PATH^{cre8}: A Tool that Facilitates the Searching for Heterologous Biosynthetic Routes

Olaa Motwalli^{1,2,#}, Mahmut Uludag^{1,#}, Ivan Mijakovic^{3,4}, Meshari Alazmi⁵, Vladimir B. Bajic¹, Takashi Gojobori^{1,6}, Xin Gao^{1*}, Magbubah Essack^{1*}.

¹ Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.

² Saudi Electronic University (SEU), College of Computing and Informatics, Madinah 41538-53307, Kingdom of Saudi Arabia.

³ Chalmers University of Technology, Division of Systems & Synthetic Biology, Department of Biology and Biological Engineering, Kemivägen 10, 41296 Gothenburg, Sweden.

⁴ Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

⁵ Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, P.O. Box 2440, Ha'il, 81411, Saudi Arabia.

⁶ Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.

*To whom correspondence should be addressed: magbubah.essack@kaust.edu.sa and xin.gao@kaust.edu.sa

#Joint first authors

Abstract

Motivation: Developing computational tools that can facilitate the rational design of cell factories producing desired products at increased yields is challenging. As the tool needs to take into account that the preferred host organism usually has compounds that are consumed by competing reactions that reduce the yield of the desired product. On the other hand, the preferred host organisms may not have the native metabolic reactions needed to produce the compound of interest; thus, the computational tool needs to identify the metabolic reactions that will most efficiently produce the desired product.

Results: In this regard, we developed the generic tool PATH^{cre8} to facilitate an optimized search for heterologous biosynthetic pathway routes. PATH^{cre8} finds and ranks biosynthesis routes in a large number of organisms, including Cyanobacteria. The tool ranks the pathways based on feature scores that reflect reaction thermodynamics, the potentially toxic products in the pathway (compound toxicity), intermediate products in the pathway consumed by competing reactions (product consumption), and host-specific information such as enzyme copy number. A comparison with several other similar tools shows that PATH^{cre8} is more efficient in ranking functional pathways. To illustrate the effectiveness of PATH^{cre8}, we further provide case studies focused on isoprene production and the biodegradation of cocaine.

Availability: PATH^{cre8} is free for academic and nonprofit users and can be accessed at <https://www.cbrc.kaust.edu.sa/pathcre8/>.

Keywords: Modeling, heterologous biosynthetic pathways, competing endogenous metabolic reactions, optimization, microbial cell factories, isoprene, bioinformatics, computer science

Co-author, Professor Vladimir Bajic, passed away suddenly on 31 October 2019.

The metabolism of microbes can be customized with synthetic biology to produce desired chemicals of interest. This idea has been proven feasible with the commercial production of enzymes, drugs, and specialty chemicals in various organisms¹⁻⁴. However, such successes are few as synthetic pathways' engineering is laborious and time-consuming due to the trial and error associated with experimental processes. Thus, accurately predicting biosynthetic pathways needed for high-yield production of desired chemicals can streamline the engineering processes. Key factors that influence accurately predicting novel biosynthetic pathways include the selection of hosts, pathways, and enzymes, amongst others. Developing computational tools that take these factors into account is not an easy task. Nonetheless, PathPred⁵, Metabolic tinker⁶, DESHARKY⁷, OptStrain⁸, RetroPath2.0⁹, XTMS¹⁰, GEMPath¹¹, FMM¹², and MRE¹³ have been developed for the exploration of heterologous pathways aiming to assist synthetic biologists. The flux balance analysis (FBA) based methods implemented in XTMS ranks pathways based on context information such as enzyme performance (gene scores), the number of reaction steps, the toxicity of intermediate products, maximum pathway yield, and thermodynamic (Gibbs energy) favorability. On the other hand, basic methods, such as those implemented in FMM, reconstruct metabolic pathways from KEGG pathway maps. The limitation of XTMS is that pathways are ranked based on several criteria, including yield that is estimated based on *Escherichia coli* model fluxes only. Another method, MRE, is host-specific, as it uses the competition of endogenous reactions specific to the chassis organisms' metabolic make-up to rank heterologous pathways. Many existing tools focus their analysis and validation results on only *E. coli* and yeast, such as XTMS and MRE. We believe such analysis and validation results should also be focused on cyanobacteria because they are known microorganisms used as a cell factory for specific products¹⁴ and have several experimentally verified biosynthetic pathways to evaluate a tool's capabilities regarding cyanobacteria as hosts. Thus, we have developed PATH^{cre8} to predict and prioritize heterologous biosynthesis routes for desired compound production in selected chassis organisms, and have validated it through the experimentally known designed pathways of *E. coli*, yeast, and cyanobacteria.

Methods

PATH^{cre8} data resources

We extracted the PATH^{cre8} compound- and reaction-related data from the KEGG COMPOUND¹⁵ and KEGG LIGAND¹⁶ databases, respectively. Reaction directionality was assigned based on 1/ known reactions in Rhea^{17, 18} and KEGG MODULE¹⁹, or 2/ predicted thermodynamic data (Gibbs energy) generated using the eQuilibrator²⁰ (using default setting). For negative Gibbs energy ($\Delta G < 0$), the forward ($\rightarrow$) reaction direction was assigned, while for positive Gibbs energy ($\Delta G > 0$), the reverse ($\leftarrow$) reaction direction was assigned, and for zero Gibbs energy ($\Delta G = 0$), the bidirectional ($\leftrightarrow$) reaction status was assigned. The bidirectional ($\leftrightarrow$) reaction status was also assigned, in cases where the above process could not provide reaction directionality.

Constructing the “Universal Metabolic Network”

We used the metabolic reactions and reaction directionality to construct a host-independent bipartite universal metabolic network (UMN)²¹. In our bipartite representation, there are only two types of nodes: 1/ compound nodes, and 2/ reaction nodes. The same type of nodes is never directly

connected. Representing edges as transition nodes allow for the labeling of edges with reaction ID. We have chosen a bipartite representation because this representation has been widely used for pathway prediction problems due to its clarity and simplicity for reaction directionality consideration, as well as; this representation fixes several issues for other representations, including a compound graph, reaction graph, and undirected graph. These issues include the representation of one compound or one reaction several times. Another issue is the connectivity since one product could be traversed from another product in the same reaction, which violates the chemical laws and increases the computational complexity and produces incorrect and illogical results. Thus, because the use of such transition nodes can lead to errors in the computational interpretation of reaction, bidirectional reactions were represented by two separate transition nodes (for example, the bidirectional reaction “R00024” is represented by two transition nodes, 1/ “fR00024” for the forward reaction, and 2/ “rR00024” for the reverse reaction (rR) as used in²¹.

