To appear in Journal of Cheminformatics, 2017

RESEARCH

A Review of Parameters and Heuristics for Guiding Metabolic Pathfinding

Sarah M Kim[1], Matthew I Peña[2], Mark Moll[1], George N Bennett[2] and Lydia E Kavraki[1*]

Emails: smk9@rice.edu, mip1@rice.edu, mmoll@rice.edu, gbennett@rice.edu, kavraki@rice.edu

Abstract

Recent developments in metabolic engineering have led to the successful biosynthesis of valuable products, such as the precursor of the antimalarial compound, artemisinin, and opioid precursor, thebaine. Synthesizing these traditionally plant-derived compounds in genetically modified yeast cells introduces the possibility of significantly reducing the total time and resources required for their production, and in turn, allows these valuable compounds to become cheaper and more readily available.

Most biosynthesis pathways used in metabolic engineering applications have been discovered manually, requiring a tedious search of existing literature and metabolic databases. However, the recent rapid development of available metabolic information has enabled the development of automated approaches for identifying novel pathways. Computer-assisted pathfinding has the potential to save biochemists time in the initial discovery steps of metabolic engineering.

In this paper, we review the parameters and heuristics used to guide the search in recent pathfinding algorithms. These parameters and heuristics capture information on the metabolic network structure, compound structures, reaction features, and organism-specificity of pathways. No one metabolic pathfinding algorithm or search parameter stands out as the best to use broadly for solving the pathfinding problem, as each method and parameter has its own strengths and shortcomings. As assisted pathfinding approaches continue to become more sophisticated, the development of better methods for visualizing pathway results and integrating these results into existing metabolic engineering practices is also important for encouraging wider use of these pathfinding methods.

Keywords: metabolic pathfinding; graph-based search; metabolic engineering

*Correspondence: kavraki@rice.edu

[1]Department of Computer Science, Rice University, 6100 Main St., Houston, TX, 77005 USA

Full list of author information is available at the end of the article

1 INTRODUCTION

Metabolic engineering is the scientific process of manipulating the metabolism of a microorganism to produce valuable compounds. Engineering microbial production involves the disruption of endogenous genes or adding genes from heterologous organisms to form pathways that tap into the natural metabolic network. There have been numerous successes of metabolic engineering, including the well publicized biosynthesis of artemisinic acid, a precursor to the antimalarial drug artemisinin [1], and thebaine, a precursor to hydrocodone and morphine [2]. In each of these cases, a pathway responsible for the production in plants was translated to a chassis microorganism, such as *E. coli* and *S. cerevisiae*, to separate the supply of these therapeutics from the plants they were sourced from. At the root of these successes is the identification of the requisite pathways and the systematic transfer of these pathways to a microbial host.

Metabolic pathfinding has clear applications to the first step in the design-build-test-learn cycle for developing biosynthetic pathways [3]. We define metabolic pathfinding as the process of identifying viable routes through a metabolic network from a starting compound to a desired target compound. Here, pathways are not limited to those that exist within a single organism, but can contain any enzymatic reactions from multiple organisms to complete a novel, heterologous pathway. To perform pathfinding we need a metabolic network that is constructed using information linking reactants to products through characterized enzymatic reactions. Several metabolic databases provide the requisite connectivity data used to construct a metabolic network structure. Of these, the Kyoto Encyclopedia of Genes and Genomes (KEGG) has been employed most frequently, likely due to being one of the first metabolic databases available with open access and a wide breadth of information. MetaCyc [4] also has descriptive entries for metabolic pathways that are attributed to many groups of organisms. Some databases, including BRENDA [5] and ExPASy [6], have more information about the enzymes including kinetics and protein structure, whereas others, such as ChEBI [7], specialize in descriptions of small molecules. New content is being continuously added to all these databases, many of which now source enzymatic reactions from thousands of organisms.

