Metabolic Engineering 13 (2011) 435–444

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/ymben

Probabilistic pathway construction

Mona Yousofshahi [a], Kyongbum Lee [b], Soha Hassoun [a,*]

[a] Department of Computer Science, 161 College Avenue, Tufts University, Medford, MA 02155, USA
[b] Department of Chemical and Biological Engineering, Tufts University, Medford, MA, USA

ARTICLE INFO

Article history:
Received 15 October 2010
Received in revised form
13 January 2011
Accepted 25 January 2011
Available online 1 February 2011

Keywords:
Pathway construction
Pathway inference
Connectivity
Probabilistic synthesis
Yield diversity

ABSTRACT

Expression of novel synthesis pathways in host organisms amenable to genetic manipulations has emerged as an attractive metabolic engineering strategy to overproduce natural products, biofuels, biopolymers and other commercially useful metabolites. We present a pathway construction algorithm for identifying viable synthesis pathways compatible with balanced cell growth. Rather than exhaustive exploration, we investigate probabilistic selection of reactions to construct the pathways. Three different selection schemes are investigated for the selection of reactions: high metabolite connectivity, low connectivity and uniformly random. For all case studies, which involved a diverse set of target metabolites, the uniformly random selection scheme resulted in the highest average maximum yield. When compared to an exhaustive search enumerating all possible reaction routes, our probabilistic algorithm returned nearly identical distributions of yields, while requiring far less computing time (minutes vs. years). The pathways identified by our algorithm have previously been confirmed in the literature as viable, high-yield synthesis routes. Prospectively, our algorithm could facilitate the design of novel, non-native synthesis routes by efficiently exploring the diversity of biochemical transformations in nature.

1. Introduction

Metabolic engineering of non-native synthesis pathways in microbial hosts has shown promise in the production or overproduction of commercially useful biomolecules, including polyesters (Aldor and Keasling, 2003), building blocks for industrial polymers (Nakamura and Whited, 2003), biofuels (Steen et al., 2010), and therapeutic natural products derived from isoprenoids (Martin et al., 2003; Pitera et al., 2007; Watts et al., 2005), polyketides (Peiru et al., 2005; Pfeifer et al., 2001), and non-ribosomal peptides (Watts et al., 2005). Isolation of these molecules from naturally occurring organisms generally suffers from low yield and can place a large environmental burden (Pitera et al., 2007). Soil dwelling microorganisms that harbor the biosynthetic enzymes for isoprenoids, polyketides and other natural product molecules typically exhibit slow growth compared to industrial workhorse organisms such as *Escherichia coli* and yeast. One promising way to address yield and growth limitations is to harness the biosynthetic capability of niche organisms into technically amenable, fast growing organisms (Pfeifer and Khosla, 2001).

In some cases, a choice for the synthesis pathway may be obvious. For example, there is only one known pathway for biosynthesis of 1,3-propanediol from glycerol (Nakamura and Whited, 2003). This pathway consists of two reactions, each catalyzed by a singular enzyme. More generally, the number of alternative pathways for a given product may be too large for experimental exploration, especially if the goal is to exploit the diversity of metabolic enzymes across many different organisms. To date, more than 1000 prokaryotic genomes have been fully sequenced and annotated. Partial or draft genomes are available for more than 6000 species. The total number of reactions listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa et al., 2006, 2010) currently exceeds 8000. In this light, computational approaches are warranted to analyze the growing number of possible metabolic and biosynthetic enzyme combinations as candidate pathways for heterologous synthesis of biomolecules.

Due to the combinatorial nature of the problem, an exhaustive search for candidate pathways is impractical even for computational approaches. To routinely analyze ever-growing and continuously updated genome-scale databases, an effective search strategy needs to address several issues. Enzymes need to be selected from a large, multi-organism database such as KEGG, MetaCyc (Caspi et al., 2008) or SEED (Overbeek et al., 2005) to form a logical reaction sequence, mapping the final product molecule to one or more reactant metabolites in the host organism. This selection process needs to take into account not only the main reactants, but also reducing equivalents and other cofactors. In the likely event that a large number of candidate pathways

* Corresponding author. Fax: +1 617 627 2227.
E-mail address: soha@cs.tufts.edu (S. Hassoun).

1096-7176/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymben.2011.01.006

have been identified, the computational analysis needs to evaluate these pathway based on a performance metric such as maximal predicted yield. The evaluation needs to also assess whether the introduction of the synthesis pathway will negatively impact the host organism's capacity for balanced growth (Feist et al., 2010).

Over the last several years, a number of heuristic approaches have been developed to predict novel pathways for degradation of xenobiotics (Moriya et al., 2010) or biosynthesis of native and non-native compounds (McShan et al., 2003; Moriya et al., 2010; Pharkya et al., 2004; Pitkänen et al., 2009). One such approach, PathMiner, seeks to build pathways that minimize the biochemical transformation cost (McShan et al., 2003). This heuristic favors reactions involving the addition of smaller functional groups, which can select against canonical modifications such as phosphorylation. PathPred is another method to construct plausible reaction pathways based on chemical structure transformation patterns of small molecules (Moriya et al., 2010). PathPred specifically exploits the KEGG RPAIR database, which contains transformation patterns for substrate–product pairs (reactant pairs) of known enzymatic reactions. The patterns are described by atom type changes at the reaction center atom and its neighboring atoms. A key advantage of PathPred is that it generates plausible pathways even when no matching compound is found for the queried molecule by utilizing pattern matches reflecting generalized reactions shared among structurally related compounds. The drawback is that the patterns need to be manually cured. OptStrain uses mixed integer programming to identify stoichiometrically balanced pathways by adding or deleting reactions to the host metabolic network (Pharkya et al., 2004). A key advantage of this approach is to couple the selection of reactions with the ranking of the synthesis pathways in terms of theoretical yields. Success of the optimization however critically depends on thoroughly pre-processing the database, which remains a non-trivial task.

There currently is a lack of data and consensus on the best pathway scoring methods. The number of pathway steps does not necessarily correlate with yield or the implementation practicality (Martin et al., 2009). Another metric for ranking the non-native pathway is metabolic burden which computes the reduction in the growth rate as a result of added reactions (Rodrigo et al., 2008). Another ranking strategy is the thermodynamic feasibility which tries to compute the change in the Gibbs free energy of the reaction along the pathways by using a group contribution method (Hatzimanikatis et al., 2005).

We present a novel method for constructing synthesis pathways using a graph-based probabilistic-search approach. Our approach is based on searching the KEGG database for pathways and using flux balance analysis (FBA) (Varma and Palsson, 1994) to rank the constructed pathways. The main challenge in this approach is to avoid exhaustive enumeration of all possible pathways as it yields an intractable number of pathways. Accordingly, we propose a probabilistic pathway construction method and we investigate three different selection criteria to mine the KEGG database in search of a synthesis pathway.