Assigning Scores to UMN edges

We assigned reaction scores (**rSC**) and compound scores (**cSC**) to the UMN (see Figure 1). The **rSC** were estimated using a combination of Gibbs energy (ΔG) and enzyme-copy number (*eCN*) for the host (see equation (1)). Only reactions present in the host organism could have *eCN* > 0. For each reaction, we calculated *eCN* in a selected host by counting the number of genes that code for a specific enzyme that catalyzes specific reactions using KEGG orthology. The *eCN* allowed us to identify host-specific enzymes that are present (in this instance given a score greater than 0) and identify which enzymes are missing or required to produce the new pathway using foreign genes (in this instance given a score of 0). So, pathways that, 1/ have more genes that encode an enzyme required for the new pathway, and 2/ require less foreign genes to be inserted, will be ranked more favorably. Thus, edge scores in the UMN are host-specific (scores change based on the selected organism). Moreover, **rSC** allows reactions with the more negative ΔG and higher *eCN* to be ranked more favorably.

$$rSC = \Delta G - eCN. \quad (1)$$

The **cSC** was estimated using a combination of compound toxicity scores (*cTOX*) and hypothetical product consumption scores (*cPC*) (see equation (7)). The *cTOX* for all compounds (which have MDL/Mol files) in our metabolic network were estimated using the Toxicity Estimation Software Tool (TEST) [<https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>] with ‘Consensus method’ for all endpoints (where endpoints are tested on five different species). Here, ‘Ames mutagenicity’ and ‘Bioconcentration factor’ were not considered as toxicity endpoints because they do not directly suggest compound toxicity. Classifying compounds as toxic is based on defined thresholds, that is, if a compound displays an LD₅₀ or LC₅₀ threshold lower than 5000 mg/kg in rats or 500 mg/L in fathead minnows, *Daphnia magna*, or *Tetrahymena pyriformis*, the compound is toxic, or if the compound is a developmental toxicant in human. The results were converted to the binary form (toxic/non-toxic); compounds without predicted toxicity were classified as non-toxic and given the value of 0. Also, we retrieved 1,410 toxic compounds from the Toxin and Toxin Target Database (T3DB)^{22,23}. These compounds were shown to be toxic in at least 11 different tested organisms. The 14 organisms used to test compound toxicity in both the TEST and T3DB databases were used to develop a method to estimate the likelihood of a compound’s toxicity based on the number of organisms in which the compound toxicity tested

Figure 1: A depiction of the methodology implemented in the PATHcre8 tool completed in four steps.

positive (available to use were 14 toxicity tested organisms that do not include microbial strains). That is, the final compound toxicity score ($cTOX$) is based on the number of organisms in which the compound toxicity is tested to be positive. $cTOX$ is calculated by the formula given below:

$$cTOX = \frac{\sum_{j=1}^{14} tox_j}{14}, \quad (2)$$

where tox_j is the binary score (0/1) that indicates compound toxicity in organism j . A compound can have a maximum toxicity score of 1 (14/14), when it is found to be toxic in all 14 organisms. On the other hand, if the compound was not found as toxic in any of the 14 organisms, it is represented by a score of 0 (0/14). Thus, the inclusion of compounds that are toxic to humans represents a small part of this fraction (1/14), and only exerts a real impact if it is toxic in a number of other organisms too. In our opinion using such a scale better suits our generic approach.

Moreover, if the user does not agree with this approach, we also include search options that allow the user to adjust the weight of the variables (α , β , γ , and δ defined below) as they see fit.

cPC were first calculated for compounds with an associated transport reaction that can transport them into or outside the cell; 597 such compounds were retrieved from the “Biochemistry” repository in the KBase²⁴ resource. The cPC were calculated as follows: The availability of each compound was set as $1/m$, where m is the number of competing pathways in which the compound participates. That is, it is assumed that each of the m outgoing edges uses an equal portion of the total compound (which is 1).

Step 1. To calculate the initial scores, we first determine cPC for: a) the 597 compounds extracted from KBase complete media; b) compounds in the UMN corresponding to nodes without incoming edges but with outgoing edges; and c) compounds associated with reverse reactions. Thus, for a node i we have:

$$initialScore_i = \frac{1}{m}, \quad (3)$$

where m is the total number of outgoing edges from compound-node (i).

Step 2. The cPC for the rest of the compounds for the node i were calculated as follows:

$$PredScore_i = \frac{\sum_{j=1}^n cHCS_j}{n}, \quad (4)$$

$$SuccScore_i = h + \left(\frac{g}{2}\right), \quad (5)$$

$$cPC_i = \frac{PredScore_i}{SuccScore_i}, \quad (6)$$

where cPC_j is the j -th predecessor of compound-node (i); the predecessor node j relative to node i , is linked to node i with the incoming edge; and n is the total number of incoming edges. $PredScore_i$ represents the total score obtained from all predecessor compound nodes to node i ; $SuccScore_i$ is the score obtained for all successor compound nodes relative to node i . Successor node relative to node i is the node that is linked to node i with the edge outgoing from node i . However, some edges may be bidirectional. In such a case, $SuccScore_i$ comprises the sum of unidirectional edges (h) and bidirectional edges (g) divided by two, as shown in (5).

Step 3. In the case where the cPC algorithm stops (could not assign cPC to any node) due to dependency, then the compound-nodes that have the maximal total number of unscored successors and predecessors in all graph components were selected and set to the *initialScore*, and the algorithm continues.

Thus, cSC allows compounds with the less toxicity score $cTOX$ and higher product consumption scores (cPC) to be ranked more favorably.

$$cSC = cTOX - cPC. \quad (7)$$

Optimization of the UMN

To rank pathways or routes, we calculated the pathway scores (PS) from the source to the target compound. We used an optimization approach to estimate the optimized values of α , β , γ , and δ weights of features representing cPC , $cTOX$, ΔG and eCN , respectively, which resulted in the values of 0.0345, 0.9548, 0.000005, 0.6988, respectively. We calculated the WeightedEdge for the PS by adding up the rSC and cSC between all compound-nodes of a path. For example, Figure 2 provides an abstract illustration of how the PS is calculated ($PS = cSC_1 + rSC_1 + cSC_2 + rSC_2$) for the path from ‘Caffeoyl-CoA’ to ‘Coniferyl aldehyde’.

Figure 2: An illustration of how path scores are calculated. The source compound id is represented with an orange oval, and the target compound is represented as a green oval, while the intermediate compound is represented as grey ovals. All other compounds that are not considered as primary metabolite precursors in the above reactions are white. The path for which the score is calculated is represented in blue.