Traditionally, researchers have manually searched existing literature and databases to design pathways. However, the rapidly growing body of metabolic information makes it difficult to effectively survey and utilize all available resources. Computational approaches have been developed to enable researchers to take advantage of these growing resources. For example, pathways for production of 1,4-butanediol, a non-natural compound, were discovered with the assistance of a pathway-identification algorithm [8]. Thousands of pathways, four to six reactions long, were generated starting from common central metabolites. Solution prioritization was required to whittle the pathways down to a manageable number to be constructed and tested in the lab resulting in a demonstration of feasibility for a novel, biocatalytic route (Figure 1A). Assisted metabolic pathfinding may aid in the more rapid discovery of synthesis pathways for other valuable products.

Assisted metabolic pathfinding aims to solve two main challenges – the challenge of efficiently speeding up the pathway search process and the challenge of selectively finding biologically feasible, novel pathways. This paper focuses primarily on the approaches of pathfinding algorithms that address these two challenges. However, improvements in the search algorithms alone are not sufficient to solve these challenges, as the quality of the pathway results is also heavily dependent on the metabolic resources utilized by the search algorithm. Advancements in metabolic pathfinding rely on advancements in techniques for expanding the metabolic search space. For example, retrosynthesis-based approaches [9, 10] can be used to build search spaces that extend beyond the data stored in curated metabolic databases. Other databases like ATLAS [11] and XTMS [12] store information on extended search spaces and even apply existing pathfinding techniques (BNICE [13, 14] and RetroPath [15], respectively) to these spaces. Metabolic pathfinding may not be the main focus of retrosynthesis algorithms and expanded databases; however, these resources are nevertheless critical for finding novel metabolic pathways and will be included in this review.

The metabolic pathfinding problem itself can be further divided into two different approaches: graph-based pathfinding and constraint-based pathfinding. This review will focus on graph-based pathfinding, which highlights the connections between compounds and reactions in the metabolic network. Graph-based approaches represent a metabolic pathway as a path that consists of an ordered series of intermediate compounds and reactions that transform some defined starting compound(s) to some defined target compound(s). Graph-based pathfinding utilizes a very well-studied data structure to represent the metabolic network, abstracting away more complicated interactions between compounds and enzymes in the cell. This abstraction enables graph-based methods to readily scale with larger metabolic networks spanning multiple organisms. However, since much of the underlying metabolic network is abstracted by the graph representation, there is a greater chance for graph-based approaches to return pathways without biological significance unless relevant parameters and heuristics are introduced to guide the search. Constraint-based methods (e.g., [16]) highlight the stoichiometry and relative rates of reactions involved in the metabolic process being studied. In many constraint-based methods, a selected set of reactions is optimized to meet a specified objective (e.g., maximizing the yield of a valuable compound) under the

