Bioinformatics, 31(7), 2015, 1016–1024
doi: 10.1093/bioinformatics/btu760
Advance Access Publication Date: 21 November 2014
Original Paper

Systems biology

Efficient searching and annotation of metabolic networks using chemical similarity

Dante A. Pertusi, Andrew E. Stine, Linda J. Broadbelt and Keith E.J. Tyo*

Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA

*To whom correspondence should be addressed.
Associate Editor: Igor Jurisica

Received on July 30, 2014; revised on October 27, 2014; accepted on November 11, 2014

Abstract

Motivation: The urgent need for efficient and sustainable biological production of fuels and high-value chemicals has elicited a wave of *in silico* techniques for identifying promising novel pathways to these compounds in large putative metabolic networks. To date, these approaches have primarily used general graph search algorithms, which are prohibitively slow as putative metabolic networks may exceed 1 million compounds. To alleviate this limitation, we report two methods—SimIndex (SI) and SimZyme—which use chemical similarity of 2D chemical fingerprints to efficiently navigate large metabolic networks and propose enzymatic connections between the constituent nodes. We also report a Byers–Waterman type pathway search algorithm for further paring down pertinent networks.

Results: Benchmarking tests run with SI show it can reduce the number of nodes visited in searching a putative network by 100-fold with a computational time improvement of up to 10^5-fold. Subsequent Byers–Waterman search application further reduces the number of nodes searched by up to 100-fold, while SimZyme demonstrates ~90% accuracy in matching query substrates with enzymes. Using these modules, we have designed and annotated an alternative to the methylerythritol phosphate pathway to produce isopentenyl pyrophosphate with more favorable thermodynamics than the native pathway. These algorithms will have a significant impact on our ability to use large metabolic networks that lack annotation of promiscuous reactions.

Availability and implementation: Python files will be available for download at http://tyolab.northwestern.edu/tools/.

Contact: k-tyo@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolic reaction networks have been the focus of intense study in recent years. Unravelling the topology of metabolic networks has been used to gain a deeper understanding of metabolism, in addition to identifying opportunities for engineering the production of drugs, fuels, and chemicals. For this task, numerous *in silico* utilities, such as metabolic flux analysis (Orth *et al.*, 2010), have been developed to optimize existing pathways, engineer novel pathways and identify drug targets by simulating the effect of remodeling the underlying metabolic network and analyzing its topology (Burgard *et al.*, 2003; Cho *et al.*, 2010; Guimerà *et al.*, 2007; Ranganathan *et al.*, 2010). However, the emergence of evidence that many enzymes exhibit promiscuous activities, acting on several different substrates or carrying out disparate reactions (Humble and Berglund, 2011; Nam *et al.*, 2012; Nobeli *et al.*, 2009) tends to obfuscate the well-defined topology of metabolic networks. The putative nodes and edges contributed by these promiscuous activities to the metabolic network both increase its overall size and number of unannotated edges.

We will focus on two significant challenges that exist in exploiting the functionality of putative metabolic networks with potentially nebulous topology. First, finding pathways in a metabolic network linking two compounds, reactions or enzymes of interest is a time-consuming task using traditional graph-searching algorithms due to the large number of nodes in extant metabolic networks (Yousofshahi *et al.*, 2011). Hence, a more efficient mode of traversing possible pathways to a specific target must be developed. Native metabolic networks, such as that of iAF1260, are composed of $\sim10^3$ compound nodes, with an additional 10^3 reaction nodes (Feist *et al.*, 2007). Comprehensive reaction databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) also have a network-like structure, and contains upwards of 15 000 compounds with ~8700 reactions (Altman *et al.*, 2013). At the largest scale, putative metabolic networks with sizes in excess of 10^7 compounds and reactions can also be generated thanks to several computational tools (Carbonell *et al.*, 2012; Cho *et al.*, 2010; Hatzimanikatis *et al.*, 2005). These algorithms use a library of generalized biochemical transformations and apply them where possible to successive generations of compounds, starting with one or more compounds in the zeroth generation. These computer-generated networks are useful because they elucidate all potential transformations that can be performed on a compound, which can be added to a growing putative metabolic network containing numerous novel pathways to a desired end product. Tools exist that can cap the size of putative networks (Kotera *et al.*, 2014) by prescribing a putative number of intermediate steps, but if a long pathway is predicted yielding networks with 10^7 nodes or more, the networks quickly become infeasible to analyze, limiting our ability to discover and design novel metabolic pathways.

A second challenge in exploiting large metabolic networks is associating a given metabolic reaction with a specific enzyme or set of enzymes capable of carrying out that reaction. Although most biochemistry literature is built on the assumption that one enzyme acts on one substrate, it is known that many similar substrates can often be catalyzed by one particular enzyme (Humble and Berglund, 2011; Nobeli *et al.*, 2009). An example of such a phenomenon is acetate secretion in *Escherichia coli* strains that have had the acetate-producing enzymes Pta/Ack and PoxB knocked out; enzyme promiscuity has been conjectured to be the responsible factor in this and similar situations with other metabolites (Perez-Gil *et al.*, 2012; Phue *et al.*, 2010). This significant problem most likely occurs on the genome scale as it is believed that ~40% of enzymes have promiscuous function (Nam *et al.*, 2012). This is most likely exacerbated in engineered cells, where enzyme and metabolite concentrations may be far away from standard concentrations found through evolution.

In addition, enzymes with similar sequences may catalyze reactions on different substrates or have different activities for the same set of substrates, thereby leading to improperly annotated genes. In general, enzymes whose activity extends to substrates other than their 'native' substrates will act on compounds with a similar structure or substructure. Determining which enzyme can fill these roles, then, requires search algorithms that consider the molecular structure of the substrates.