In general,

$$PS_i = \sum_{j=1}^{length(path)} weightedEdge_j.$$

The objective function that we minimized was the sum of ranking positions of known designed pathways in the training set. The training set used for optimization training is 1/ Pinocembrin (C09827) from L-Phenylalanine (C00079), with *E. coli* as the host organism, 2/ Hexadecanoyl-

[acp] (C05764) from Malonyl-CoA (C00083), with cyanobacteria as the host organism, and 3/Tryptophan (C00078) from Chorismate (C00251), with yeast as the host organism.

The optimization problem was defined as follows:

$$\min_{\alpha, \beta, \gamma, \delta} \sum_{j=1}^3 \text{rankpos}_j \quad 1 > \alpha, \beta, \gamma, \delta > 0$$

$$cSC = (\beta * cTOX - \alpha * cPC),$$

$$rCS = (\gamma * \Delta G - \delta * eCN).$$

Simulated Annealing solver available in MATLAB Optimization Toolbox was used to find an optimized minimum.

Biosynthetic pathway route search

For route searches (see Figure 1) based on the user specifying both source and target compounds, top k promising routes are obtained through the K-shortest loopless path algorithm applied to the bipartite graph, based on the integration of Yen's algorithm²⁵ with A* heuristic, developed by the PathLinker project²⁶. Shortest path algorithms have been applied on bipartite graphs²¹ since the shortest path algorithm starts with a source node and traverses it regardless of its topological representation, trying to find the target node. For searches based on the user specifying only a target compound, the default shortest path algorithm available as part of the NetworkX library is applied to the bipartite graph²⁷.

Some weights were negative in the initial PATH^{cre8} graphs, for example, as a result of negative ΔG . The negative weights were normalized by dividing all negative numbers with the absolute value of the largest negative number. In this way, the lowest potential value becomes '-1'. To make all weights positive, '1' was added to both the negative and the positive weights. Hence, the values were shifted to have zero as the value reflecting the most negative ΔG . Hence, our pathway scoring function favors the lower score, which supports the more negative ΔG , the better.

Since the shortest path algorithm identifies the routes with the minimum weight without considering the length of paths, we also added an ad hoc penalty factor of 0.7 to all edge weights to encourage shorter routes to be recognized as the top-ranked. We used the default value of 5 for the k parameter, but users can specify a different value using the path-search input form. We also used a default value of 11 for the cutoff parameter; this means the path searching function stops when the search exceeds five reactions and six metabolites in a route. This default value was determined based on several test runs that showed this cutoff produces the best results. Also, to retrieve meaningful route searches, PATH^{cre8}, by default, excludes several side compounds that are highly consumed or produced by many reactions, such as cofactors (see Supplementary Table S1).

Web interface

The PATH^{cre8} web interface provides input forms that allow the user to specify route searches through, 1/ source-target inputs, and 2/ target-only input, as well as general PATH^{cre8} exploration

searches via compound and reaction IDs/keywords, etc. (see User Manual). The input form for route searches has additional options for the alpha, beta, gamma and the delta parameters that when selected PATH^{cre8} server recalculates the weights of the selected KEGG metabolite/reaction graph with the parameter values provided before executing the requested route search. The route search results are visualized in graphs using the Cytoscape.js JavaScript library, and also provided in tabular format. For the graph layout, we used cytoscape.js-dagre layout [<http://doi.org/10.5281/zenodo.1211727>]. We enriched the tabular results with reaction/compound/enzyme data collected from KEGG and other dataset collections (MetaNetX²⁸, IntEnz²⁹, UniProt³⁰, and ModelSEEDdb [<https://github.com/ModelSEED/ModelSEEDDatabase>], and which were all indexed with MongoDB in PATH^{cre8}). When initiating a path-search request, PATH^{cre8} also executes path-search on IntEnz, BIGG³¹, Rhea^{17, 18}, and ModelSEEDdb metabolite graphs. Results of these additional route-searches show the connections of the given source-target compounds in other reaction sets that could help users to interpret PATH^{cre8} graphs better. Construction of all the metabolite graphs requires finding the reactant-product pairs in the reactions listed in these datasets. The IntEnz dataset has structured reaction data; PATH^{cre8} uses MongoDB query operators to get the reactant-product pairs directly in the query results. MetaNetX and ModelSEEDdb reactions are in semi-structured text form; PATH^{cre8} employed open-source parsers to find reactant-product pairs in these reactions, namely [<https://github.com/mmundy42/cobrababel>, <https://github.com/ModelSEED/ModelSEEDDatabase>].

Graphical overview of the PATH^{cre8} organisms' metabolite graphs is available through the statistics page. An input form on this page allows users to select organisms and to specify the limits for pruning the highly connected metabolites and the small subgraphs. Organism graphs are displayed using the Euler layout algorithm for Cytoscape.js [<http://doi.org/10.5281/zenodo.1211731>]. To achieve views readable by eye, we also deleted the reaction nodes in the original PATH^{cre8} graphs and added reactions as edge data between the metabolite nodes. Cytoscape.js-navigator library was used to provide a bird's eye view pan and zoom control [<http://doi.org/10.5281/zenodo.1308989>].

Results and Discussion

PATH^{cre8} algorithm and its computational performance

Figure 1 summarizes the PATH^{cre8} algorithm (also see the Materials and Methods section). In the first processing step (S1), the PATH^{cre8} algorithm uses the input data (including compound data, reactions, and reaction directionality) to construct a universal metabolic network (UMN). The UMN comprises 7,964 unique compound nodes (that represent both substrates and products) and 9,896 reaction nodes. The 9,896 reaction nodes comprise 1,664 reaction nodes with a defined reaction direction and 8,232 bidirectional nodes (that include all nodes with undefined reaction directionality). In the second processing step (S2), the algorithm assigns reaction scores (*rSC*) and compounds scores (*cSC*) to the UMN, that is, we calculated feature scores through optimization. Then, we used these scores to generate an optimized weighted network (S3). In the last processing step (S4), the algorithm uses the edge weights that were calculated as the sum of *rSC* and *cSC* between all compound-nodes in the path of interest from source to the target compound, to produce the pathway scores (PS) (see Materials and Methods section). PATH^{cre8} uses the PS to predict and rank heterologous metabolic pathways with the completion of these four steps (S1-S4). We used

this optimization approach because it is reasonable to assume that not all scoring features (such as the *cPC*, *cTOX* used to calculate *cSC*, as well as ΔG and *eCN* used to calculate *rSC*) have the same level of influence when searching for the most promising heterologous biosynthesis routes.