2. Methods

2.1. Pathway construction

We develop a graph-based, probabilistic search technique of the KEGG database to identify *non-native* synthesis pathways for a given product metabolite. We define a non-native synthesis pathway as a sequence of non-native reactions beginning with any native metabolite and ending with the specified product metabolite. The product metabolite may or may not be a native metabolite. Pathways are constructed as a graph, specifically a tree, by adding metabolite nodes and reaction edges selected from the KEGG database. The KEGG database was chosen for its breadth of coverage of metabolic pathways across many organisms.

Tree construction proceeds recursively, starting from the target metabolite, *i.e.* synthesis product, as the root of the tree (Fig. 1). A single reaction is selected from a list of candidate reactions in the KEGG database that involve the target metabolite as a main product. Selection occurs probabilistically based on a weighting scheme determined by the connectivity of the candidate reactions' metabolites (see Section 2.2). The type of selection scheme is passed to the algorithm as a free parameter. The selected reaction is then added to the tree and represented by an edge. This edge expands the tree by attaching new nodes representing the reactant metabolites and cofactors of the selected reaction. The construction thus proceeds in a depth-first fashion. Each of these nodes is a new root for the recursion, unless the corresponding metabolite or cofactor is already present in the host organism or was previously added to the tree. Details of the algorithm are provided in Fig. 2.

Because there is a practical limit to the number of heterologous genes that can be inserted into a typical host organism such as *E. coli* (Peiru et al., 2005), we set a limit on the number of reactions that can be used to construct a pathway. The length limit is thus used to obtain candidate pathways of practical length, rather than to rank-order or otherwise evaluate pathway quality. In the present study, the length limit was set to 23 reactions, which reflects state-of-the-art with respect to the number of simultaneous gene insertions (Peiru et al., 2005). When the addition of a reaction to the tree violates this limit, the search algorithm backtracks and proceeds by adding to the tree another reaction that has not been previously explored, effectively exploring an alternative pathway. If none of these alternative routes satisfy the pathway length limit, the algorithm further backtracks and continues from there. The algorithm

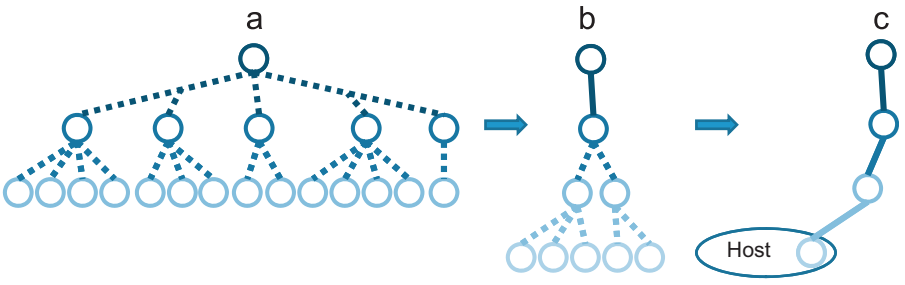

Fig. 1. Schematic illustration of the probabilistic search. The dashed and solid lines show the possible routes and selected reactions, respectively. The tree expansion terminates when a metabolite found that is native to the host network: (a) all possible reaction choices to generate a target metabolite, two reactions away from the target; (b) only one reaction is explored in depth-first fashion and (c) recursive exploration terminates at a metabolite within the host network.

```
Algorithm: Probabilistic Pathway Construction
procedure PROB_PATHWAY_CONSTRUCTION (in target metabolite, in selection scheme, out Pathway)
begin
    Call CONSTRUCT_PATH (target metabolite, selection scheme, pathway)
    Perform flux balance analysis on pathway
    if the calculated yield obtained from pathway is less than that of the native pathway
        pathway ← NULL
    end if
end

procedure CONSTRUCT_PATH (in metabolite, in selection scheme, out pathway)
begin
    if pathway length is greater than length limit
        pathway ← NULL
        return
    end if
    for each reaction r in KEGG database containing metabolite
        if r already exists in the pathway
            rWeighting ← 0
        else if
            Set rWeighting based on the selection scheme
        end if
    end for
    Randomly select a reaction based on rWeighting
    Add the selected reaction to pathway
    for each reactant metabolite, m, of the selected reaction
        if m is already in the host or in pathway
            continue
        else if
            Call CONSTRUCT_PATH (m, selection scheme, pathway)
        end if
    end for
end
```

Fig. 2. Pseudo-code for the probabilistic pathway construction algorithm. Pathways are constructed recursively starting from the target metabolite which is assigned as the root node of the tree. The tree is expanded at each recursion by adding an edge which represents a randomly selected reaction among all candidate reactions linked to the nodal metabolite. The constructed pathway is evaluated by calculating the maximum yield of the target metabolite using flux balance analysis.

finishes when all permitted-length branches of the tree terminate in a metabolite that is native to the host organism. Due to the probabilistic nature of selecting the reactions, the completed tree does not exhaustively enumerate all possible pathways. Rather, each tree represents a single pathway from the target metabolite to one or more required reactant metabolites (including cofactors) that are native to the host organism. Therefore, the search is iterated many times to explore a diverse number of possible pathways.

As previously observed (Blum and Kohlbacher, 2008), a small subset of the reactions in the KEGG database are annotated as 'unclear' and/or lack corresponding enzyme commission number entries. Such reactions were excluded from the search.

To evaluate the effectiveness of our probabilistic pathway construction approach, we compare the probabilistic searches against an exhaustive search, which constructs all possible pathways in the form of a single tree. Tree construction proceeds recursively, similar to the probabilistic search, except that the algorithm adds all of the possible pathways. The output of the search is thus a set of pathways, rather than a single pathway, that satisfies the length limit and terminates at a metabolite in the host. For the tree shown in Fig. 1a, the exhaustive search recursively explores *all* possible additions to the tree. Due to the prohibitive computational cost associated with exhaustive search, the pathway length limit was set to 10 reactions (as opposed to 23, which is the limit set for the probabilistic search).

2.2. Probabilistic reaction selection

We explore three different selection schemes based on metabolite connectivity of candidate reactions: high-degree connectivity, low-degree connectivity and uniform. Here, degree connectivity refers to the number of reactions in which a metabolite participates. The results of the three different selection schemes are compared based on the likelihood of identifying the pathway with the highest predicted yield.

2.2.1. High connectivity

In this scheme, we use weighted probabilities to bias the selection in favor of reactions involving high-degree metabolites. It has been observed that scale-free networks include hub nodes of high degree through which lower degree nodes connect (Barabási and Albert, 1999). For example, if A, B and C are three

metabolites with A having a high-degree connectivity and B and C having low degrees of connectivity, a path from B to C is likely to proceed through A rather than directly. To verify that the metabolites in the KEGG multi-organism database constitute the nodes of a scale-free network, we characterized the degree distribution by counting the number of times each metabolite participates in a distinct reaction. This analysis did not consider the directionality of the reactions, as most of the reactions are reversible. A log-scale histogram (Fig. 3) showed that the degree distribution indeed followed a power law similar to other evolved, scale-free networks (Barabasi, 2009) with a scaling exponent value of −2.04. Motivated by this connectivity property of the KEGG database, we weighted the selection probabilities of candidate reactions to favor pathways whose intermediates are hub metabolites.