Searching these networks with a basic depth- or breadth-first search algorithm is time intensive and impractical for moderately long pathways, and several algorithms have been developed to explore the space of known reactions in a probabilistic manner (Yousofshahi *et al.*, 2011; Rodrigo *et al.*, 2008). In the past, much success has been reported in using 2D chemical fingerprints in exploring the substrate and inhibitory promiscuities of cytochrome P450s (Cheng *et al.*, 2011; Nath *et al.*, 2010; Terfloth *et al.*, 2007) and in assessing the verisimilitude of enzyme-substrate complexes (Faulon *et al.*, 2008). We present here a molecular structure-biased search algorithm based on SMARTS-type (*Daylight Theory Manual*, 2011) (Supplementary Fig. S1) fingerprint similarity measured using the Tanimoto coefficient (Rogers and Tanimoto, 1960) (τ) to guide the network search, making the use of traditional pathway search algorithms a more tractable undertaking. Similar successful approaches have been taken to enable rapid searching of extant metabolic networks such as KEGG for links to known pathway mapping (Hattori *et al.*, 2010) and validation of biochemical reactions in large databases (Félix and Valiente, 2007). Our approach applies the chemical similarity comparison on-the-fly to foster efficient convergence on a target molecule and subsequent annotation of the edges in the resultant network.

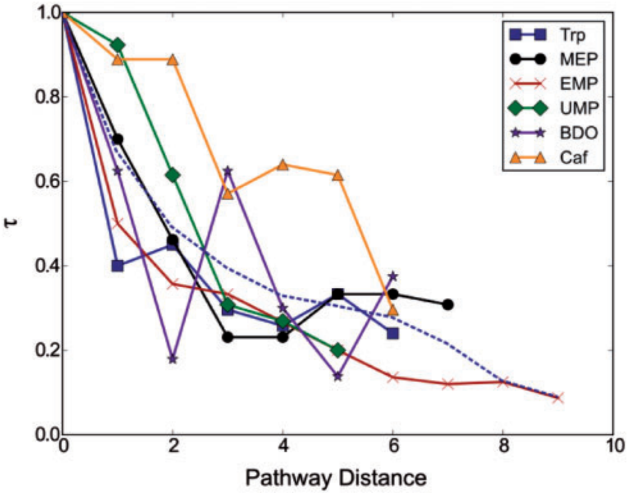

Fig. 1. The Tanimoto coefficient moves towards 1 as the pathway distance between a 'product' node and an 'intermediate' approaches 0, but can have deviations from this trend in cases where large cofactor modifications are carried out. The pathways shown are tryptophan synthesis (Trp), the MEP pathway, glycolysis (EMP), uridinyl synthesis (UMP), a synthetic 1,4-butanediol pathway (BDO) (Yim *et al.*, 2011), and caffeine synthesis (Caf). Pathway distance is defined as the number of transformations linking two metabolites. Tanimoto coefficient shows a general increasing trend on average as a function of pathway distance (dashed line)

In a metabolic network, nearby metabolites are more chemically similar than distant metabolites (Fig. 1). Therefore, proceeding towards a particular product via a metabolic pathway should see the intermediates generally becoming more similar to the final product. We have used this property of increasing chemical similarity in metabolic pathways to bias our search through metabolic networks. This algorithm of searching for increased chemical similarity cannot only be used for reaction/metabolite bipartite networks, but also can be utilized to search for promiscuous substrates an enzyme may act on within a metabolic network.

An additional improvement to pathway searching utilizes the searching algorithm developed by Byers and Waterman (1984) to expedite pathway searching by ensuring that only pathways that have the capability of meeting certain additional criteria are explored (Fig. 2). For the work in this article, we utilize the criteria of path length with the assumption that short paths are, in general, easier to engineer into a cell. However, this technique is capable of expediting pathway searching by utilizing additional criteria such as the number of reactions that are novel to an organism, the thermodynamic favorability of a pathway, or the number of energy carriers expended by the pathway. To our knowledge, this is the first use of this technique to search for pathways in putative metabolic networks.

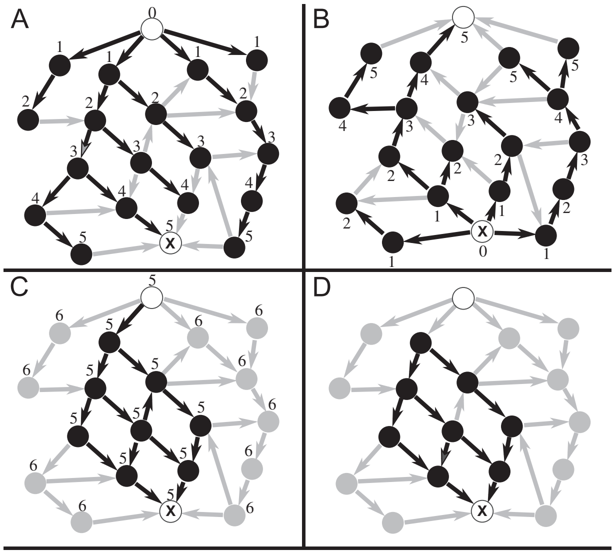

Fig. 2. The starting and target compound for the pathway search are shown by a white and a node marked with an "X", respectively. In this example, the maximum path length allowed is five. First, a breadth-first search is run to determine the distance of each node from the starting node (**A**). The black arrows indicate a shortest path to each node. Next, all of the edges are reversed and a breadth-first search is run again to determine the distance of each node from the target node (**B**). The sum of these two distances is calculated. If this sum is greater than five for any node, then that node and all its associated edges are removed (**C**). Finally, a depth-first search is run on the reduced network to find all pathways of length five or less (**D**). In this example, four such pathways exist and are illustrated by black arrows

Using these methods, we present a fast and efficient strategy for exploring metabolic networks generated by making use of known biological transformations to suggest alternative pathways based on the expansion of known enzymatic activities. As a test case, we apply the tools developed to a problem that was previously impractical using the Biochemical Network Integrated Computational Explorer (BNICE) (Hatzimanikatis *et al.*, 2005; Henry *et al.*, 2010a; Li *et al.*, 2004), a program that creates a metabolic network by simulating all combinations of possible biochemical reactions on a starting compound. Searching this large network is intractable with previous graph search approaches. We show that our new algorithms overcome these limitations by predicting a novel pathway linking methylerythritol phosphate (MEP) and isopentenyl pyrophosphate (IPP). The results should enable the analysis of metabolic networks that are orders of magnitude larger than current networks. Applying these tools will allow for the development of many novel synthetic routes to previously unattainable chemicals.