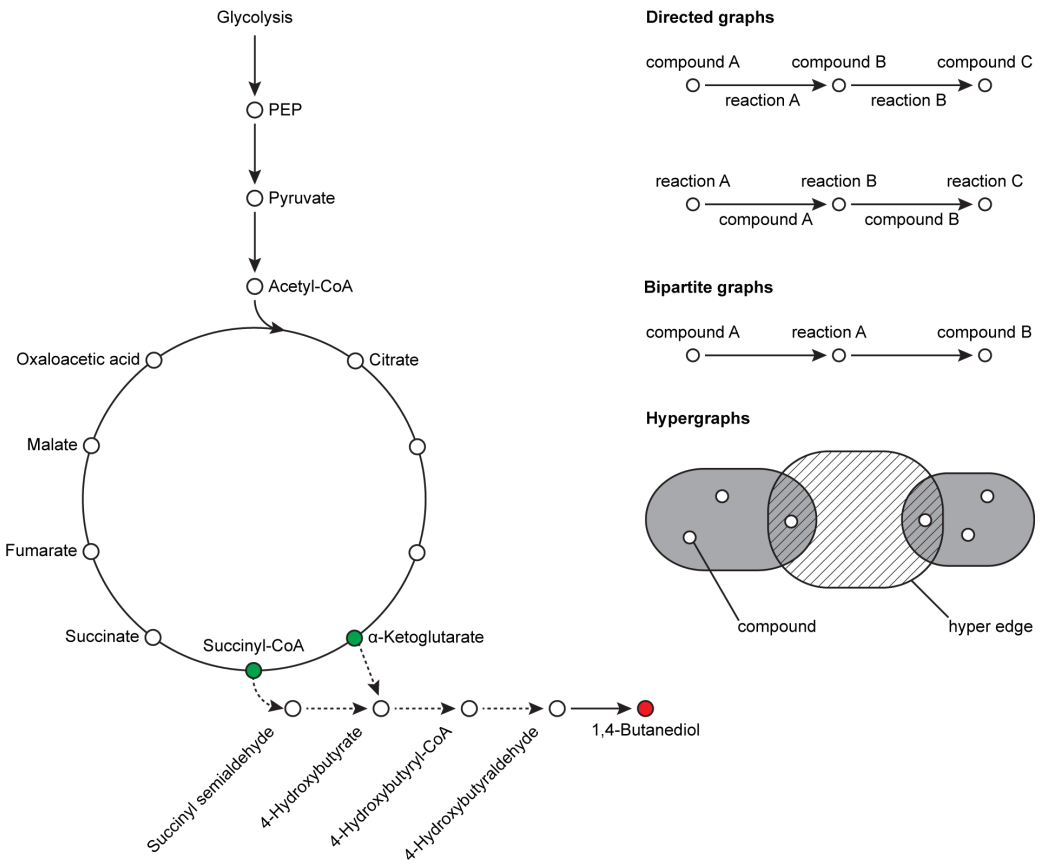


Figure 1 **A** 1,4-Butanediol is an example of a non-natural compound that is produced using a heterologous pathway generated by an pathway-identification algorithm. These heterologous genes from organisms differing from the host are dashed. Intermediates of the citric acid cycle were used as starting compounds in a pathway search restricted by path length toward the desired target, 1,4-butanediol. **B** Graph-based networks can be described as either directed, bipartite, or hypergraphs. In directed graphs, nodes (open circles) and edges (arrows) can represent either compounds or reactions. In the case of bipartite graphs, the pathway will be composed of nodes that alternate between compounds and reactions. For hypergraphs, groups of multiple compounds can be linked through a single hyperedge that may represent a reaction rule that links compounds that are associated with multiple additional reaction rules.

steady state assumption, meaning that there is no net increase or decrease of metabolites within the studied system. For constraint-based methods, elementary flux modes or extreme pathways can serve as the representation of a metabolic pathway [17, 18, 19]. Unlike graph-based paths which may only include the main compounds and reactions in a pathway, elementary flux modes and extreme pathways provides a more complete summary of the requisite intermediate compounds and enzymes while conforming to steady-state constraints. Overall, constraint-based methods tend to offer a more accurate model of a known metabolic network, such as one from a well-studied organism like E. coli. However, this approach is not yet able to computationally scale to very large metabolic networks [20]. Though algorithms have been developed to identify viable pathways using elementary mode analysis [20, 21], we choose to focus specifically on graph-based pathfinding to examine how parameters and heuristics can be used to efficiently guide the search in large-scale metabolic networks.

A metabolic network can be described as connections between compounds and the enzymes catalyzing reactions between compounds, which lends itself well to graph representation. There are many different ways a metabolic network can be represented as a graph (Figure 1B). One of the simplest ways is for the nodes in a graph to represent the compounds in the metabolic network, and the edges to represent the reactions or enzymes that connect one compound to another. This representation is used in several earlier pathfinding algorithms [22, 23, 24]. It is also possible for the nodes in a metabolic graph to represent the enzymatic reactions and the edges to represent the intermediate compounds, as done in MetaRoute [25]. Another possible graph

representation of the metabolic network is for both compounds and reactions to be represented as nodes in a bipartite graph, where edges represent the connections between compounds and reactions. This representation is used in a few algorithms [26, 27]. A third possible graph representation is the hypergraph, where multiple compounds (i.e., the reactants) can be connected to multiple target compounds (i.e., the products) with a single hyperedge (the reaction). Unlike other graph representations, the hypergraph representation can connect two different groups of compounds with a single reaction hyperedge, which allows more details about each reaction (i.e., all intermediate compounds involved) to be shown explicitly in the representation [20]. The hypergraph representation is used in several pathfinding and retrosynthesis algorithms [20, 28, 29, 12, 30]. Node and edge weights based on relevant parameters (e.g., atom mappings, compound similarity, reaction thermodynamics, and organism-specific information) can be introduced to any of the above graph representations to guide the pathfinding search towards more biologically relevant results.