The probability of selecting a reaction is proportional to its relative weight normalized by the sum of the weights of all reactions. Mathematically put, $Prob(selecting\ R_i) = W_{R_i} / \sum W_{R_j}$. As an example, consider the hypothetical reactions shown in Fig. 4. Metabolites B through E have the following degree connectivity values: deg(B)=4, deg(C)=3, deg(D)=2 and deg(E)=1. The weights for the reactions are W_{R1}=min(4, 3)=3, W_{R2}=2 and W_{R3}=min(4, 1)=1. In this example, the probabilities of selecting reactions 1, 2 and 3 are 0.5, 0.33 and 0.17, respectively.

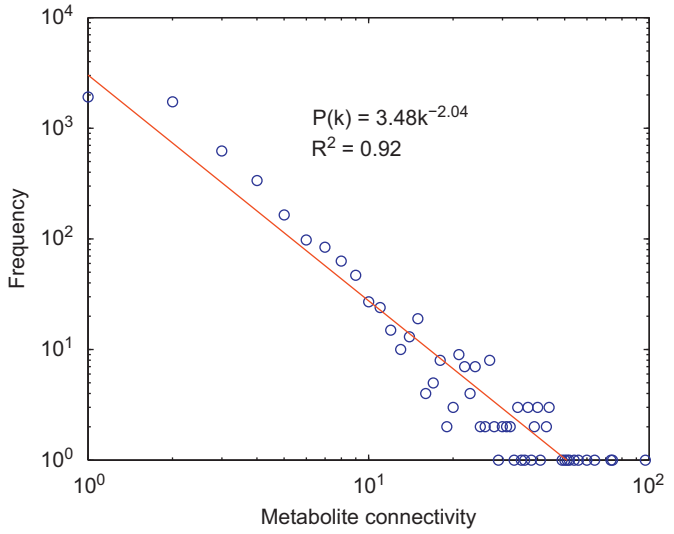

Fig. 3. Degree connectivity distribution of metabolites in the KEGG database exhibiting a power-law distribution in which the probability of finding a metabolite with connectivity k is proportional to $k^{-2.04}$. The blue circles depict the actual distribution and the red line represents the fitted power-law distribution.

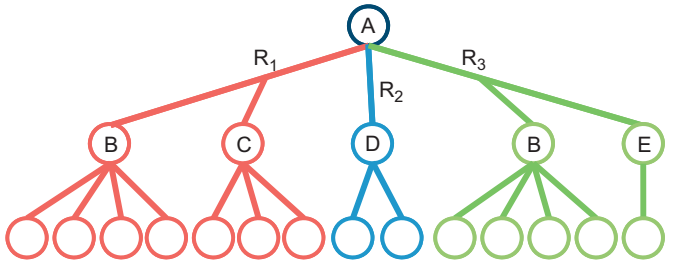

Fig. 4. Metabolite connectivity-based weighting scheme. The nodes and edges represent metabolites and reactions, respectively. The three reactions R1, R2 and R3 producing the metabolite A are defined as follows. R_1: A⇔B+C, R_2: A⇔D and R_3: A⇔B+E. Based on these definitions, the algorithm assigns weights of 3, 2 and 1 to reactions R1, R2 and R3, in order.

It should be noted that a clear distinction between cofactors and main reactants is not always possible without manually inspecting the reaction definition. One way to discriminate cofactors is based on the degree connectivity, which is generally higher than other metabolites. Therefore, we determine the weight of a reaction based on the metabolite with the lowest degree connectivity as calculated by the 'min' function in the formulas above. Also, all of the metabolites in a reaction are *and*-related, *i.e.* different pathway branches should be constructed for all of them.

2.2.2. Low connectivity

In this scheme, we bias the selection in favor of reactions involving low-degree metabolites. The idea of identifying pathways using low-connectivity metabolites has been used previously to infer meaningful pathways in biochemical networks. One such pathway identification algorithm is Metabolic PathFinding, which first assigns metabolites a weight equal to their connectivity and then performs a search for the path with the smallest cumulative weight, thus reducing the likelihood of including currency metabolites such as ATP or H_2O (Croes et al., 2005). Another algorithm, MetaRoute, assigns a modified weight to each reaction based on metabolite degree and then performs a weighted path search to compute the first k-shortest paths between two given metabolites (Blum and Kohlbacher, 2008). Metabolic PathFinding, MetaRoute, and path-pruning methods (Gerlee et al., 2009) are motivated by the idea that low-connectivity metabolites define the major pathways of carbon (or nitrogen) transfer in biochemical networks. In the present study, the low-connectivity selection probabilities for the reactions are calculated similar to the high-connectivity scheme, except that the inverse of the smallest metabolite degree in a reaction is used as the weight for the reaction.

2.2.3. Uniform

Finally, we investigate a selection scheme where each reaction is assigned the same weight. This scheme does not favor either high- or low-degree metabolites as pathway intermediates, and thus should return the most diverse set of pathways.

2.3. Yield calculation

We evaluate each pathway by calculating the maximum yield of the desired product, subject to constraints, using flux balance analysis (FBA) (Becker et al., 2007). In the present study, the maximum yield is used as an overall performance metric of the entire synthesis pathway. The yield also takes into account several (but not all) important biochemical and biophysical constraints of the host organism. Indeed, other pathway construction algorithms have relied on metrics such as thermodynamic favorability and structural similarity of reaction steps (Cho et al., 2010) to prune the search space. In the present study, the focus is on exploring the diversity of possible synthesis pathways, which proceeds independently from the evaluation of the pathways.

As the base model for FBA, we used a genome-scale model of *E. coli* metabolism (iAF1260) (Feist et al., 2007). Reactions selected by the search algorithm to constitute a plausible pathway were then added to the base model to generate the modified strain model. Upper and lower flux bounds for the added reactions were set to 1000 and −1000 mmol/gDW/h, respectively. All other constraints were kept at the same default values of the base model as described in Feist et al. (2007). In brief, the glucose uptake upper and lower bounds were 1000 and −8 mmol/gDW/h, respectively, oxygen uptake bounds were 1000 and −18.5 mmol/gDW/h and

the ATP maintenance requirement was set to 8.39 mmol/gDW/h (Feist et al., 2007). The FBA objective was to maximize the flux forming the desired product subject to the constraint that the modified host strain produces at least 80% of the wild-type biomass yield. Pathways leading to zero product fluxes were considered non-viable. Viable pathways were rank ordered according to the maximum product yield. For comparisons with literature values, product yields were expressed as fluxes normalized by the corresponding glucose uptake flux or dry cell weight (DCW).