The processing time for the route-search and ranking step were measured by randomly selecting 100 connected pairs of source and target compounds and executed path search with a path length cutoff parameter of 10 and 10 as the number of maximum paths. The average return time was about 3.4 seconds per compound pair on a computer with Intel Xeon 2.67 GHz processors (all searches are done sequentially without parallelization).

Comparison with known heterologous pathway prediction tools

The tabulated features of the state-of-the-art heterologous pathway prediction tools (MRE, FMM, XTMS, see Table 1) show that PATH^{cre8} has additional features that increase its functionality and likely prediction accuracy. Concerning functionality, PATH^{cre8} allows for route-searches using multiple organisms similar to MRE and FMM. Based on our tests (see Table 2), PATH^{cre8} performs Cyanobacteria-related and *E. coli*-related route searches better than any of the tools we used in the comparison. In the yeast-related route searches, PATH^{cre8} and MRE perform the same, while FMM and XTMS were not successful (see Table 2). Moreover, PATH^{cre8} supports route searches via the input of specific source and target compounds similar to MRE and FMM, and the input of “target compounds only”, similar to XTMS/RetroPath. However, XTMS route searches based on the “target compound only” is generated through the implementation of Extended Metabolic Space using molecular signatures, whereas PATH^{cre8} route searches based on compound and reaction identifiers. Thus, one would expect that XTMS should produce better target compound only search results than PATH^{cre8}. However, XTMS implements flux balance analysis (FBA) that requires organism-specific data (that are only available for well-studied organisms such as *E. coli*) to determine yield. Thus, even though XTMS is purportedly generic, it is biased towards *E. coli* because it ranks pathways based on estimated yield scores obtained for *E. coli* only. Nonetheless, this “target only search option” allows users to possibly identify novel source compounds for the production of the target compound. Also, a search option that allows the weight of the variables to be tuned has been added so that advanced users can perform exploratory tasks.

Concerning prediction accuracy, route search prediction errors increase when “reaction directionality” is not well defined. Thus, PATH^{cre8} uses not only KEGG (as used in FMM) and ΔG (used in MRE) to determine “reaction directionality”, but also used Rhea, the manually curated resource of biochemical reactions, which was not used by MRE, FMM or XTMS. The direction of 4,984 PATH^{cre8} graph reactions is set based on Rhea, 563 reaction directions are set based on the KEGG module, and 1,979 reaction directions are inferred via ΔG , while the remaining reaction directions that were undefined are labeled bidirectional.

Here too, features used to rank route searches affect the easy identification of viable alternative routes. To clarify, it is easier to identify viable alternative routes if they are scattered in the top 10-20 ranked routes as opposed to routes scattered in the top 100 ranked routes. In this regard, ranking scores are calculated based on ΔG and competing for endogenous reactions for MRE and the linking of KEGG pathway maps based on the presence or absence of reactions in specific species for FMM. On the other hand, the ranking score for PATH^{cre8} is the PS, which is the sum of the *rSC* and *cSC*, where *rSC* is estimated using a combination of ΔG and *eCN* for the host (see equation (1)). Thus, only reactions present in the host organism could have *eCN* > 0. In this manner, reactions present in an organism obtain a higher score. However, since the shortest path

algorithm ranks pathways with the minimum score, to favor paths with high copy numbers, we multiplied the eCN by -1. This scoring increases the chances that the routes with, 1/ higher eCN that could contribute to higher yield, and 2/ pathways with most of the required enzymes (less heterologous reactions) ranks higher.

Similarly, we also integrated cSC that is estimated using a combination of $cTOX$ and cPC . The $cTOX$ allows the routes that include toxic compounds to obtain a higher value to increase the chances that top rank routes would not include toxic compounds. We also calculated cPC that allows intermediate products in pathways that are consumed by competing reactions to be penalized for increasing the chances that top rank routes would include the compounds with less competing reactions. These parameters ensure that the accuracy of route predictions is coupled with the identification of routes more suitable for engineering. The optimization approach we used for ranking routes also adds to prediction accuracy as it increases the chances that the viable alternative routes are top-ranked, by ensuring known heterologous pathways/reactions are among the top-ranked and that weights assigned to features at least reflect a reasonable ranking of known heterologous pathway among the k top-ranked routes.

Evaluating the ability of PATH^{cre8} to perform route-searches

We evaluated the ability of PATH^{cre8} to predict and prioritize heterologous pathways using 20 experimentally confirmed heterologous pathways. Of the 20 pathways, ten pathways were evaluated using *E. coli* K-12 MG1655 as a host organism, two pathways using *Saccharomyces cerevisiae* as the host organism, and seven pathways using cyanobacterial strains as host organisms. Table 2 shows that all 20 pathways were found in the top 15 ranked heterologous pathways, indicating that PATH^{cre8} is identifying and prioritizing experimentally confirmed heterologous pathways well. Moreover, when comparing performances of PATH^{cre8} and other similar open-access web tools such as FMM, XTMS, and MRE (Table 2), it is clear that all these tools are good at predicting and prioritizing pathways where *E. coli* is the host organism. Specifically, as part of its 36 top-ranked routes, MRE was able to generate the known yeast pathways. It, however, missed two *E. coli* and four cyanobacterial pathways. Also, FMM could not retrieve two cyanobacterial pathways, four *E. coli* pathways, and all of the yeast pathways. XTMS could not retrieve four heterologous pathways in *E. coli*, even though it is focused on *E. coli*, and it missed nine of the other heterologous pathways. On the other hand, PATH^{cre8} retrieved all pathways as part of its 4 top-ranked routes. These results suggest that PATH^{cre8} is more generic and prioritizes more accurate predictions than FMM, XTMS, and MRE. It is the only heterologous pathway prediction tool that efficiently generates possible heterologous pathways for cyanobacterial hosts as well. This is of interest as cyanobacteria have been recognized as a suitable host organism for the production of free fatty acids (FFA) because they use carbon dioxide for FFA production and are capable of secreting FFAs. In this regard, the *in silico* screening method called FFASC^{32, 33} has been developed for the selection of appropriate cyanobacterial hosts. However, adequate knowledge of the combination of pathway and enzymes required for high-yield FFA production is lacking, and PATH^{cre8} may be able to assist research in this regard.