2 Materials and methods

2.1 Data sources and fingerprints

In order to analyze the thermodynamics and chemical similarity for native pathways, a metabolic model—iAF1260—was downloaded from the ModelSEED project (Henry *et al.*, 2010b), and processed with the NetworkX (Hagberg *et al.*, 2008) Python. Chemical similarity indexing was carried out with the 'FP4' fingerprint type available from the OpenBabel (O'Boyle *et al.*, 2008, 2011) cheminformatics package. The chemical fingerprints are 1D binary arrays where each position corresponds to a different substructure query. The value in each position is '1' if the substructure corresponding to that position is present in the molecule being fingerprinted and is '0' otherwise. Fingerprint pairs are scored according to Equation (1), where $\vec{A}$ and $\vec{B}$ represent two molecular fingerprints being compared.

$$\tau = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|^2 + \|\vec{B}\|^2 - \vec{A}\cdot\vec{B}} \tag{1}$$

Gibbs' free energy of formation (ΔG_f) values was predicted using group contribution theory (Jankowski et al., 2008; Mavrovouniotis, 1990, 1991) and was used to calculate Gibbs' free energy of reaction (ΔG_r) values. The thermodynamics calculations were made in BNICE and were set to correct for coupled reactions and pH (Henry et al., 2006, 2007) and oxygenase activity (VanBriesen, 2001). Searches for alternative pathways through metabolic networks after application of SimIndex (SI) were made using a depth first search.

2.2 SimIndex

Similarity indexing was used to reduce the number of nodes searched in a metabolic network by biasing the search toward nodes representing chemically similar molecules. SI takes a starting compound and target compound as inputs. During the network search, all nodes (molecules) that are the daughters of a particular node starting node X are compared with the target molecule via τ (τ^{dt}). The parent node X is also compared with the target compound (τ^{pt}). Daughter nodes that fail to meet the acceptance criterion $\tau^{dt} > \tau^{pt}$ are not searched because they represent graph steps that decrease in chemical similarity to the target molecule. In subsequent rounds of the search, where the daughter nodes are more than one step removed from X and hence may have disparate parents, the τ^{pt} used for the comparison is given by the average τ^{pt} as shown Equation (2), where i is the number of steps removed from X and n_i is the number of nodes i steps removed from X.

$$\tau^{pt} = \frac{1}{n_i}\sum_{j=1}^{n_i} \tau_j^{pt} \tag{2}$$

In benchmarking this method, it was often necessary to relax the rejection criterion by a tolerance k such that a node is not searched if $\tau^{dt} > k{*}\tau^{pt}$ in order that our method could successfully recapitulate native pathways. A k-value of 0 forces all nodes to be searched, making it equivalent to a breadth-first search. SI runs were performed on eight-processor high-memory nodes with 64 GB of RAM, and 2.4 GHz Intel Xeon E5620 processors.

2.3 Byers–Waterman pathway searching

Searches for alternative pathways through metabolic networks generated using SI were made using a modified version of the algorithm proposed by Byers and Waterman (1984) for finding all near-optimum pathways in a weighted network. Briefly, this algorithm works as follows. First, the shortest path is found between each node in the network and the target node. The length of this path *f(n)* is saved as a property of each node *n*. A standard depth-first search algorithm is then run with the following modification. Before the algorithm traverses any edge it first calculates the sum of the weights of the edges currently in the path, the weight of the edge being traversed, and *f(m)* where *m* is the destination node of the edge. If this value is greater than some user specified maximum cost, the edge is not traversed. This allows for the efficient discovery of all pathways *near* the optimum pathway where nearness is defined by the user.

For all the pathway searches performed in this work the criteria utilized to define optimality was the path length. Utilization of this criterion simplified the Byers–Waterman searching algorithm since each edge cost is identically one. This allowed for the implementation of the algorithm in the following manner. First, a breadth-first search was run from the starting compound and the distance of each node in the network from that node, d_s, was recorded. Next, a similar process was performed to determine the distance from each node in the network to the target compound for the pathway, d_t. Next, the following inequality was tested for each node i in the network.

$$d_s + d_t > N \tag{3}$$

Here, N is the maximum allowable length of the pathways. Any node for which this inequality is true is removed from the network because it is not on any path of length less than or equal to N. Finally, a depth-first search is performed to find all the pathways on this reduced network. This significantly increases the speed of pathway discovery by reducing the nodes which must be explored by the depth-first search.

2.4 SimZyme

SimZyme proposes putative edges between enzymes and metabolites based on the similarity of a given metabolite to other metabolites the enzymes is known to act upon. For SimZyme, we query the BRENDA enzyme database (Scheer *et al.*, 2011), which lists all documented substrate-product pairs for each fully elucidated enzyme as our training data. Once a pathway through a putative metabolic network is selected, all that is known about a given step in the pathway is a generalization of the chemistry performed by several different enzymes (Henry, 2007) and bears what is most often the first three digits of the Enzyme Commission (EC) designation of the enzymes used to create the chemical operator. Based on this information, it is necessary to select real enzymes that might be able to perform the appropriate chemistry on a given metabolite in the pathway. To accomplish this, a query metabolite is compared against all substrates listed in the BRENDA database for any EC numbers mapped to a particular biochemical transformation. For SimZyme, the mapping procedure compared reactions to be mapped to those produced by generalized transformations in the BNICE software (Henry, 2007). Known specific reactions reproduced by each generalized transformation are mapped to those generalized transformations. The output provided ranks enzymes for their potential to act upon the query metabolite by the similarity of the enzymes' known substrates to the query metabolite. We selected enzymes whose substrates most closely resembled our pathway intermediates to propose specific enzymes to use, relying on promiscuous enzyme activity to catalyze the reaction.

Benchmarking SimZyme was done using a leave-one-out cross validation. The sampling process proceeds by selecting a random compound associated with the queried biochemical transformation 100 times. The compounds selected were in SMILES form converted from the names in the database—compounds selected whose structures could not assigned canonical SMILES were omitted from the analysis. A naïve enzyme-substrate database is constructed by removing all instances of the chosen compounds one at a time from the full database. The naïve database is then queried as described above, with the deleted compound as the query in each iteration, the results were compiled and analyzed.