3. Results and discussion

To analyze the effectiveness of our algorithm, we examined the synthesis of several native and non-native metabolites, which have previously been identified as commercially useful targets for overproduction, using *E. coli* as the host organism. Moreover, experimentally determined yield or titer data were available in the published literature, thus providing references for comparison. The test compounds belonged to four groups: precursors for natural products with therapeutic activity (isopentenyl diphosphate and taxa-4,11-diene); alcohols used as building blocks for polymer synthesis and other commercial applications (1,3-propanediol and 2,3-butanediol); a complex carbohydrate precursor for value added chemicals (*myo*-inositol); and lipid biofuels (fatty acid methyl and ethyl esters and triacylglycerol).

The analysis compared the performance of the probabilistic search algorithm for different weighting schemes (uniform, high connectivity, and low connectivity) based on the average yield (defined as the average value of yields obtained from repeated runs) of the synthesis pathways. The probabilistic search with uniform weighting was also compared against an exhaustive search in terms of sampling efficiency as reflected in the yield diversity of the pathways. This comparison also involved an analysis of the computational cost, which showed that the runtime of the exhaustive search increases exponentially with respect to the number of reactions in the *pathway*, whereas the runtime of the probabilistic search scales linearly with the number of reaction in the *database*. Finally, we compared the yield results calculated by the probabilistic and exhaustive searches against experimentally obtained values reported in the literature.

3.1. Yield results for different weighting schemes

Due to the probabilistic nature of our algorithms, meaningful interpretation of the search requires a large number of iterations. To estimate the number of iterations needed to identify viable, high-yield pathways, we executed the probabilistic search for varying numbers of iterations ranging from 100 to 1500 and recorded the maximum product yield obtained for each iteration number. To also determine an average yield, we repeated this process 500 times for each iteration number. The results of this analysis for fatty acid methyl esters are shown in Fig. 5. The *overall* maximum product yield remained constant for all iteration numbers. The *average* maximum product yield increased steadily with the iteration number, and gradually leveled off, reaching a plateau around iteration number 1500. Similar trends were observed for all other test cases (data not shown). These trends suggest that the likelihood of finding a pathway with the overall maximum yield increases with the iteration number, but only up to a point. Once a sufficiently large iteration number has been reached, the likelihood of finding a pathway with the overall maximum yield remains essentially unchanged.

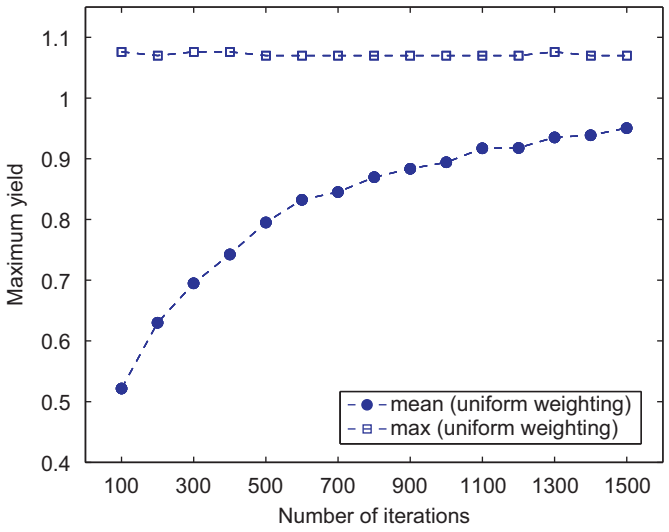

Fig. 5. Dependence of the overall and average maximum yields on the number of iterations for the fatty acid methyl ester test case. At each iteration number, the probabilistic search was repeated 500 times. The overall maximum refers to the largest of the 500 maximum yields for the iteration number calculated using flux balance analysis (FBA). The average maximum yield refers to the arithmetic mean of the 500 FBA yields.

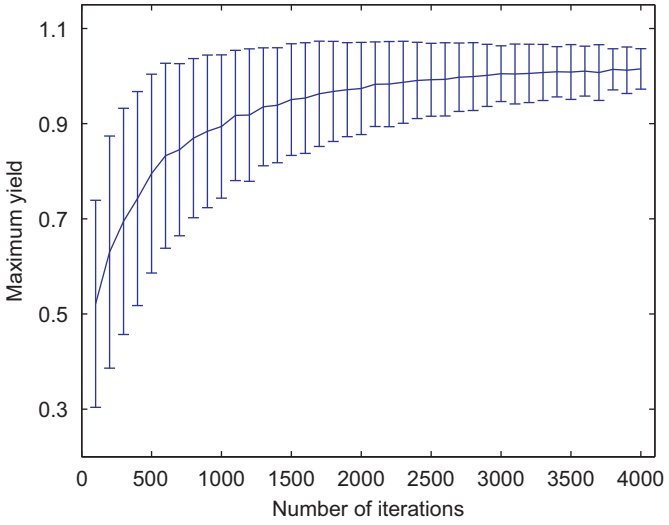

Fig. 6. Average yield (solid line) and standard deviation (error bars) vs. number of iterations for the fatty acid methyl ester test case.

In addition to the overall maximum yield, we also recorded the standard deviation of the maximum yields for the 500 repeats at the various iteration numbers. To more clearly visualize the trend, we plotted the standard deviations for iteration numbers up to 4000 (Fig. 6). The decreasing standard deviations suggest that increasing the iteration number improves the predictability of the search outcomes resulting from repeated runs. Given that we calculate and record the maximum yield for each run in a batch of repeats, the convergence in yield trends toward the overall maximum.

To examine the impact of the weighting scheme on search performance, we repeated the probabilistic search with uniform, low-connectivity and high-connectivity selection of reactions. Comparisons of overall and average maximum yields for fatty acid ethyl esters (FAEEs) are shown in Fig. 7. The uniform weighting scheme consistently outperformed the connectivity-based weighting schemes, as it needed fewer iterations to identify

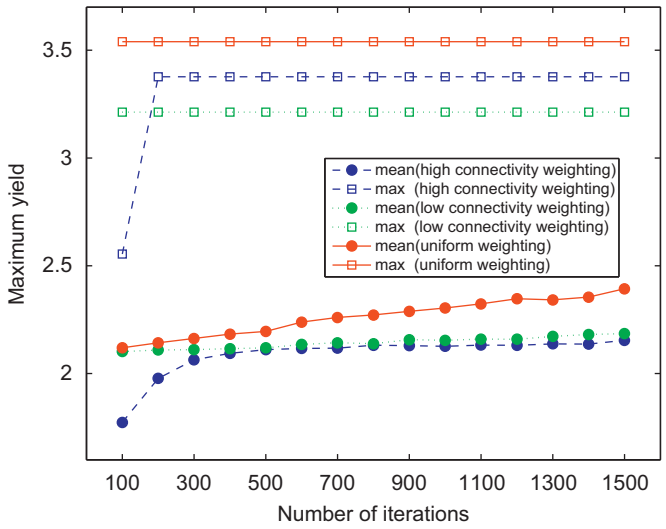

Fig. 7. Yield comparisons for fatty acid ethyl esters obtained using connectivity-based and uniform weighting schemes.