Biosynthesis of Isoprene

Proof-of-concept has been established for terpene-based renewable biofuels^{34, 35}, bringing forth an interest in increasing metabolic flux that feeds the terpenoid biosynthetic pathways thereby increasing terpenoid production yield³⁶⁻³⁸. This goal was achieved with the heterologous expression of the archaeal/eukaryotic terpenoid biosynthetic mevalonic acid (MVA) pathway. This pathway can operate in parallel with the native methylerythritol-phosphate (MEP) pathway in mutant *Synechocystis* sp. PCC 6803^{34, 39} to increase the prenyl diphosphate precursors (isopentenyl-diphosphate (IPP) and dimethylallyl diphosphate (DMAPP)) for isoprene and other terpenoids⁴⁰. Thus, we set out here to assess, for the first time, if PATH^{cre8}, MRE, FMM, and XTMS can predict and prioritize this pathway using cyanobacteria as a host. Therefore, we applied the following steps: 1/ a compound native to cyanobacteria, Acetoacetyl-CoA (C00332) was chosen as the source compound to search for the target compound IPP (C00129), and 2/ (R)-Mevalonate (C00418) was chosen as the source compound to search for the target compound Isoprene (C16521), both of which are not native to cyanobacteria. The Acetoacetyl-CoA to IPP pathway (see Figure 3) was retrieved by PATH^{cre8}, MRE, and FMM, as one of their top 15 ranked pathways, but XTMS was not able to retrieve this pathway. Also, the experimentally proven (R)-Mevalonate to Isoprene pathway via Isopentenyl phosphate (C20345) is retrieved by PATH^{cre8} as the top-ranked pathway⁴¹. These results highlight an explicit limitation of FMM, which is that it only retrieved the heterologous pathway where the source compound is native to the host. On the other hand, MRE could not retrieve the (R)-Mevalonate to Isoprene pathway via Isopentenyl phosphate. It was able to retrieve the (R)-Mevalonate to Isoprene pathway via (R)-5-Diphosphomevalonate (C01143) (see Figure 3) but not within the top 50 ranked pathways. Based on this analysis (and Table 2), the results suggest that PATH^{cre8} is generating alternative known heterologous pathways, not generated by some of the other methods.

Figure 3: An illustration of the known Mevalonate (MVA) pathway (with alternative routes) generated as top-ranked pathway/reactions by PATH^{cre8}. The Mevalonate pathway via Isopentenyl phosphate is depicted in black.

Biodegradation of Cocaine

Cocaine (also known as coke) is an addictive tropane alkaloid compound used as an illegal stimulant or as an anesthetic medicine extracted from the leaves of coca plants (*Erythroxylum* spp.)⁴². Thus, we set out to assess, for the first time, if PATH^{cre8} can be used for biodegradation, rather than bioproduction. As an example, we define cocaine (C01416) as the source and pseudoecgonyl-CoA (C12450) as the target, PATH^{cre8} determines the biodegradation pathway of cocaine as the top-ranked pathway (see Figure 4). PATH^{cre8} shows that in bacteria, cocaine breakdown is initiated by cocaine esterase, creating ecgonine methyl ester. Ecgonine methyl ester is then broken down to form ecgonine, which leads to pseudoecgonine. Finally, it can be seen that the pseudoecgonine results in the target compound pseudoecgonyl-CoA. This cocaine biodegradation pathway was confirmed through experimental testing⁴². The biodegradation pathway prediction system EAWAG-BBD⁴³ (http://eawag-bbd.ethz.ch/coc/coc_map.html), also which presents the same biodegradation pathway predicted by PATH^{cre8}. On the other hand, MRE, FMM, and XTMS could not retrieve the biodegradation/bioproduction pathway for cocaine using *Pseudomonas fluorescens* as a host.

Figure 4: An illustration of the known cocaine biodegradation pathway (with alternative routes) generated as top-ranked pathway/reactions by PATH^{cre8}.

Concluding Remarks

The reason behind developing the PATH^{cre8} method and tool was that available approaches were not generic enough to allow for effective pathway analyses with different host organisms. We have developed PATH^{cre8} as a generic biosynthetic route search method and associated tool for multiple organisms. PATH^{cre8} is a preliminary study that focuses on ranking potential heterologous pathways that have yet to be experimentally tested. This will help in narrowing down potential pathways for high-yield production of desired chemicals. The success of our approach has been demonstrated by using cyanobacteria as a host. Genericity of our approach has been achieved by optimizing features based on *E. coli*, *S. cerevisiae*, and cyanobacterial strains. Several known heterologous pathways are generated by PATH^{cre8} successfully amongst the 4 top ranked routes. This ensures biosynthetic pathways with the most promise is top ranked in all hosted organisms. For assignment of reaction directionality, reactions from Rhea and KEGG MODULE were used. Subsequently, Gibbs energy was used to assign directionality to the remaining reactions. This ensures that less false positive pathways are being generated by PATH^{cre8} in comparison with the other methods, which generalized the reaction direction by assigning bidirectionality to all reactions.

A combination of selected features that are used by MRE and XTMS are employed in PATH^{cre8}. These features ensure that top-ranked pathways contain compounds with smaller toxicity scores, require less energy and involve less competitive pathways. To quantify the effects of different features to the selection of suitable pathways/reactions, and in that way to optimize the selection of these pathways/reactions, we set up the problem as constrained optimization. Testing with cyanobacteria with the optimized weights has shown that our approach is sound.

We have not implemented weights to the metabolite graphs generated from datasets other than KEGG. As future work, we intend to add weights to these graphs based on similar features we used in optimized PATH^{cre8} graphs; this should provide a more fair comparison of results across different metabolite graphs. Also, the current tab-based presentation of multiple path-search results and neighbor graphs does not allow quick comparison, so we intend to look for alternative ways for easy comparison of the results from multiple graphs. Also, many tools depend on KEGG as their primary source of data, limiting the discovery of novel compounds. Similarly, PATH^{cre8} relies on the KEGG and Rhea databases, and suffer from the same limitation. Nevertheless, KEGG is known to increase the number of verified reactions consistently, which would minimize this limitation over time. Furthermore, we will also consider integrating additional metabolic reaction databases when they become available.

We provide biosynthesis of isoprene and the biodegradation of cocaine case studies, as well as a compare PATH^{cre8} results with the state-of-the-art methods using a benchmark dataset (20 known pathways) to show it performs better in this capacity i.e. retrieving known heterologous pathways. Retrieving known heterologous pathways does not directly show that the new heterologous pathways introduced by PATH^{cre8}, MRE, FMM, and RetroPath are valid. However, it increases confidence in the performance of the tool, as the logic used is if the tool does not produce the heterologous pathways known to be valid through experimental evidence, then how confident can we be that other alternative heterologous pathways introduced by the tool could be valid/correct as well. As future work, we intend to experimentally validate a few of the predicted pathways to increase confidence in this tool's performance.