3 Results

3.1 Benchmarking SI

A Python-based script was developed to carry out the SI algorithm which biases a network search by prioritizing nodes that are more similar to the product as described in Section 2 (Fig. 3). We investigated two aspects of SI: its ability to find a target compound of interest, as well as its ability to complete this task while searching fewer nodes within a putative metabolic network than a breadth-first search. The NetGen module of the biochemical network exploration software BNICE (Hatzimanikatis *et al.*, 2005) was used to generate putative metabolic networks.

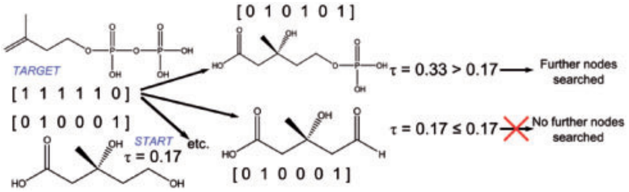

Fig. 3. SI modulates the number of compounds that are allowed to proceed to the next generation. In a given generation, the average Tanimoto coefficient of the parent molecules with the target is calculated. The Tanimoto coefficient of each daughter compound with the target is also calculated. The progeny of daughter compounds with a lower similarity with respect to the target compound than the parent compounds are not searched

We first tested the performance of SI in identifying native metabolic pathways in searching of putative metabolic networks, namely the mevalonate (MVA) pathway from R-MVA to IPP and the pentose phosphate pathway from β-D-glucose-6-phosphate (G6P) to D-arabino-3-hexulose-6-phosphate (A6P). Values of the chemical similarity bias parameter k were varied from 0.8 (weak bias for chemical similarity to 1.0 (strong bias, see 'Materials and methods') for runs on both networks to determine if the chemical similarity bias would find the native pathways to the target compound. Values of k below 0.8 yielded networks with a number of compounds nearly equal to that of a NetGen run without SI and thus were not used for benchmarking. Across all tested tolerances, the target compounds were found. Furthermore, pathways linking MVA and IPP (three steps) and G6P and A6P (four steps) were composed of compounds found in the native MVA pathway and pentose phosphate pathway, respectively.

The MVA pathway search with SI searched 1.5- to 5-fold fewer compounds between $k = 0.8$ and $k = 1.0$ compared with the breadth-first style NetGen search (Supplementary Fig. S2A). Average values of τ for the set of all compounds equidistant from start compound increased as the distance from the start compound increased (Supplementary Fig. S2B). The decrease in compounds searched was accompanied by a 3- to 27-fold decrease in computational time taken to generate the putative sub-network over the same range of chemical similarity biases (Supplementary Fig. S2C). Similarly, the number of compounds searched in the putative network surrounding G6P drastically decreased by 16- to 117-fold over the range of tolerances (Fig. 4A). As with the MVA to IPP search, the average value of τ for all compounds an equal distance from the start compound showed a general increasing trend as the distance from the start compound increased (Fig. 4B). The computational time for searching the network around G6P decreased 10- to 10 000-fold over the tolerance range (Fig. 4C).

3.2 Benchmarking Byers–Waterman pathway search

SI shows impressive efficiency at reducing the size of the generated-subnetwork. However, the use of additional requirements on the pathways can even further reduce the number of nodes that must be explored. To illustrate this, we found all pathways from G6P to A6P in the $k = 0.8$ network with lengths of three, four and five. Utilizing the Byers–Waterman search algorithm (BWPS), the number of nodes explored during the pathway search was reduced compared with the depth-first search by two to three orders of magnitude. The computational time gains begin to become evident in searches for pathways of length four, with a 5-fold time improvement. Searches for paths of length five are completed with a 10-fold reduction in computational time. This decrease in number of nodes visited did not, however, reduce the number of pathways found; in both the depth-first search and BWPS search, the same number of pathways was found for each specified pathway length. SI shows impressive efficiency at reducing the size of the generated-subnetwork. However, the use of additional requirements on the pathways can even further reduce the number of nodes that must be explored. To illustrate this, we found all pathways from G6P to A6P in the $k = 0.8$ network with lengths of three, four and five. Utilizing the BWPS, the number of nodes explored during the pathway search was reduced compared with the depth-first search by two to three orders of magnitude. The computational time gains begin to become evident in searches for pathways of length four, with a 5-fold time improvement. Searches for paths of length five are completed with a 10-fold reduction in computational time.