viable, high-yield pathways. This improvement in the search performance suggested that the high-yield pathways involve intermediates with both high *and* low metabolite connectivity. This indeed was the case for fatty acid ethyl esters. The pathway with the highest maximum product yield (3.58 mmol/gDW/h) was identified only when the reactions were selected based on a uniformly random probability. The connectivity values of metabolites in the reaction sequence for this pathway were 3, 8, 6, 44, 7, 24, 11, 22, 7 and 8, in order. The maximum product yield calculated using the high-connectivity weighting scheme was 3.37 mmol/gDW/h. The metabolite connectivity values for this reaction sequence were 3, 8, 6, 44, 7, 24, 11, 22 and 12, in order, differing only slightly from the higher yield pathway at the end of sequence. As a consequence of its bias for reactions involving high-degree metabolites (in this case, bias for the reaction involving a product with degree 12 over degree 7) the high-connectivity weighting scheme favors the lower yield sequence, whereas the uniform weighting scheme is equally likely to select either sequence.

Similar results were obtained for triacylglycerol, with greater average maximum yields calculated by the uniform weighting scheme compared to the connectivity weighting schemes (Supplementary Fig. 13). In the case of fatty acid methyl esters, the results of the low-connectivity weighting method were similar to those of the uniform weighting method (Supplementary Fig. 12), suggesting that there are some highest-yielding pathways (1.065 mmol/gDW/h) which proceed through metabolites with low-connectivity values. Since the uniform probabilistic method is able to find not only the pathways found by the low-connectivity method but also other pathways with the same yield, which otherwise cannot be identified using low-connectivity weighting due to some high-connectivity metabolites in them, uniform probabilistic method outperforms the low-connectivity probabilistic search in this case. One possible reason the uniform weighting scheme outperforms the connectivity-based schemes could be due to interactions with the host metabolic network. Integrating a high-connectivity pathway with the host could subject many native pathways to competition with the non-native pathway. Likewise, a pathway with low-connectivity metabolites could add to the scarcity of metabolites with one or few native routes of production.

The three weighting schemes returned similar performances for all other test cases (Table 1 and Supplementary Figs. 7–11), where the synthesis pathways did not exhibit significant variations in the obtained yield.

Table 1
Effect of the weighting scheme on search performance.

Metabolite name	Weighting scheme	Normalized average maximum yield
Isopentenyl diphosphate	Uniform	1
	High-connectivity	1
	Low-connectivity	1
myo-Inositol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
Taxa-4(5),11(12)-diene	Uniform	1
	High-connectivity	1
	Low-connectivity	1
1,3-Propanediol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
(*R*,*R*)-2,3-Butanediol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
Fatty acid ethyl esters	Uniform	0.63
	High-connectivity	0.60
	Low-connectivity	0.60
Fatty acid methyl esters	Uniform	0.84
	High-connectivity	0.77
	Low-connectivity	0.84
Triacylglycerol	Uniform	0.55
	High-connectivity	0.50
	Low-connectivity	0.41

The normalized average maximum yield was calculated by dividing the average maximum yield with the overall maximum yield. Average and overall maximum yields were determined for runs of 1000 iterations repeated 500 times as described in the text.

3.2. Sampling efficiency

For every test case, the probabilistic search with the uniform weighting scheme identified viable synthesis pathways supporting a non-zero yield of the target product and at least 80% of the maximum wild-type biomass flux. A summary of the search results is shown in Table 2. In general, the exhaustive search returned a greater number of pathways than the probabilistic search, despite the lower length limit (10 vs. 23 reactions). Equally small numbers of pathways were identified for taxadiene and 1,3-propanediol (2 and 1, respectively), presumably reflecting the involvement of singular reactions. In the cases of isopentenyl diphosphate (IPP) and *myo*-inositol, the number of pathways identified by the probabilistic search was greater than the exhaustive search. In the cases of the lipids, the exhaustive search generated a larger number of pathways than the probabilistic search, with the fold differences in the number of pathways ranging from 30 to 58.

Despite the differences in the total number of pathways, the maximal yields calculated by the probabilistic (uniform weighting) and exhaustive searches were identical in all test cases, except for fatty acid methyl esters, where the difference was less than 1%.

For the sake of completeness, we also compared the results obtained from the uniform probabilistic and exhaustive search with length limit of 10 reactions (Supplementary Table 1). For some of the test cases, the number of identified pathways using the probabilistic method decreased. However, the maximum calculated fluxes were the same, because the search found

Table 2
Summary of search results obtained using uniform probabilistic and exhaustive methods.

Metabolite name	Native yield/rate	Method	No. of pathways	Max. yield/rate	Pathway lengths (pathways with yields larger than 95% of the max. yield)
Isopentenyl diphosphate[a]	301.13 mg/gDW/h	Uniform probabilistic search Exhaustive search Literature	11 9 NA	314.24 mg/gDW/h 314.24 mg/gDW/h 27.4 g/L amorphadiene (Newman et al., 2006) (1.95 mg/gDW/h)	4, 5, 8, 10, 19 4, 5, 8, 10 NA
Myo-inositol[b]	0 mol/molglucose	Uniform probabilistic search Exhaustive search Literature	71 42 NA	0.2 g/g glucose 0.2 g/g glucose 0.23 g/L (Moon et al., 2009) (0.08 g/g glucose)	2, 7, 8, 9, 10, 11, 13 2, 7, 8, 9, 10 NA
Taxadiene[c]	Non-native	Uniform probabilistic search Exhaustive search Literature	2 2 NA	0.06 g/g glucose 0.06 g/g glucose 1.3 mg/L (Huang et al., 2001) (0.06 mg/g glucose)	2, 5 2, 5 NA
1,3-Propanediol	Non-native	Uniform probabilistic search Exhaustive search Literature	1 1 NA	2.19 mmol/gDW/h 2.19 mmol/gDW/h 2.3 mmol/gDW/h (Burgard et al., 2003)	2 2 NA
2,3-Butanediol	Non-native	Uniform probabilistic search Exhaustive search Literature	9 9 NA	0.11 g/g glucose 0.11 g/g glucose 0.31 g/g glucose (Yan et al., 2009)	2, 3, 4 2, 3, 4 NA
Fatty acid ethyl esters[d]	Non-native	Uniform probabilistic search Exhaustive search Literature	19 1092 NA	3.58 mmol/gDW/h 3.58 mmol/gDW/h 647 mg/L (Steen et al., 2010) (0.34 g/g glucose)	8, 9, 10, 11 8, 9, 10 NA
Fatty acid methyl esters[e]	Non-native	Uniform probabilistic search Exhaustive search Literature	45 1353 NA	1.07 mmol/gDW/h 1.08 mmol/gDW/h 0.3 g/gDW (Jakobsen et al., 2008)	7, 8, 9 7, 8, 9 NA
Triacylglycerol	Non-native	Uniform probabilistic search Exhaustive search Literature	51 2900 NA	1.65 mmol/gDW/h 1.65 mmol/gDW/h 0.06 mol/mol glucose (Famili and Schilling, 2006)	8, 9 8, 9 NA