Funding

This work has been supported by the King Abdullah University of Science and Technology (KAUST) Base Research Fund BAS/1/1606-01-01, BAS/1/1059-01-01, BAS/1/1624-01-01, as well as through the Awards Nos. FCC/1/1976-02-01, FCC/1/1976-17-01, FCC/1/1976-16-01, and FCC/1/1976-26-01 from the KAUST Office of Sponsored Research (OSR). Also, the Novo Nordisk Foundation grant NNF10CC1016517 awarded to IM.

Conflict of Interest. None declared.

Supporting Information. The side compounds excluded from all predicted pathways.

Table 1. Feature summary of heterologous pathway design methods with open access to associated web applications

	PATH^{cre8}	MRE	FMM	XTMS
Chassis	many organisms	many organisms	many organisms	only <i>E. coli</i>
Metabolic data source	KEGG reactions	verified KEGG reactions only	KEGG reactions	MetaCyc
Data Representation	directed bipartite graph	directed graph	directed graph	directed hypergraphs
Reaction directionality	KEGG MODULE, Rhea, Gibbs energy	Gibbs energy	KEGG pathway maps	Gibbs energy
Features	toxicity, enzyme copy number, Gibbs energy, <i>c</i> PC (product consumption scores)	Gibbs energy	KEGG pathway maps	gene scores, number of reaction steps, toxicity, yield, Gibbs energy
Ranking score	optimization	normalized Boltzmann weights	-	Optimization
Consideration of host metabolic infrastructure	based on competing endogenous reactions	based on competing endogenous reactions	not in ranking	based on FBA
Search input	only target or both source and target compounds	source and target compounds	source and target compounds	only target compound

Table 2: Known heterologous biosynthesis pathways generated by PATH^{cre8} and other tools. For PATH^{cre8}, 5 reaction steps were used as a cutoff. Abbreviations: organism (Org.), cyanobacteria (Cyan.), references (Ref.), *Pseudomonas fluorescens* (Pfo.).

Source	Target	Org.	Steps	Native: Foreign	pathway rank				Ref.
					PATH ^{cre8}	MRE	FMM	XTMS	
(R)-Mevalonate (C00418)	Isoprene (C16521)	Cyan.	5	(4:1)	2	1	No	No	40
Acetoacetyl-CoA (C00332)	Isopentenyl diphosphate (C00129)	Cyan.	5	(5:0)	3	1	1	No	40
3-Oxostearoyl-[acp] (C16219)	Octadecanoic acid (C01530)	Cyan.	4	(3:1)	1	No	1	No	44
3-Oxostearoyl-[acp] (C16219)	Oleic acid (C00712)	Cyan.	5	(3:2)	1	No	1	No	44
Malonyl-[acp] (C01209)	Palmitic acid (C00249)	Cyan.	5	(4:1)	1	No	1	No	44
Malonyl-[acp] (C01209)	Lauric acid (C02679)	Cyan.	5	(4:1)	1	36	1	No	44
Long-chain acyl-CoA (C02843)	Wax ester (C01629)	Cyan.	2	(1:1)	3	No	No	No	45
L-Tyrosine (C00082)	Naringenin (C00509)	<i>E. coli</i>	4	(0:4)	1	1	1	5	13, 46
Glycerol (C00116)	1,3-Propanediol (C02457)	<i>E. coli</i>	2	(0:2)	1	1	1	No	13, 47
L-Tyrosine (C00082)	Resveratrol (C03582)	<i>E. coli</i>	3	(0:3)	1	1	1	No	13, 48
D-Xylose (C00181)	Xylitol (C00379)	<i>E. coli</i>	2	(1:1)	3	3	2	No	13, 49
Tryptophan* (C00078)	Histidine* (C00135)	<i>E. coli</i>	4	(3:1)	4	9	No	No	50

Methylmalonate (C02170)	(R)-Methylmalonyl-CoA (C01213)	<i>E. coli</i>	1	(0:1)	1	No	No	1	51
3-Oxopropanoate* (C00222)	Malonyl-CoA* (C00083)	<i>E. coli</i>	1	(0:1)	1	1	1	2	52, 53
Acetyl-CoA (C00024)	Malonyl-CoA (C00083)	<i>E. coli</i>	1	(1:0)	1	1	No	4	52, 54-56
Malonate* (C00383)	Malonyl-CoA* (C00083)	<i>E. coli</i>	1	(0:1)	1	No	1	3	52
Malonyl-CoA (C00083)	Naringenin (C00509)	<i>E. coli</i>	2	(0:2)	1	1	No	4	52, 56, 57
Glycerol (C00116)	(R)-Propane-1,2-diol (C02912)	Yeast	5	(2:3)	4	2	No	No	13, 58
(R)-5-Phosphomevalonate (C01107)	Artemisinic acid (C20309)	Yeast	5	(3:2)	2	2	No	No	59
Cocaine (C01416)	Pseudoecgonyl-CoA (C12450)	Pfo.	4	(0:4)	1	No	No	No	[42]

References

- Nielsen, J.; Archer, J.; Essack, M.; Bajic, V. B.; Gojobori, T.; Mijakovic, I. Building a bio-based industry in the Middle East through harnessing the potential of the Red Sea biodiversity. *Appl Microbiol Biotechnol* **2017**, *101* (12), 4837-4851.
- Paddon, C. J.; Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat Rev Microbiol* **2014**, *12* (5), 355-67.
- Tsuruta, H.; Paddon, C. J.; Eng, D.; Lenihan, J. R.; Horning, T.; Anthony, L. C.; Regentin, R.; Keasling, J. D.; Renninger, N. S.; Newman, J. D. High-level production of amorpha-4,11-diene, a precursor of the antimalarial agent artemisinin, in *Escherichia coli*. *PLoS One* **2009**, *4* (2), e4489.