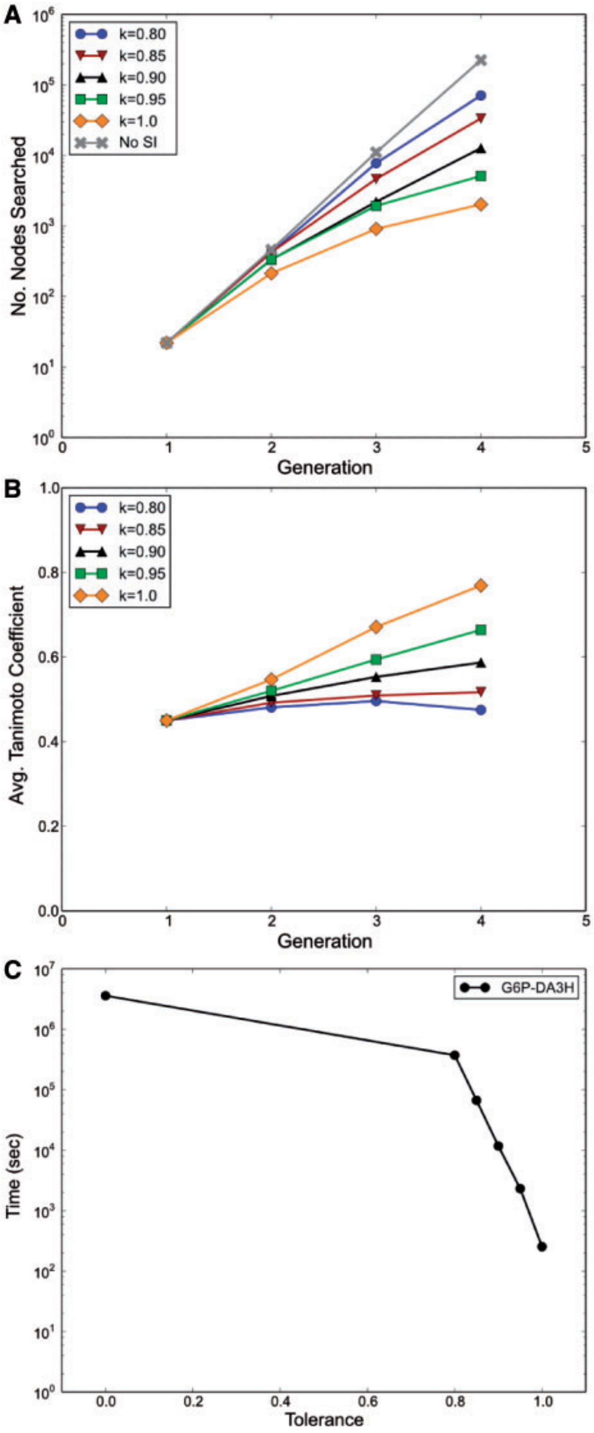

Fig. 4. In benchmarking a pathway search from G6P to A6P (four steps) with SI, the number of compounds searched at each level decreases over three orders of magnitude with varying tolerance (*k*) with SI (**A**). The average Tanimoto coefficient at each level approaches one with similar behavior to that observed in the three-generation run (**B**). Computational time decreases over five orders of magnitude, with the tolerance of zero representing a regular NetGen run that terminated upon reaching a time limit (**C**)

This decrease in number of nodes visited did not, however, reduce the number of pathways found; in both the depth-first search and BWPS search, the same number of pathways was found for each specified pathway length.

3.3 Benchmarking SimZyme

To validate SimZyme, which predicts promiscuous reactions, we examine the accuracy with which it could correctly assign enzymes that will act on a substrate given a training set of substrates of enzymes with a similar reaction type. The leave-one-out validation using the BRENDA enzyme promiscuity database (Scheer *et al.*, 2011) is based solely on the chemical similarity to other substrates on which the member enzymes show activity. The basis for our benchmarking is that the sampling of compounds that are substrates of enzymes in a given class is a Bernoulli process where success is defined as finding the first correct enzymes within the top n best-scoring enzymes that catalyze chemically similar substrates as scored by τ. The proportion of successful trials in a Bernoulli process follows a beta distribution, which is bounded on the interval (0, 1); 90 and 95% confidence intervals (CIs) were constructed around the mean proportion assuming this distribution as the probability density function (Fig. 5). We selected four types of biological transformations on which to test SimZyme. SimZyme succeeded in scoring the correct enzyme in the top three 90.1% of the time on average across the four classes tested; the correct enzyme received the highest score 74% of the time.

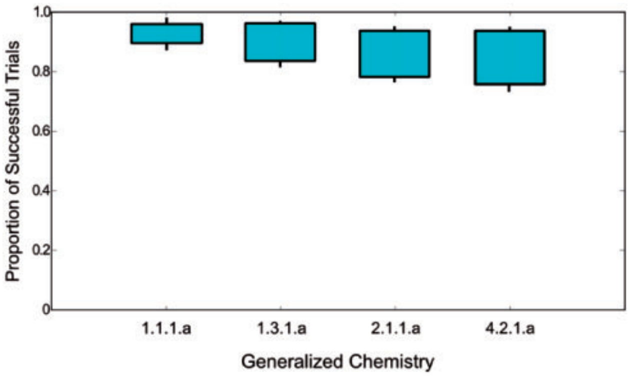

Fig. 5. CIs for the accuracy of SimZyme in matching substrates with enzymes performing a desired reaction that accept that substrate in the top three highest scores for four types of chemistry. The boxes represent a confidence level of 90%, while the whiskers show a confidence level of 95%

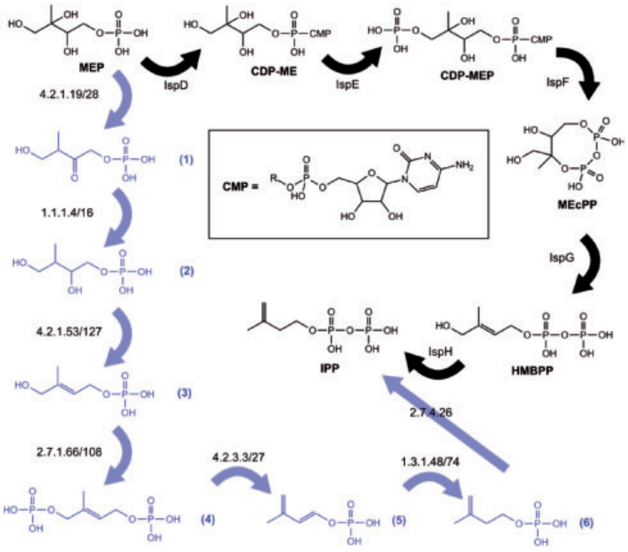

Fig. 6. An alternative MEP pathway (blue) proceeds to IPP in seven steps and was constructed using NetGen runs aided by SI. The intermediate step connecting (2) and (3) was found by manual search. The pathway to (2) was found with SI in 2 generations, while the pathway connecting (3) to IPP was found in a four-generation SI run

3.4 Case study: A *de novo* MEP pathway

In order to demonstrate the utility of both the improved pathway searching and SimZyme in proposing novel pathways extant in putative metabolic networks, we endeavored to elucidate a *de novo* pathway—which we define as proceeding through either partially or entirely through novel metabolites—linking MEP with IPP. This pathway is of particular interest for metabolic engineering as it has a better theoretical yield than the commonly used MVA pathway, but it has a number of problematic features (Ajikumar *et al.*, 2010; Carlsen *et al.*, 2013). This is chiefly due to the inefficiency of the iron-sulfur cluster (ISC) assembly machinery required to assemble the active forms of the proteins IspG (Lee *et al.*, 2010) and IspH (Gräwert *et al.*, 2004), two key enzymes in the native bacterial pathway (Fig. 6). Another feature of the pathway that proves problematic in bacteria is the limitation of intermediate efflux, primarily methylerythritol cyclodiphosphate, which is the product of the reaction catalyzed by the enzyme IspF (Zhou *et al.*, 2012), whose ΔG_r is an unfavorable +26 kcal/mol.