For comparisons with FBA calculations, the reported titer values were converted as follows:
[a]IPP: (27.4 g/L/88 gDW/L)/160 h=1.95 mg/gDW/h.
[b]*myo*-Inositol: 10−7 g/L=3 g/L (glucose consumed); 0.23 g/L/3 g/L=0.08 g/g glucose.
[c]Taxadiene: 1.3 mg/L/20 g/L=0.06 mg/g glucose (We assumed all glucose in LB medium is consumed).
[d]FAEE: 647 mg/L/2 g/L=0.34 g/g glucose.
[e]FAME: The yield was reported for thraustochytrid *Aurantiochytrium* sp. strain T66 (as opposed to *E. coli*).
Results are shown for 1000 iterations of the probabilistic search and one single run of the exhaustive search for each test case. The total number of pathways includes only those with specific productivities greater than the wild-type organism represented by the base model. The maximal yield refers to the pathway with the highest ratio of product flux to biomass flux. Product fluxes were calculated using FBA with the constraint that the biomass flux exceeds 80% of the wild-type.

pathways with lengths less than 10 reactions and the same highest yield (Table 2).

The similarity of the maximal yields suggested that the two search methods identified at least partially overlapping sets of pathways. To evaluate the extent of the overlap, we compared the yield distributions resulting from the two methods for each test case (Supplementary Figs. 1–6). In the case of fatty acid methyl esters (Fig. 8a,b), the yield distributions were essentially identical, even though the total number of pathways returned by the probabilistic search was less than 4% of the total returned by the exhaustive search. Similar trends were observed for all other test cases, suggesting that the probabilistic search representatively sampled the space of possible pathways in the KEGG database.

We also compared the yield distributions for fatty acid methyl esters obtained from different weighting schemes (Figs. 8a,c,d). In all three cases, similar patterns were generated, suggesting that there is no correlation between connectivity and yield distribution. In conclusion, using uniform weighting proves superior in finding highest-yielding pathways and does not miss some potential high-yield pathways otherwise undiscovered using other schemes.

3.3. Runtime comparisons

To examine the computational efficiency of the probabilistic search, we compared its runtime against the exhaustive search. The runtime for the exhaustive search grew exponentially with the number of reactions used in the pathway (Fig. 9), rendering the algorithm intractable for longer pathways. For example, a single run of the exhaustive search with a pathway length limit of 23 was projected to require a runtime exceeding 400 years on a workstation with four Quad-Core 2.3 GHz processors (AMD Optron 8356) and 64 GB of physical memory. The runtime for the probabilistic search shows a linear dependence on the number of reactions in the KEGG database. With the identical pathway length limit of 23, 1000 iterations of the probabilistic search required a runtime of 6 min on the same workstation.

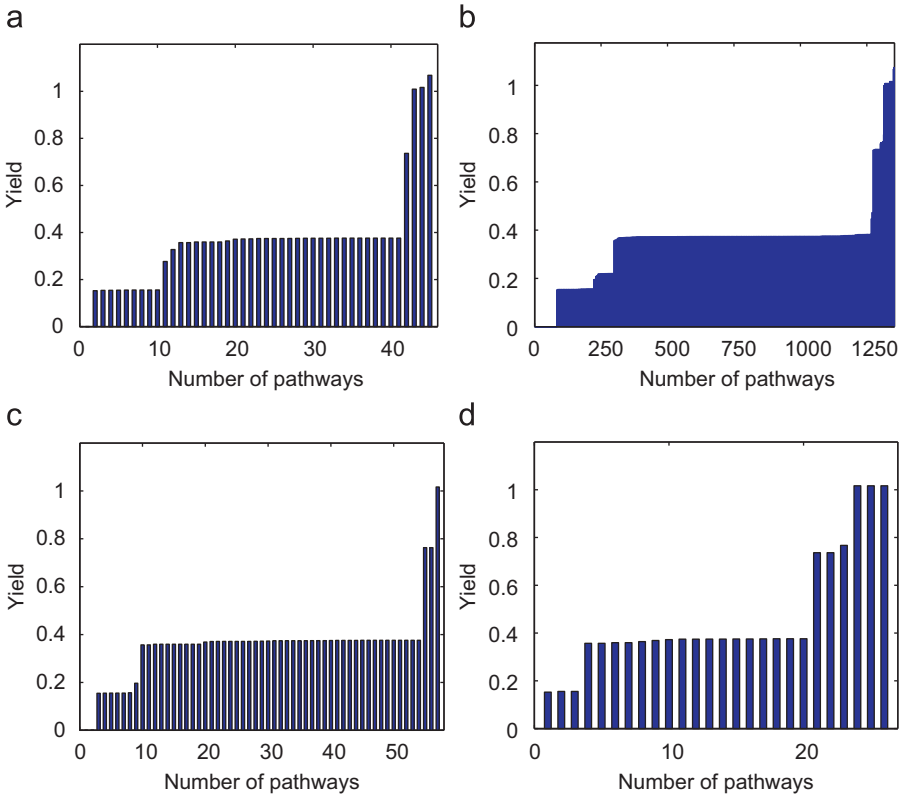

Fig. 8. Yield distributions for fatty acid methyl esters obtained using uniform probabilistic search (a), exhaustive search (b), high-connectivity probabilistic search (c) and low-connectivity probabilistic search (d). The histograms for the probabilistic searches represent the cumulative results of 1000 iterations. The histogram for the exhaustive search reflects a single run.

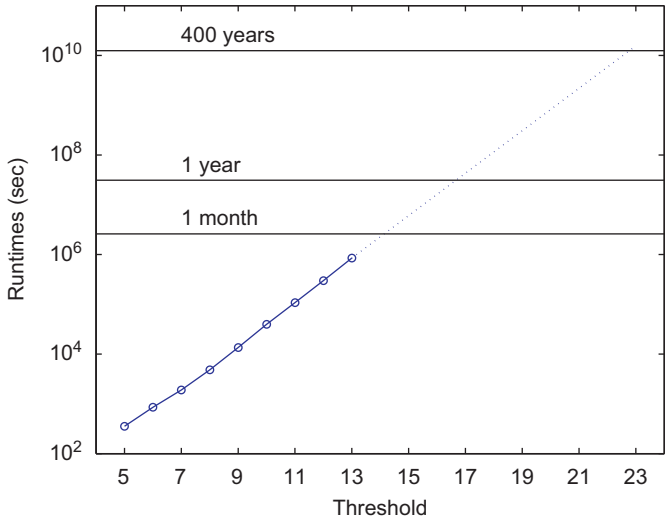

Fig. 9. Runtimes of the exhaustive search as a function of pathway length limit. Data shown are for the fatty acid methyl esters test case. Symbols indicate recorded values from simulations performed on a four Quad-Core 2.3 GHz AMD Optron 8356 with 64 GB of physical memory. The dashed line extrapolates the runtimes for length limits up to 23 reactions, which is the maximum number of steps allowed for the probabilistic search. Extrapolation is performed based on linear regression of log-transformed runtime data against length limit.