4. Westfall, P. J.; Pitera, D. J.; Lenihan, J. R.; Eng, D.; Woolard, F. X.; Regentin, R.; Horning, T.; Tsuruta, H.; Melis, D. J.; Owens, A.; Fickes, S.; Diola, D.; Benjamin, K. R.; Keasling, J. D.; Leavell, M. D.; McPhee, D. J.; Renninger, N. S.; Newman, J. D.; Paddon, C. J. Production of amorphadiene in yeast, and its conversion to dihydroartemisinin acid, precursor to the antimalarial agent artemisinin. *Proc Natl Acad Sci U S A* **2012**, *109* (3), E111-8.
5. Moriya, Y.; Shigemizu, D.; Hattori, M.; Tokimatsu, T.; Kotera, M.; Goto, S.; Kanehisa, M. PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res* **2010**, *38* (Web Server issue), W138-43.
6. McClymont, K.; Soyer, O. S. Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways. *Nucleic Acids Res* **2013**, *41* (11), e113.
7. Rodrigo, G.; Carrera, J.; Prather, K. J.; Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **2008**, *24* (21), 2554-6.
8. Pharkya, P.; Burgard, A. P.; Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* **2004**, *14* (11), 2367-76.
9. Delepine, B.; Duigou, T.; Carbonell, P.; Faulon, J. L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab Eng* **2018**, *45*, 158-170.
10. Carbonell, P.; Parutto, P.; Herisson, J.; Pandit, S. B.; Faulon, J. L. XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res* **2014**, *42* (Web Server issue), W389-94.
11. Campodonico, M. A.; Andrews, B. A.; Asenjo, J. A.; Pálsson, B. O.; Feist, A. M. Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. *Metab Eng* **2014**, *25*, 140-58.
12. Chou, C. H.; Chang, W. C.; Chiu, C. M.; Huang, C. C.; Huang, H. D. FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res* **2009**, *37* (Web Server issue), W129-34.
13. Kuwahara, H.; Alazmi, M.; Cui, X.; Gao, X. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. *Nucleic Acids Res* **2016**, *44* (W1), W217-25.
14. Zahra, Z. C., D.H.; Lee, H. Cyanobacteria: Review of Current Potentials and Applications. *Environments* **2020**, (7(2)), 13.
15. 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* (Database issue), D109-14.
16. Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* **2002**, *30* (1), 402-4.
17. Alcantara, R.; Axelsen, K. B.; Morgat, A.; Belda, E.; Coudert, E.; Bridge, A.; Cao, H.; de Matos, P.; Ennis, M.; Turner, S.; Owen, G.; Bougueleret, L.; Xenarios, I.; Steinbeck, C. Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res* **2012**, *40* (Database issue), D754-60.
18. Morgat, A.; Axelsen, K. B.; Lombardot, T.; Alcantara, R.; Aimo, L.; Zerara, M.; Niknejad, A.; Belda, E.; Hyka-Nouspikel, N.; Coudert, E.; Redaschi, N.; Bougueleret, L.; Steinbeck, C.; Xenarios, I.; Bridge, A. Updates in Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res* **2015**, *43* (Database issue), D459-64.
19. Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **2014**, *42* (Database issue), D199-205.
20. Flamholz, A.; Noor, E.; Bar-Even, A.; Milo, R. eQuilibrator--the biochemical thermodynamics calculator. *Nucleic Acids Res* **2012**, *40* (Database issue), D770-5.

21. Faust, K.; Croes, D.; van Helden, J. Prediction of metabolic pathways from genome-scale metabolic networks. *Biosystems* **2011**, *105* (2), 109-21.
22. Lim, E.; Pon, A.; Djoumbou, Y.; Knox, C.; Shrivastava, S.; Guo, A. C.; Neveu, V.; Wishart, D. S. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res* **2010**, *38* (Database issue), D781-6.
23. Wishart, D.; Arndt, D.; Pon, A.; Sajed, T.; Guo, A. C.; Djoumbou, Y.; Knox, C.; Wilson, M.; Liang, Y.; Grant, J.; Liu, Y.; Goldansaz, S. A.; Rappaport, S. M. T3DB: the toxic exposome database. *Nucleic Acids Res* **2015**, *43* (Database issue), D928-34.
24. Arkin, A. P.; Cottingham, R. W.; Henry, C. S.; Harris, N. L.; Stevens, R. L.; Maslov, S.; Dehal, P.; Ware, D.; Perez, F.; Canon, S.; Sneddon, M. W.; Henderson, M. L.; Riehl, W. J.; Murphy-Olson, D.; Chan, S. Y.; Kamimura, R. T.; Kumari, S.; Drake, M. M.; Brettin, T. S.; Glass, E. M.; Chivian, D.; Gunter, D.; Weston, D. J.; Allen, B. H.; Baumohl, J.; Best, A. A.; Bowen, B.; Brenner, S. E.; Bun, C. C.; Chandonia, J. M.; Chia, J. M.; Colasanti, R.; Conrad, N.; Davis, J. J.; Davison, B. H.; DeJongh, M.; Devoid, S.; Dietrich, E.; Dubchak, I.; Edirisinghe, J. N.; Fang, G.; Faria, J. P.; Frybarger, P. M.; Gerlach, W.; Gerstein, M.; Greiner, A.; Gurtowski, J.; Haun, H. L.; He, F.; Jain, R.; Joachimiak, M. P.; Keegan, K. P.; Kondo, S.; Kumar, V.; Land, M. L.; Meyer, F.; Mills, M.; Novichkov, P. S.; Oh, T.; Olsen, G. J.; Olson, R.; Parrello, B.; Pasternak, S.; Pearson, E.; Poon, S. S.; Price, G. A.; Ramakrishnan, S.; Ranjan, P.; Ronald, P. C.; Schatz, M. C.; Seaver, S. M. D.; Shukla, M.; Sutormin, R. A.; Syed, M. H.; Thomason, J.; Tintle, N. L.; Wang, D.; Xia, F.; Yoo, H.; Yoo, S.; Yu, D. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* **2018**, *36* (7), 566-569.
25. Yen, J. Y. Finding the K Shortest Loopless Paths in a Network. *Management Science* **1971**, *17* (11).
26. Ritz, A., Poirel, C., Tegge, A. et al. Pathways on demand: automated reconstruction of human signaling networks. *npj Systems Biology and Applications* **2016**, *2* (16002).
27. Hagberg Aric, S. P., S Chult Daniel Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SCIPY 08)*, Pasadena, CA USA, 2008.
28. Ganter, M.; Bernard, T.; Moretti, S.; Stelling, J.; Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* **2013**, *29* (6), 815-6.
29. Fleischmann, A.; Darsow, M.; Degtyarenko, K.; Fleischmann, W.; Boyce, S.; Axelsen, K. B.; Bairoch, A.; Schomburg, D.; Tipton, K. F.; Apweiler, R. IntEnz, the integrated relational enzyme database. *Nucleic Acids Res* **2004**, *32* (Database issue), D434-7.
30. The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2017**, *45* (D1), D158-D169.
31. Schellenberger, J.; Park, J. O.; Conrad, T. M.; Palsson, B. O. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **2010**, *11*, 213.
32. Motwalli, O.; Essack, M.; Salhi, A.; Hanks, J.; Mijakovic, I.; Bajic, V. B. BioPS: System for screening and assessment of biofuel-production potential of cyanobacteria. *PLoS One* **2018**, *13* (8), e0202002.
33. Motwalli, O.; Essack, M.; Jankovic, B. R.; Ji, B.; Liu, X.; Ansari, H. R.; Hoehndorf, R.; Gao, X.; Arold, S. T.; Mineta, K.; Archer, J. A.; Gojobori, T.; Mijakovic, I.; Bajic, V. B. In silico screening for candidate chassis strains of free fatty acid-producing cyanobacteria. *BMC Genomics* **2017**, *18* (1), 33.