We first attempted to circumvent the terminal step in the pathway catalyzed by the ISC-containing enzyme IspH; we explored putative metabolic networks using both (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate and its singly phosphorylated analog, shown in Figure 6 as species **3**. We found a set of four transformations with suitable thermodynamics linking species **3** with IPP in the putative network searched with SI with $k = 0.85$. Having found a potential latter half of a *de novo* pathway for IPP production, we then searched the putative metabolic network surrounding the metabolite MEP for potential synthetic routes to species **3**. SI did not include species **3** in its truncated putative network of depth four around MEP, but did yield a species **2** that can potentially reach **3** by a reaction not included in the set of biochemical transformations used in generating the putative network. The failure to directly find a connection was because the appropriate biochemistry had not been encoded in the underlying software. Taken together, these putative metabolites form a seven-step pathway (Fig. 6) linking MEP and IPP via a vastly superior thermodynamic landscape (Fig. 7).

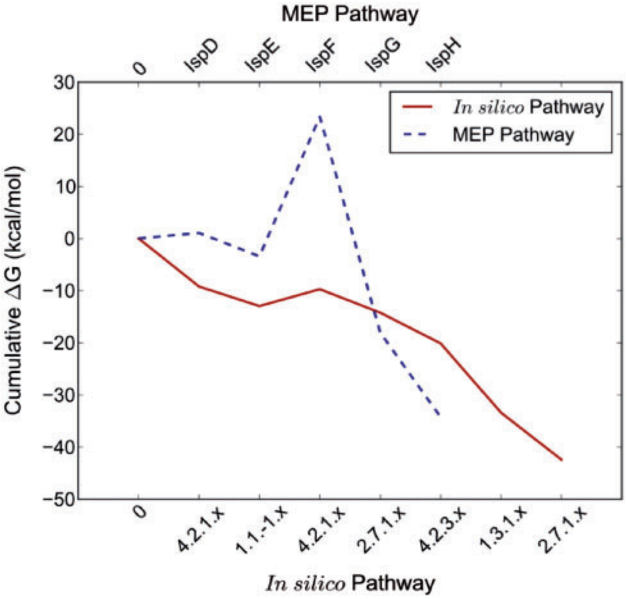

Fig. 7. The thermodynamic landscape of the original MEP pathway beginning from MEP has unfavorable thermodynamics at IspF (dashed line). The proposed alternative begins at MEP and proceeds with improved thermodynamics to IPP (solid line). The thermodynamics calculations take into account coupled reactions involving cofactors, such as ATP hydrolysis

We then used SimZyme to assist in assigning specific enzymes, making certain that the ones chosen were not dependent on ISCs for activity. We selected high-ranking results whose chemistry matched the transformations given in the putative networks (Table 1). The chemistry required to link intermediates **2** and **3** is of the type performed by the 4.2.1 class of enzymes, and so we determined that these two pathways could potentially be linked via promiscuous activity of enzymes that were determined by SimZyme. Additionally, the final transformation in the putative network that connects species **6** with IPP can be carried out with the native activity of isopentenyl phosphate kinase (EC 2.7.4.26). Because of SI, BWPS and SimZyme, we are now able to fully propose metabolic pathways for long and complicated chemistry.

4 Discussion

As the focus of metabolic engineering shifts to more exotic secondary metabolites, the exploration of metabolic networks is a vital first step in the design of novel biosynthetic pathways in common chassis organisms. It is highly useful to elucidate routes connecting two nodes in organism networks, metaorganism networks and putative networks of all possible biochemical transformations via *in silico* simulations (Ajikumar *et al.*, 2010; Choi and Lee, 2013; Henry *et al.*, 2010b; Li *et al.*, 2004; Yim *et al.*, 2011). As either type of network has the potential to be large and highly-connected—particularly if edges are added to account for the effects of enzyme promiscuity—efficient searching of the available pathways through the network is key for enumerating all possible routes linking two nodes in the network.

In linear native pathways, each successive compound in a given pathway has, in general, a greater similarity to the 'target' than the previous compound in line. We propose that this constraint is integral to reducing putative pathways existing in a given metabolic network, and may mirror the natural biological imperative to keep

Table 1. Enzyme predictions for novel MEP pathway

EC class	Reaction type	Best match(es)
4.2.1	Carbon-oxygen hydro-lyase	propanediol dehydratase
1.1.1	NAD(P)H-dependent CH-OH oxidoreductase	(R,R)-butanediol dehydrogenase galactitol 2-dehydrogenase
4.2.1	Carbon-oxygen hydro-lyase	oleate hydratase linalool dehydratase
2.7.1	Primary alcohol kinase/ phospotransferase	undecaprenol kinase dolichol kinase
4.2.3	Carbon-oxygen phospholyase	isoprene synthase
1.3.1	NAD(P)H-dependent CH-CH oxidoreductase	15-oxoprostaglandin 13-oxidase 2-alkenal reductase
2.7.4	Phosphate phosphotransferase	isopentenyl phosphate kinase[a]

[a]Used for its native reaction rather than a promiscuous reaction.

metabolic pathways to the fewest possible steps. Observed exceptions to this trend occur when relatively small molecules are activated with the attachment of a large cofactor—such as coenzyme A, acyl carrier protein, or cytidine—that can cause the similarity of an intermediate to sharply decrease. Representing these modifications as single atoms or single features in a fingerprinting routine has the potential to rectify this perceived downturn in chemical similarity along a metabolic pathway. In the event of pathways involving a condensation, it is also possible to observe exceptions to this trend.

The breadth-first style illumination of these large putative networks, achieved through iterative applications of a pre-defined set of reaction rules to a seed compound and its progeny, stands to benefit from similarity-based improvements to allow deeper exploration into chemical space. This could also be expanded to include non-biological chemical networks that are similarly large and would benefit from improved network navigation algorithms. This real-time generation of putative networks is necessary since, unlike metabolomics databases based on individual organisms, databases of putative networks are uncommon and existing ones tend to be focused on pharmaceutical targets in humans (Menikarachchi *et al.*, 2013; Peach *et al.*, 2012). With SI, the nodes eligible for addition to a pathway through a network at each step are reduced to those nodes which are more similar (within a tolerance) to the target compound than the previous cohort of nodes. The on-the-fly application of the similarity search as the nodes in the network are illuminated differentiates SI from existing approaches previously used to validate proposed reactions in biochemical databases (Félix and Valiente, 2007).