3.4. Experimental support

We next examined the quality of the results from the probabilistic search (uniform weighting) by comparing the FBA calculations against published data on experimentally obtained yields. In general, direct comparisons of the predicted yields against published values were not possible. The immediate output of FBA is specific productivity (rate of target production normalized to dry cell weight) during balanced growth, which can then be normalized to glucose uptake or another flux. Typically reported values are volumetric productivity (product titer and cell density) measurements obtained from shake flask or fed-batch experiments (Newman et al., 2006; Steen et al., 2010). In some cases, quantitative details regarding the carbon source and other culture parameters needed to derive the specific productivity or yield were not reported. In such cases, representative values were used based on a survey of the literature. For example, we used the default value of 0.4 absorbance unit per gDCW/L to convert the reported OD 600 readings into cell concentrations. The details of converting the reported titer values varied from case to case, and are therefore described separately for each case in the caption of Table 2.

In the case of IPP, the search algorithm identified 11 distinct pathways, all of which led to higher maximal yields compared to the native pathway. Comparisons with published data suggested that the predicted maximal yield is two orders of magnitude greater than previously observed yields. Similarly large differences were found between the FBA calculation and reported values for the other natural product, taxadiene. In the case of the alcohols and *myo*-inositol, the FBA results were of comparable magnitude as previously achieved production rates (Burgard et al., 2003; Moon et al., 2009; Yan et al., 2009). In the case of the lipids, comparisons with published values were further confounded by the generic representation of hydrocarbon side chains and unspecified stoichiometries of the relevant reactions in the KEGG database. Computing mass yields was especially

problematic for synthesis pathways involving chain elongation reactions catalyzed by multi-functional enzyme complexes such as fatty acid synthase.

In addition to yield, we also examined the compositions of the pathways resulting from the probabilistic search. For every test case, the computational search is able to reconstruct viable pathways involving reactions whose insertion has been shown to improve the yield. In the case of IPP, the search results included the mevalonate pathway, which has been shown to be a preferred synthesis route in *E. coli* (Pitera et al., 2007). Two pathways were identified for taxadiene, one of which started with IPP, as previously shown in Huang et al. (2001). The other started with farnesylfarnesylgeraniol and had the same yield. In the case of *myo*-inositol, several pathways started from acetyl-CoA as reported in Na et al. (2010). The search also identified several other pathways that started from D-glucose-6-phosphate, acetaldehyde, pyruvate, propanoyl-CoA or malonyl-CoA. The highest yield belonged to the pathway which started with acetyl-CoA. Other identified pathways had slightly less yields. Only one pathway was identified for 1,3-propanediol, which started from glycerol, consistent with the analysis in a review (Nakamura and Whited, 2003). The synthesis pathways for (*R*,*R*)-2,3-butanediol could begin with a variety of metabolites native to *E. coli*, including (*S*)-2-acetolactate, 3-methyl-2-oxobutanoic acid, (*R*)-2,3-dihydroxy-3-methylbutanoate (all three had the highest yield), thiamin diphosphate and pyruvate. Of these, the pathway starting from pyruvate has already been demonstrated experimentally in *E. coli* (Yan et al., 2009). Pathways producing fatty acid ethyl esters (FAEEs) started with different metabolites, including 1,2-diacyl-*sn*-glycerol with the highest yield of 3.58 mmol/gDW/h, (3*R*)-3-hydroxyacyl-[acyl-carrier protein], acetyl-CoA, phosphatidylethanolamine, 1-acyl-*sn*-glycerol 3-phosphate, 2-acyl-*sn*-glycero-3-phosphoethanolamine and phosphatidate, all with lesser yields. Depending on the length of the desired hydrocarbon side chain, fatty acid methyl esters (FAMEs) could be produced from various glycerolipids and phospholipids such as (3*R*)-3-hydroxyacyl-[acyl-carrier protein] with the highest yield of 1.07 mmol/gDW/h, CDP-diacylglycerol, phosphatidylethanolamine, choline, 1,2-diacyl-*sn*-glycerol, phosphatidate and 1-acyl-*sn*-glycerol 3-phosphate. In the case of triacylglycerol, most pathways involved acyl-CoA (Saha et al., 2006) starting with (3*R*)-3-hydroxyacyl-[acyl-carrier protein] which generated the highest yield of 1.65 mmol/gDW/h, phosphatidate, phosphatidylethanolamine and 1-acyl-*sn*-glycerol 3-phosphate. Other pathways, which did not involve acyl-CoA, began with CDP-diacylglycerol, choline and 1,2-diacyl-*sn*-glycerol.

4. Conclusion

We designed a probabilistic graph-based search algorithm to identify novel, non-native synthesis pathways for metabolite overproduction using heterologous hosts. Importantly, the algorithm considers not only the main reactants, but also the cofactors needed for the biosynthesis. The probabilistic search for pathways is based on uniform weighting and the degree of metabolite connectivity determined from the KEGG database. Results demonstrate that uniform weighting outperforms the connectivity weighting in terms of average maximum yield with the same number of iterations. The probabilistic method is much faster (~minutes, independent of the pathway length) than the exhaustive search algorithm (~years for the longest pathways) and takes into account all possible pathways in a probabilistic way.

Using this method, we were able to reproduce experimentally obtained pathways reported in the literature as in the cases of IPP, *myo*-inositol, taxadiene, 1,3-propanediol, (*R*,*R*)-2,3-butanediol, fatty acid ethyl and methyl esters and triacylglycerol. The corresponding maximum yields are also comparable with those reported in the literature. We also compared the yield results of this method with those of an exhaustive search method which looks for all possible pathways leading to the target metabolite production in a higher rate. Our calculations show that for reasonable number of iterations (~1000 with a runtime on the order of minutes) the results of both methods are comparable.

Our approach has the potential to be integrated with ensemble modeling (Contador et al., 2009) to develop a kinetic model or to be extended based on genomic-scale mapping (Warnecke et al., 2010). The approach presented in this paper does not consider the issue of host integration. It is possible that a high-yield pathway identified by the probabilistic search generates potentially undesirable byproducts or other side effects. It is also possible that the heterologous genes needed to functionally express the synthesis pathway interact with the host genome through regulatory mechanisms (Kim and Reed, 2010). We intend to address this issue in our future work by investigating the interactions between the native and added genes.

