin produced earlier by engineered *E. coli* cells

Publication	*Pinocembrin (mg/L)*
(Kaneko et al., 2003a)	0.75
(Hwang et al., 2003)	0.752
(Miyahisa et al., 2005)	58
(Leonard et al., 2007)	429
(Leonard et al., 2008)	710
(Wu et al., 2013)	40.02

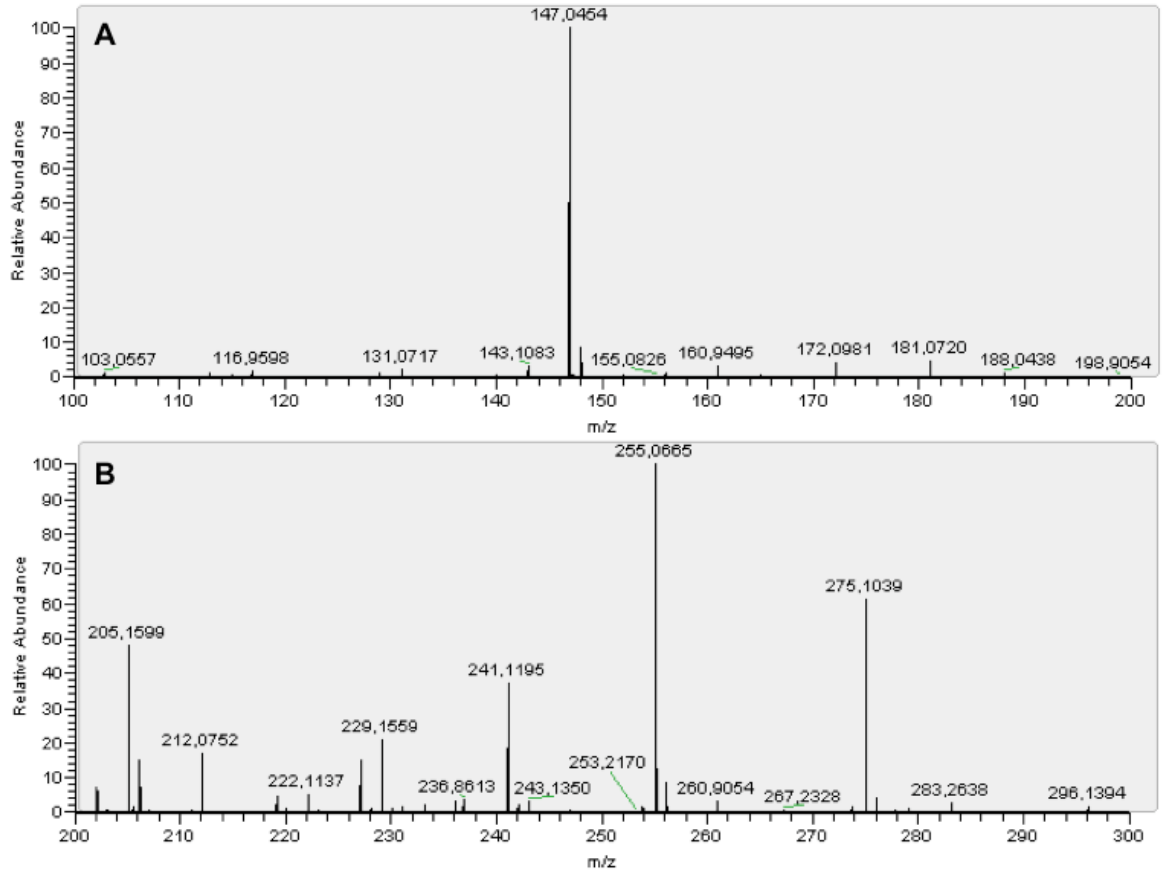

Figure S1. ESI mass spectrum obtained in the negative ionization mode (see section 2.6). A: m/z range 100-200; trans-cinnamate $[M-H]^-$ was detected at m/z 147.0454 (accuracy: 1.9 ppm). B: m/z range 200-300; pinocembrin $[M-H]^-$ was detected at m/z 255.0665 (accuracy: 0.7 ppm).