We attempt to select enzymes that will comprise the novel pathways discovered using similarity-based methods in the SimZyme utility. The problem of identifying enzymes with high promiscuity and determining if they will act on intermediates in novel pathways has been previously addressed using amino acid sequence data (Nath and Atkins, 2008) or 3D protein structure (Wu *et al.*, 2011), and others limit the scope of their promiscuity search to relatively small reaction libraries (Cho *et al.*, 2010). Similarity-based methods have the advantage of being relatively quick for comparing a proposed substrate against known substrates cataloged in a large database. The small amount of information required for SimZyme compared with other methods makes it an attractive choice for broad application. Compared with more intensive methods of determining enzyme promiscuity, a similarity-based approach coupled with the enzyme-substrate data from a large database such as BRENDA (Scheer *et al.*, 2011) is suitable for rapid *in silico* annotation of novel biosynthetic pathways to value-added compounds within biochemical networks as demonstrated by the *de novo* MEP pathway. BRENDA offers a larger amount of substrates for comparison to a proposed metabolite, as it enumerates observed experimental interactions in addition to consensus substrates as found in SIMCOMP/SUBCOMP searches (Hattori *et al.*, 2010) and KEGG-based approaches(Cho *et al.*, 2010). Although the BRENDA web service does support a substructure search, SimZyme performs a different, similarity-based substrate search that allows more results, which could reveal novel substrate-enzyme complex possibilities, expanding the scope of known enzyme promiscuity.

In order to demonstrate the utility of chemical similarity-based methods for navigating metabolic networks, we attempted a redesign of an existing metabolic pathway: the MEP pathway in *E.coli*. The MEP pathway consists of seven reactions, beginning with condensation of glyceraldehyde-3-phosphate with pyruvate, and culminating in the formation of the isoprenoid precursor IPP. Interest in introducing this pathway into yeast is due to its improved stoichiometry over the eukaryotic IPP-producing MVA pathway and the wealth of diverse product that can be produced from downstream reactions on IPP (Carlsen *et al.*, 2013). However, two of the pathway enzymes contain ISCs that cannot be efficiently loaded into the recombinant enzymes (Carlsen *et al.*, 2013; Gräwert *et al.*, 2004; Lee *et al.*, 2010) and it is possible that oxidative damage to ISCs in *E.coli* may also complicate production (Partow *et al.*, 2012). The redesigned pathway replaced the five steps linking the intermediate MEP with IPP with seven enzymatic steps that do not rely on ISC proteins. A putative network expanding radially from species **3** in the novel pathway contains on the order of 10^4 nodes; applying SI reduces that figure to on the order of 10^2 nodes, and application of BWPS reduces that figure even further by another order of magnitude, drastically increasing the speed of a depth-first search for a pathway linking **3** with IPP, while at the same time allowing for a diverse population of compounds to be considered.

SimZyme subsequently identifies candidate enzymes to carry out the reactions proposed in the novel metabolic pathway. As SimZyme ranks enzymes purely based on substrate similarity, the possibility exists that an enzyme performing a different type of reaction on the intermediate in the proposed pathway will be proposed. Algorithmically, this is difficult to account for because the ensemble of BRENDA entries searched is tied to the biotransformations used by the putative network-generating software, but errors are rare in the highest-scoring results and can easily be identified by inspection when they occur. In the case of the novel MEP pathway, SimZyme provided at least one candidate enzyme in the top three results with a reaction outcome congruent with the reaction predicted in the putative network.

Similarity based search tools as developed here are relatively fast, easily implementable methods to expedite the design of a novel metabolic pathways. In this article, we demonstrate the utility of a fingerprint similarity-based method for efficiently navigating a putative network from a starting compound to a product of interest; however, the concept of a similarity based, modified best-first search is easily extensible to native metabolic networks with orders of magnitude fewer nodes. Tools for creating putative networks can generate enormous amounts of data that must be sifted through in order to find pathways from a designated starting compound to a target of interest. Existing depth-first search algorithms that select branches at random are adequate for most current native metabolic networks and small meta-networks, but as the capability to generate larger putative networks grows, such algorithms will become inefficient. Increased efficiency over our method may be possible using a 3D fingerprint approach, since the 2D fingerprints used for this study encode the functional groups found via SMARTS search in a molecule and the implicit configuration within the groups, but does not capture the 3D geometry and connectivity of the molecule or the spatial relationships among the several functional groups. Future implementations of the methods described herein may show improved efficacy if based on a more complex 3D fingerprint similarity approach that can more precisely describe 3D characteristics, which are important in determining if a molecule is shaped similarly enough to known substrates of a given enzyme to be able to fit into the corresponding enzyme's active site.

A systematic workflow for efficient assembly of novel pathways *in silico* is an important approach for simplifying transition from computer to bench-top in modern metabolic engineering efforts. The increasingly large size of known metabolic networks has the benefit of increasing the palette of biochemical transformations that can be considered for inclusion in *de novo* biosynthetic pathways. On the other hand, this presents a challenge that merits a scalable and rapid solution. With the development of similarity-based methods to improve both the search for potentially useful metabolic pathways and the annotation of putative novel edges between nodes within large networks, the implementation of experiments to confirm the practicality of *in silico* predictions can be hastened.

Funding

NIH Biotechnology Training [T32 GM008449, DP]; DOE Computer Science Fellowship (AS); Bill and Melinda Gates Foundation [OPP1044008].

Conflict of interest: The authors have filed a provisional patent on the described MEP pathway jointly with Manus Biosynthesis, Inc. KT is a scientific adviser for Manus Biosynthesis, Inc. and holds stock in the company. The research described was not funded by Manus Biosynthesis, Inc.

References

Ajikumar,P.K. *et al*. (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science*, **330**, 70–74.

Altman,T. *et al*. (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

Burgard,A.P. *et al*. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng*., **84**, 647–657.

Byers,T.H. and Waterman,M.S. (1984) Determining all optimal and near-optimal solutions when solving shortest path problems by dynamic programming. *Oper. Res*., **32**, 1381–1384.