This review covers the techniques supporting graph-based metabolic pathfinding algorithms and the heuristics that guide pathway discovery from networks, enzymatic reactions, and chemical structures to a specific host organism context (Figure 1B). We will begin with a description of the structure of the metabolic network in terms of (1) graph connectivity, which refers to the number of connections each node has across the network, and (2) path length, or the number of transformative steps that separate any two compounds in the network (Section 2). Then, the role of compound structure (Section 3) and reaction specific information (Section 4) in identifying feasible, novel pathways will be discussed. Next, we briefly describe the role of organism-related information (Section 5). We conclude the paper with a discussion of the limitations and implications for future directions for metabolic pathfinding (Section 6). By describing the advantages and disadvantages of features used in current pathfinding approaches, we hope to guide interested users to the algorithms that suit their needs while summarizing the latest research for developers.

2 METABOLIC NETWORK STRUCTURE

Properties of the metabolic network representation can be used to guide and constrain the search problem and rank the resulting pathways. The properties that have been used in the literature are the connectivity of the network and the length of pathways found. The individual compounds and reactions of a pathway can also be assigned weights based on biochemical and network-based properties.

2.1 Graph connectivity

The graph-based representation of the network makes it intuitive to gravitate towards graph-based features and constraints, particularly graph connectivity (Figure 2A). Many approaches identify highly connected compound nodes in the graph, or hub compounds, which appear in many different reactions. Identifying hub compounds can suggest potential currency metabolites, or side compounds that are used as energy or electron providers but are not incorporated into the final product compounds (e.g., NADH, ATP, etc.). As such, pathways routing through currency metabolites tend to not be biologically meaningful, and for many algorithms, these currency compounds are manually removed [13, 14, 31, 24, 32]. In Croes et al. [31], the weight of compound vertices is set equal to the degree of the compound in the network, biasing the search against going through highly connected compounds. Croes et al. compared this weighted graph search with an unweighted graph and a filtered graph (where 36 highly connected pool metabolites were removed), and found that weighted graph search performed better (85% correspondence with annotated pathways) than the unweighted graph search (30%) and filtered graph search (65%). Croes et al. also suggested that the small world property of metabolic networks described by Wagner and Fell in 2001 [33] is an artifact of having currency metabolites in unweighted metabolic graphs, which make compounds in the metabolic network seem more tightly connected. This is also suggested by several other papers [34, 35, 36].

In Faust et al. [26], different weighting schemes for compounds and reactant pairs (RPAIRs) were compared amongst each other. The weighting schemes included weighting compounds by degree, as described by Croes et al. in 2006, and weighting RPAIRs by their classification type. The RPAIR classification can be treated as a ranking for how relevant the pair of compounds are in the reaction. For example, if an RPAIR is classified as "main," the compounds involved in the RPAIR are considered the main chemical transformation that occurs in the reaction, whereas a RPAIR classified as "cofac" or "ligase" may describe compounds that serve as metabolite compounds or facilitators of the reaction. Faust et al. introduces higher weights for RPAIR classifications that are considered less relevant to the reaction, favoring pathways that include more RPAIRs classified as "main." According to this study, searches using the Croes et al. weighting for compounds found better results than searches without compound weighting, while using RPAIR classification weighting showed no significant improvement in search results.

In MetaRoute [25], the weight of the compound vertices is set to the sum of the out-degree of the compound and the context weight of the in-going reaction nodes. The context weight is based on the degree of the side compounds involved in the reaction. The context weighting gives rare compounds a high weight and common compounds a low weight, encouraging paths to go through reactions that use common compounds as side compounds.