Acknowledgment

This work was supported by the National Science Foundation under Grant no. 0829899.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at doi:10.1016/j.ymben.2011.01.006.

References

Aldor, I.S., Keasling, J.D., 2003. Process design for microbial plastic factories: metabolic engineering of polyhydroxyalkanoates. Curr. Opin. Biotechnol. 14, 475–483.
Barabasi, A., 2009. Scale-free networks: a decade and beyond. Science 325, 412–413.
Barabási, A., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.
Becker, S.A., Feist, A.M., Mo, M.L., Hannum, G., Palsson, B.O., Herrgard, M.J., 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat. Protocols 2, 727–738.
Blum, T., Kohlbacher, O., 2008. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics 24, 2108–2109.
Burgard, A., Pharkya, P., Maranas, C., 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84, 647–657.
Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P., Karp, P.D, 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 36, D623–D631.
Cho, A., Yun, H., Park, J., Lee, S., Park, S., 2010. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Syst. Biol. 4, 35.
Contador, C.A., Rizk, M.L., Asenjo, J.A., Liao, J.C., 2009. Ensemble modeling for strain development of L-lysine-producing *Escherichia coli*. Metab. Eng. 11, 221–233.
Croes, D., Couche, F., Wodak, S.J., van Helden, J., 2005. Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res. 33, W326–W330.
Famili, I., Schilling, C.H., 2006. Multicellular metabolic models and methods, 20060147899 〈http://www.freepatentsonline.com/y2006/0147899.html〉.
Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.O., 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3.
Feist, A.M., Zielinski, D.C., Orth, J.D., Schellenberger, J., Herrgard, M.J., Palsson, B.Ø., 2010. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. Metab. Eng. 12, 173–186.
Gerlee, P., Lizana, L., Sneppen, K., 2009. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. Bioinformatics 25, 3282–3288.

Hatzimanikatis, V., Li, C., Ionita, J.A., Henry, C.S., Jankowski, M.D., Broadbelt, L.J., 2005. Exploring the diversity of complex metabolic networks. Bioinformatics 21, 1603–1609.

Huang, Q., Roessner, C.A., Croteau, R., Scott, A.I., 2001. Engineering *Escherichia coli* for the synthesis of taxadiene, a key intermediate in the biosynthesis of taxol. Bioorg. Med. Chem. 9, 2237–2242.

Jakobsen, A., Aasen, I., Josefsen, K., Strøm, A., 2008. Accumulation of docosahexaenoic acid-rich lipid in thraustochytrid *Aurantiochytrium* sp. strain T66: effects of N and P starvation and O_2 limitation. Appl. Microbiol. Biotechnol. 80, 297–306.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M., 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–D360.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–357.

Kim, J., Reed, J., 2010. OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst. Biol. 4, 53.

Martin, C.H., Nielsen, D.R., Solomon, K.V., Prather, K.L.J., 2009. Synthetic metabolism: engineering biology at the protein and pathway scales. Chem. Biol. 16, 277–286.

Martin, V.J.J., Piteral, D.J., Withers, S.T., Newman, J.D., Keasling, J.D., 2003. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat. Biotechnol. 21, 796–802.

McShan, D.C., Rao, S., Shah, I., 2003. PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics 19, 1692–1698.

Moon, T.S., Yoon, S., Lanza, A.M., Roy-Mayhew, J.D., Prather, K.L.J., 2009. Production of glucaric acid from a synthetic pathway in recombinant *Escherichia coli*. Appl. Environ. Microbiol. 75, 589–595.

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., Kanehisa, M., 2010. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 38, W138–W143.

Na, D., Kim, T.Y., Lee, S.Y., 2010. Construction and optimization of synthetic pathways in metabolic engineering. Curr. Opin. Microbiol. 13, 363–370.

Nakamura, C.E., Whited, G.M., 2003. Metabolic engineering for the microbial production of 1,3-propanediol. Curr. Opin. Biotechnol. 14, 454–459.

Newman, J.D., Marshall, J., Chang, M., Nowroozi, F., Paradise, E., Pitera, D., Newman, K.L., Keasling, J.D., 2006. High-level production of amorpha-4,11-diene in a two-phase partitioning bioreactor of metabolically engineered *Escherichia coli*. Biotechnol. Bioeng. 95, 684–691.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, N., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V., 2005. The subsystems approach to genome annotation and its use in the project to Annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702.

Peiru, S., Menzella, H.G., Rodriguez, E., Carney, J., Gramajo, H., 2005. Production of the potent antibacterial polyketide Erythromycin C in *Escherichia coli*. Appl. Environ. Microbiol. 71, 2539–2547.

Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E., Khosla, C., 2001. Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. Science 291, 1790.

Pfeifer, B.A., Khosla, C., 2001. Biosynthesis of polyketides in heterologous hosts. Microbiol. Mol. Biol. Rev. 65, 106–118.

Pharkya, P., Burgard, A.P., Maranas, C.D., 2004. OptStrain: a computational framework for redesign of microbial production systems. Genome Res.

Pitera, D.J., Paddon, C.J., Newman, J.D., Keasling, J.D., 2007. Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. Metab. Eng. 9, 193–207.

Pitkänen, E., Jouhten, P., Rousu, J., 2009. Inferring branching pathways in genome-scale metabolic networks. BMC Syst. Biol. 3, 103.

Rodrigo, G., Carrera, J., Prather, K.J., Jaramillo, A., 2008. DESHARKY: automatic design of metabolic pathways for optimal cell growth. Bioinformatics 24, 2554–2556.

Saha, S., Enugutti, B., Rajakumari, S., Rajasekharan, R., 2006. Cytosolic triacylglycerol biosynthetic pathway in oilseeds. Molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. Plant Physiol. 141, 1533–1543.

Steen, E.J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., del Cardayre, S.B., Keasling, J.D., 2010. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. Nature 463, 559–562.

Varma, A., Palsson, B.O., 1994. Metabolic flux balancing: basic concepts. Nat. Biotechnol. 12, 994–998.

Warnecke, T.E., Lynch, M.D., Karimpour-Fard, A., Lipscomb, M.L., Handke, P., Mills, T., Ramey, C.J., Hoang, T., Gill, R.T., 2010. Rapid dissection of a complex phenotype through genomic-scale mapping of fitness altering genes. Metab. Eng. 12, 241–250.

Watts, K.T., Mijts, B.N., Schmidt-Dannert, C., 2005. Current and emerging approaches for natural product biosynthesis in microbial cells. Adv. Synth. Catal.

Yan, Y., Lee, C., Liao, J.C., 2009. Enantioselective synthesis of pure (*R*,*R*)-2,3-butanediol in *Escherichia coli* with stereospecific secondary alcohol dehydrogenases. R. Soc. Chem. Org. Biomol. Chem. 7, 3914–3917.