Carbonell,P. *et al*. (2012) Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst. Biol*., **6**, 10.

Carlsen,S. *et al*. (2013) Heterologous expression and characterization of bacterial 2-C-methyl-D-erythritol-4-phosphate pathway in Saccharomyces cerevisiae. *Appl. Microbiol. Biotechnol*., **97**, 5753–5769.

Cheng,F. *et al*. (2011) Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J. Chem. Inf. Model*., **51**, 2482–2495.

Cho,A. *et al*. (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol*., **4**, 35.

Choi,Y.J. and Lee,S.Y. (2013) Microbial production of short-chain alkanes. *Nature*, **502**, 571–574.

Daylight Theory Manual. (2011) Daylight Chemical Information Systems, Inc., Santa Fe, New Mexico, pp. 19–25.

Faulon,J.-L. *et al*. (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics*, **24**, 225–233.

Feist,A.M. *et al*. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol*., **3**, 121.

Félix,L. and Valiente,G. (2007) Validation of metabolic pathway databases based on chemical substructure search. *Biomol. Eng*., **24**, 327–335.

Gräwert,T. *et al*. (2004) IspH protein of *Escherichia coli*: studies on iron-sulfur cluster implementation and catalysis. *J. Am. Chem. Soc*., **126**, 12847–12855.

Guimerà,R. *et al*. (2007) A network-based method for target selection in metabolic networks. *Bioinformatics*, **23**, 1616–1622.

Hagberg,A.A. *et al*. (2008) Exploring network structure, dynamics, and function using network. In: *Proceedings of the 7th Python in science conference (SciPy 2008), Pasadena, CA, 19–24 August 2008*, pp. 11–15.

Hattori,M. *et al*. (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res*., **38**, 652–656.

Hatzimanikatis,V. *et al*. (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.

Henry,C.S. (2007) Computational thermodynamic and biosynthetic analysis of genome-scale metabolic models. (Order No. 3256045, Northwestern University). ProQuest Dissertations and Theses, 227–227 p. (304830966).

Henry,C.S. *et al*. (2007) Thermodynamics-based metabolic flux analysis. *Biophys. J*., **92**, 1792–1805.

Henry,C.S.*et al*. (2010a) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng*., **106**, 462–473.

Henry,C.S. *et al*. (2010b) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol*., **28**, 977–982.

Henry,C.S. *et al*. (2006) Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys. J*., **90**, 1453–1461.

Humble,M.S. and Berglund,P. (2011) Biocatalytic Promiscuity. *Eur. J. Org. Chem*., **2011**, 3391–3401.

Jankowski,M.D. *et al*. (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J*., **95**, 1487–1499.

Kotera,M. *et al*. (2014) Metabolome-scale prediction of intermediate compounds in multistep metabolic pathways with a recursive supervised approach. *Bioinformatics*, **30**, i165–i174.

Lee,M. *et al*. (2010) Biosynthesis of isoprenoids: crystal structure of the [4Fe-4S] cluster protein IspG. *J. Mol. Biol*., **404**, 600–610.

Li,C. *et al*. (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci*., **59**, 5051–5060.

Mavrovouniotis,M.L. (1990) Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng*., **36**, 1070–1082.

Mavrovouniotis,M.L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem*., **266**, 14440–14445.

Menikarachchi,L.C. *et al*. (2013) In silico enzymatic synthesis of a 400,000 compound biochemical database for nontargeted metabolomics. *J. Chem. Inf. Model*., **53**, 2483–2492.

Nam,H. *et al*. (2012) Network context and selection in the evolution to enzyme specificity. *Science*, **337**, 1101–1104.

Nath,A. and Atkins,W.M. (2008) A quantitative index of substrate promiscuity. *Biochemistry*, **47**, 157–166.

Nath,A. *et al*. (2010) Quantifying and predicting the promiscuity and isoform specificity of small-molecule cytochrome P450 inhibitors. *Drug Metab. Dispos*., **38**, 2195–2203.

Nobeli,I. *et al*. (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol*., **27**, 157–167.

O'Boyle,N.M. *et al*. (2011) Open Babel: An open chemical toolbox. *J. Cheminform*., **3**, 33.

O'Boyle,N.M. *et al*. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J*., **2**, 5.

Orth,J.D. *et al*. (2010) What is flux balance analysis? *Nat. Biotechnol*., **28**, 245–248.

Partow,S. *et al*. (2012) Reconstruction and evaluation of the synthetic bacterial MEP pathway in Saccharomyces cerevisiae. *PLoS One*, **7**, e52498.

Peach,M.L. *et al*. (2012) Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software. *Future Med. Chem.*, **4**, 1907–1932.

Perez-Gil,J. *et al*. (2012) Mutations in *Escherichia coli* aceE and ribB genes allow survival of strains defective in the first step of the isoprenoid biosynthesis pathway. *PLoS One*, **7**, e43775.

Phue,J.-N. *et al*. (2010) Acetate accumulation through alternative metabolic pathways in ackA (-) pta (-) poxB (-) triple mutant in *E. coli* B (BL21). *Biotechnol. Lett.*, **32**, 1897–1903.

Ranganathan,S. *et al*. (2010) OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.*, **6**, e1000744.

Rodrigo,G. *et al*. (2008) DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, **24**, 2554–2556.

Rogers,D.J. and Tanimoto,T.T. (1960) A computer program for classifying plants. *Science*, **132**, 1115–1118.

Scheer,M. *et al*. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, D670–D676.

Terfloth,L. *et al*. (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.*, **47**, 1688–1701.

VanBriesen,J.M. (2001) Thermodynamic yield predictions for biodegradation through oxygenase activation reactions. *Biodegradation*, **12**, 265–281.

Wu,D. *et al*. (2011) A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *J. Chem. Inf. Model.*, **51**, 1634–1647.

Yim,H. *et al*. (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol.*, **7**, 445–452.

Yousofshahi,M. *et al*. (2011) Probabilistic pathway construction. *Metab. Eng.*, **13**, 435–444.

Zhou,K. *et al*. (2012) Metabolite profiling identified methylerythritol cyclodiphosphate efflux as a limiting step in microbial isoprenoid production. *PLoS One*, **7**, e47513.