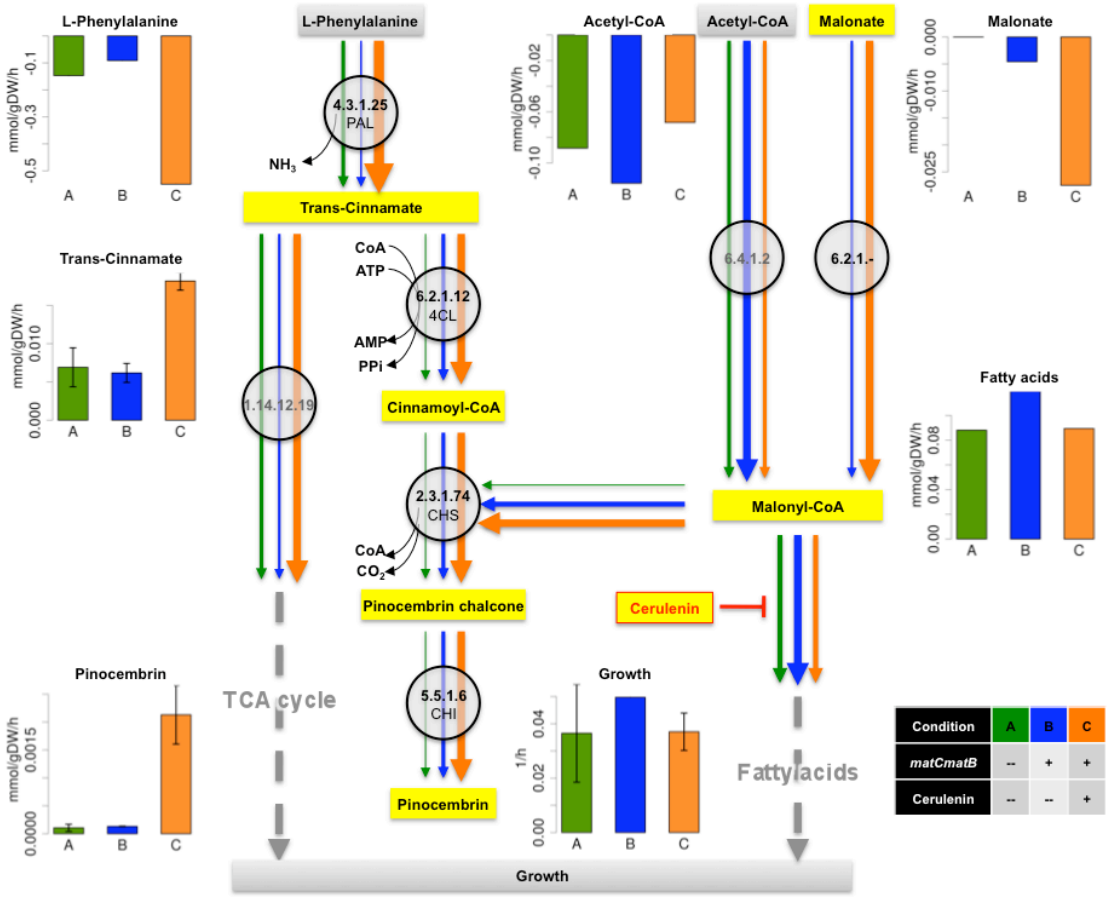

Figure S2. Estimated fluxes (bars) and measured fluxes (bars with error bars) in the pinocembrin-producing *E. coli* strains (A) HHHH construct; (B) HHHH + *matCmatB*; (C) HHHH + *matCmatB* + cerulenin supplementation. Arrow widths correspond to relative fluxes for each simulation. Bar plots are shown for consumption of main precursors (phenylalanine, acetyl-coA, malonate), production of fatty acids and growth, as well as for accumulation of trans-cinnamate and pinocembrin in the medium. Endogenous and heterologous *E. coli* metabolites are displayed in grey and yellow boxes respectively.

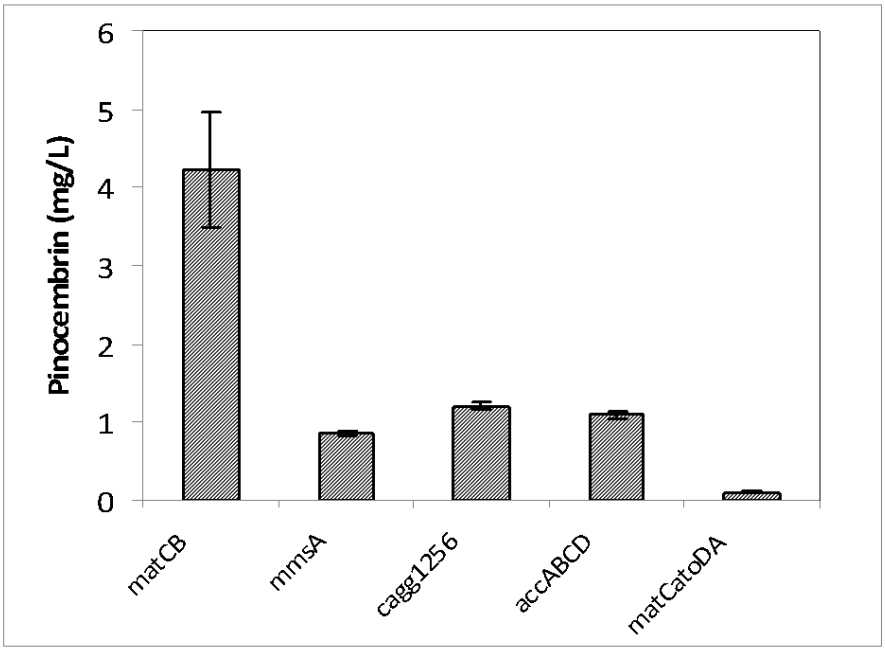

Figure S3. Pinocembrin titers measured 24 hours after inducing strains carrying the HHHH construct and a malonyl-CoA-producer pathway, in the absence of cerulenin.

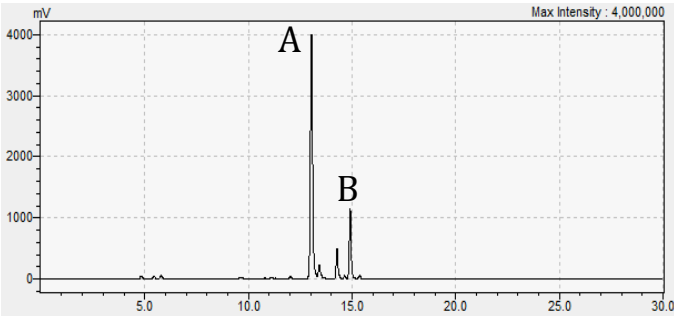

Figure S4. HPLC chromatogram of extracted medium detected at 290 nm. Peaks corresponding to (A) trans-cinnamate and (B) pinocembrin have a retention time of 13.1 and 14.9 min, respectively.