The connectivity of a graph is very simple to compute, and it is no surprise that it has been used by several metabolic

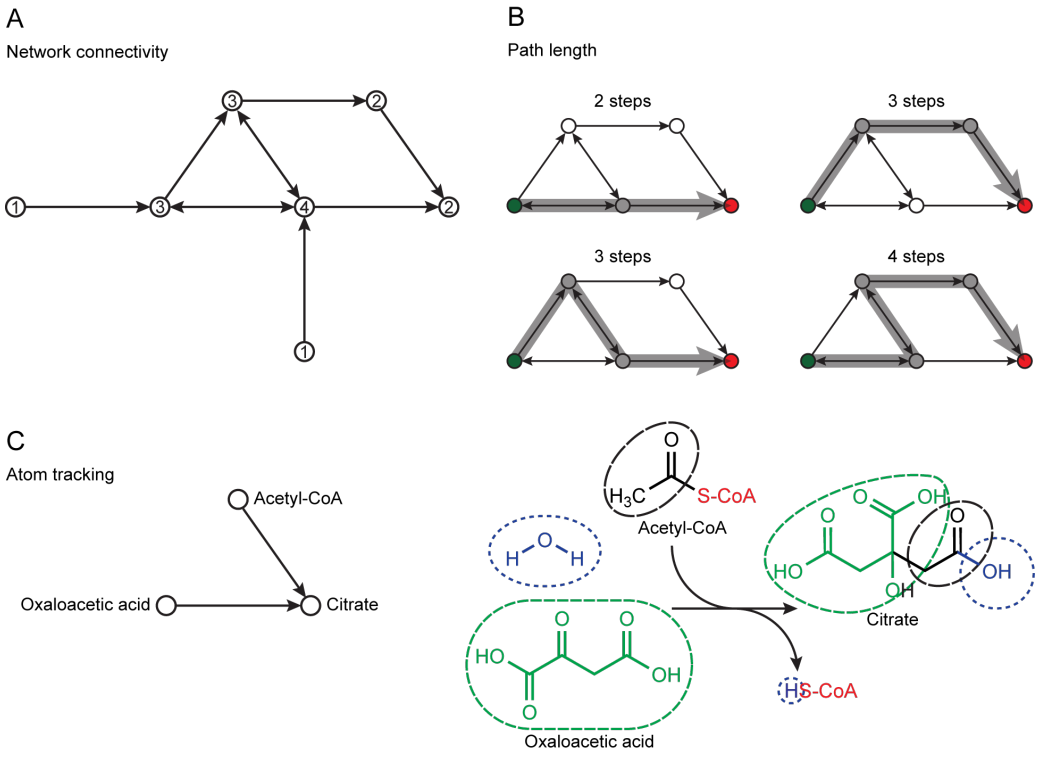


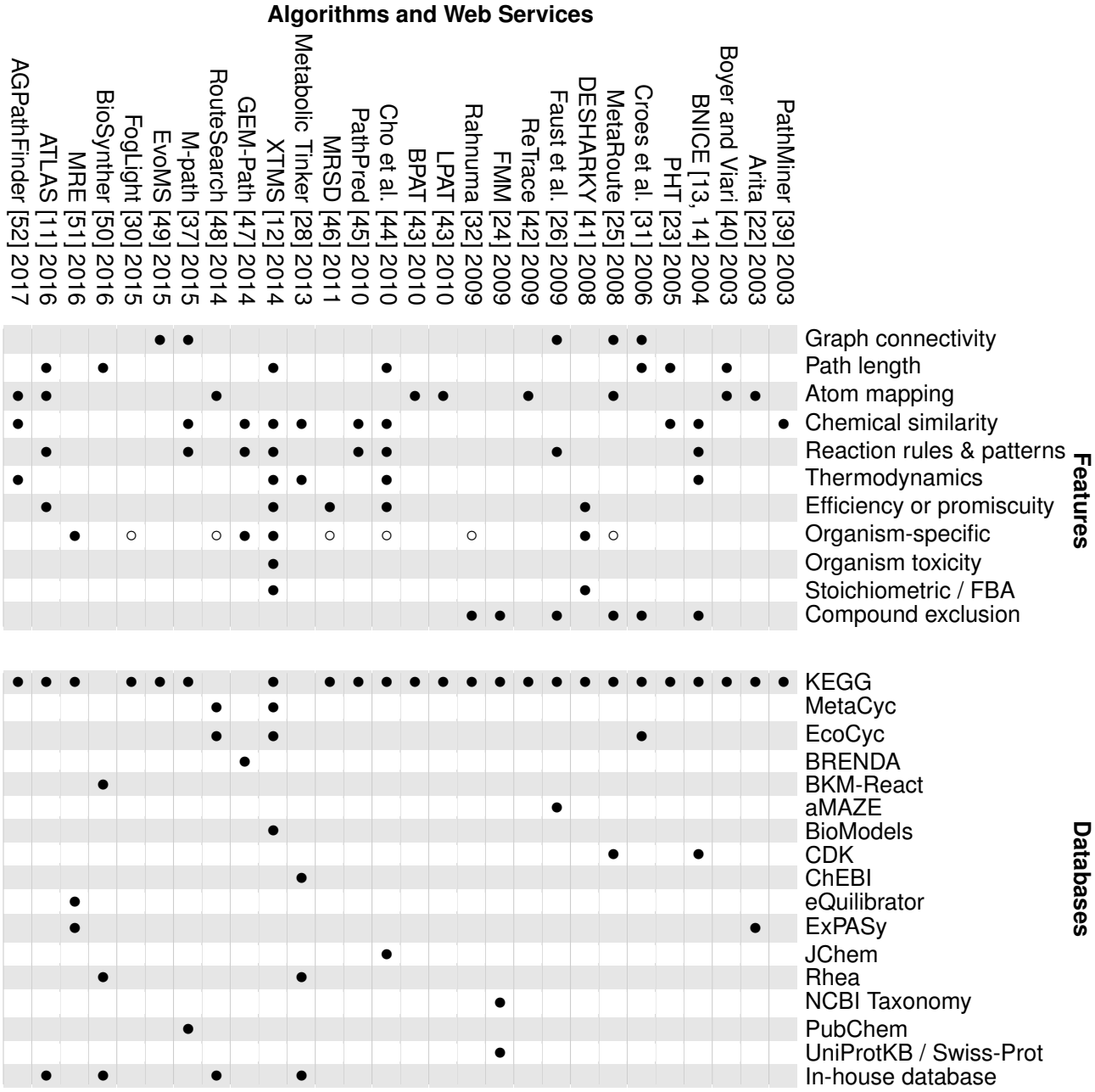
Figure 2 **A** The degree of connectivity for a node (vertex) within a metabolic network can be described by the number edges that are incident to it. In this case, the degree for each node, or the total number of incoming and outgoing edges in the directed graph, is provided in each node. **B** The path length can be easily determined by counting the number of edges that must be traversed to complete a pathway from the starting (green) node to the target (red) node. This is directly analogous to the number of enzymatic steps that must be engineered to complete a biosynthetic pathway. **C** Atom mapping is a highly detailed interpretation of chemical transformation that occurs along each edge of the metabolic network. Here, the first step of the citric acid cycle can be both described as a simple network with two edges and three nodes, or as a balanced chemical reaction where atoms can be tracked between reactants and their corresponding products. Notice that the network ignores water and coenzyme A (CoA), whereas the atom mapped reaction can be used to identify what atoms from the reactants acetyl-CoA and oxaloacetic acid contributed to the product, citrate.

pathfinding approaches. Despite its simplicity, connectivity can be used to effectively infer some biochemical information about the metabolic network. However, excluding features of the metabolic network based on connectivity alone may not reflect known biochemical properties. For example, excluding highly connected compounds to avoid currency compounds may also exclude compounds that play a significant role in pathways (e.g., pyruvate). Unlike other algorithms, M-path by Araki et al. [37] uses hub compounds as a launch point to speed up the search. The approach identifies 139 compounds involved in eight or more reactions as hub compounds and introduces the reactions between the start compound and the hub compounds as the first steps in the search. Araki et al. refers to a paper by Barabasi and Oltvai [38], which suggests that highly connected compounds that are not currency metabolites are critical in linking together many compounds in the metabolic network. By including these highly connected compounds as first intermediates, the M-path algorithm can shorten the number of reaction steps needed to reach the target compound and improve the performance of the search.