34. Lindberg, P.; Park, S.; Melis, A. Engineering a platform for photosynthetic isoprene production in cyanobacteria, using *Synechocystis* as the model organism. *Metab Eng* **2010**, *12* (1), 70-9.
35. Peralta-Yahya, P. P.; Ouellet, M.; Chan, R.; Mukhopadhyay, A.; Keasling, J. D.; Lee, T. S. Identification and microbial production of a terpene-based advanced biofuel. *Nat Commun* **2011**, *2*, 483.
36. Bentley, F. K.; García-Cerdán, J.G.; Chen, H.C. et al. Paradigm of Monoterpene (β -phellandrene) Hydrocarbons Production via Photosynthesis in Cyanobacteria. *BioEnergy Research* **2013**, *6* (3), 917-929.
37. Leonard, E.; Ajikumar, P. K.; Thayer, K.; Xiao, W. H.; Mo, J. D.; Tidor, B.; Stephanopoulos, G.; Prather, K. L. Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc Natl Acad Sci U S A* **2010**, *107* (31), 13654-9.
38. Martin, V. J.; Pitera, D. J.; Withers, S. T.; Newman, J. D.; Keasling, J. D. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat Biotechnol* **2003**, *21* (7), 796-802.
39. Bentley, F. K.; Melis, A. Diffusion-based process for carbon dioxide uptake and isoprene emission in gaseous/aqueous two-phase photobioreactors by photosynthetic microorganisms. *Biotechnol Bioeng* **2012**, *109* (1), 100-9.
40. Bentley, F. K.; Zurbriggen, A.; Melis, A. Heterologous expression of the mevalonic acid pathway in cyanobacteria enhances endogenous carbon partitioning to isoprene. *Mol Plant* **2014**, *7* (1), 71-86.
41. Wang, Q.; Quan, S. & Xiao, H. Towards efficient terpenoid biosynthesis: manipulating IPP and DMAPP supply. *Bioresour. Bioprocess* **6** (6 (2019)).
42. Lister, D. L.; Sproule, R. F.; Britt, A. J.; Lowe, C. R.; Bruce, N. C. Degradation of cocaine by a mixed culture of *Pseudomonas fluorescens* MBER and *Comamonas acidovorans* MBLF. *Appl Environ Microbiol* **1996**, *62* (1), 94-9.
43. Gao, J.; Ellis, L. B.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Res* **2010**, *38* (Database issue), D488-91.
44. Liu, X.; Sheng, J.; Curtiss, R., 3rd. Fatty acid production in genetically modified cyanobacteria. *Proc Natl Acad Sci U S A* **2011**, *108* (17), 6899-904.
45. Li, F.; Wu, X.; Lam, P.; Bird, D.; Zheng, H.; Samuels, L.; Jetter, R.; Kunst, L. Identification of the wax ester synthase/acyl-coenzyme A: diacylglycerol acyltransferase WSD1 required for stem wax ester biosynthesis in *Arabidopsis*. *Plant Physiol* **2008**, *148* (1), 97-107.
46. Santos, C. N.; Koffas, M.; Stephanopoulos, G. Optimization of a heterologous pathway for the production of flavonoids from glucose. *Metab Eng* **2011**, *13* (4), 392-400.
47. Tang, X.; Tan, Y.; Zhu, H.; Zhao, K.; Shen, W. Microbial conversion of glycerol to 1,3-propanediol by an engineered strain of *Escherichia coli*. *Appl Environ Microbiol* **2009**, *75* (6), 1628-34.
48. Mei, Y. Z.; Liu, R. X.; Wang, D. P.; Wang, X.; Dai, C. C. Biocatalysis and biotransformation of resveratrol in microorganisms. *Biotechnol Lett* **2015**, *37* (1), 9-18.
49. Cirino, P. C.; Chin, J. W.; Ingram, L. O. Engineering *Escherichia coli* for xylitol production from glucose-xylose mixtures. *Biotechnol Bioeng* **2006**, *95* (6), 1167-76.
50. Alifano, P.; Fani, R.; Lio, P.; Lazcano, A.; Bazzicalupo, M.; Carlomagno, M. S.; Bruni, C. B. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol Rev* **1996**, *60* (1), 44-69.

51. Boghigian, B. A.; Zhang, H.; Pfeifer, B. A. Multi-factorial engineering of heterologous polyketide production in *Escherichia coli* reveals complex pathway interactions. *Biotechnol Bioeng* **2011**, *108* (6), 1360-71.
52. Feher, T.; Planson, A. G.; Carbonell, P.; Fernandez-Castane, A.; Grigoras, I.; Dariy, E.; Perret, A.; Faulon, J. L. Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering. *Biotechnol J* **2014**, *9* (11), 1446-57.
53. Osterhout, R. E., et al. Methods for enhancing microbial production of specific length fatty alcohols in the presence of methanol. *Google Patents* **2016**.
54. 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* (37), 28593-8.
55. 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* (12), 3877-86.
56. Miyahisa, I.; Kaneko, M.; Funa, N.; Kawasaki, H.; Kojima, H.; Ohnishi, Y.; Horinouchi, S. Efficient production of (2S)-flavanones by *Escherichia coli* containing an artificial biosynthetic gene cluster. *Appl Microbiol Biotechnol* **2005**, *68* (4), 498-504.
57. Kaneko, M., et al. Heterologous production of flavanones in *Escherichia coli*: potential for combinatorial biosynthesis of flavonoids in bacteria. *J Ind Microbiol Biotechnol* **2003**, *30* (8), 456-61.
58. Jeon, E., et al. Development of a *Saccharomyces cerevisiae* strain for the production of 1, 2-propanediol by gene manipulation. *Enzyme and Microbial Technology*. *Enzyme and Microbial Technology* **2009**, *45* (1), 42-47.
59. Ro, D. K.; Paradise, E. M.; Ouellet, M.; Fisher, K. J.; Newman, K. L.; Ndungu, J. M.; Ho, K. A.; Eachus, R. A.; Ham, T. S.; Kirby, J.; Chang, M. C.; Withers, S. T.; Shiba, Y.; Sarpong, R.; Keasling, J. D. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **2006**, *440* (7086), 940-3.