2.2 Path length

Pathfinding algorithms often optimize for pathways with the smallest number of enzymatic steps, as these pathways tend to require less manipulation in a metabolic engineering context (Figure 2B). Many pathfinding algorithms set a maximum path length [31] or give the user an option to specify a maximum path length [40, 23, 48]. Pitkanen et al. [42] uses path length as part of the pathfinding heuristic to limit the search in the underlying networks. Pathways can also be ranked based on path length (e.g., algorithms finding *k*-shortest paths [27, 53]). Ranking by path length is often a byproduct of the applied graph search algorithm (i.e., *k*-shortest paths) and used to organize pathway results. In order to distinguish pathfinding methods that actively include path length as a constraint or heuristic from methods that only use path length to rank pathway results, the latter cases were not marked as using path length in Table 1. Since

Table 1 Algorithms are arranged by publication date, and closed circles denote the features used in the respective pathfinding algorithms along with the databases they draw information from. Open circles used in the organism-specific column indicate that users may input weights or parameters to make the algorithm organism-specific, but the algorithm itself does not provide options for the user to select for specific organisms.



path length is an inherent property of the solution pathway, no additional computation is required for obtaining this information. However, strongly biasing the search towards shorter pathway results ignores longer pathway results that may be equally valid for lab testing, such as the pathway for synthesizing a precursor of opioids that took 23 enzymatic steps in lab [2].

3 STRUCTURE OF COMPOUNDS

Most network representations include both structures of compounds and reactions, along with parameters that give additional information on both these parts. The chemical structure of compounds in the metabolic network can be useful in inferring the existence of a biochemical reaction between compounds, as biochemical reactions tend to have products that structurally resemble one or more reactants. Structural information can be represented at different levels of detail, which introduces a trade-off between the accuracy

of the similarity measure and the computational complexity of the overall metabolic pathfinding problem.

3.1 Atom tracking

At the finest level of detail, algorithms can track changes on atomic level (Figure 2C). Retaining as many of the atoms from the start compound in the target compound automatically excludes currency metabolites that contribute no atoms to the final product, which helps exclude pathway results that are biochemically infeasible. Also, conserving as much of the atomic structure of compounds in each reaction step can help to select pathways that are more biologically feasible. This method was first introduced by Arita in 2003 [22], which aims to conserve at least one atom from start compound to target using *k*-shortest paths. The MetaRoute algorithm [25] also uses this approach. Building on this approach, new algorithms aimed to conserve multiple atoms. Pitkanen et al. [42] uses a heuristic to maximize the number of carbons transferred during a reaction, while also minimizing the path length. This encourages the inclusion of reactions that transfer more carbon atoms in the final branched pathway results. In Heath et al. [27], the pathway must conserve a minimum number of carbon atoms from start to target compound. A search to find the maximum number of conserved carbon atoms will start with the total number of carbon atoms in either the start or target compound and then decrement this number by one if no pathways are found that conserve that number of atoms. In Boyer and Viari [40], pathways must conserve a minimum number of atoms which do not necessarily need to be carbons. In the initial carbon flux path algorithm proposed by Pey et al. [53], any reactions not involving a carbon exchange between its main reactant and product were removed from the search space. Pey et al. later updated their carbon flux paths algorithm to include atom tracking [54] to insure carbons from the start compound were eventually incorporated into the target compound. RouteSearch [48] maximizes atoms conserved throughout the pathway using a heuristic scoring function. This score accounts for five different atom types (carbon, oxygen, nitrogen, phosphorus, and sulfur), and each type of atom can be assigned a different weight. More recently, atom group tracking has been introduced by AGPathFinder [52]. Instead of tracking single atoms, this algorithm tracks groups of adjacent atoms connected by bonds. This avoids the computational cost of tracking individual atoms, but still captures much of the information gained by atom tracking. Incorporating atomic level information into the search ensures that at least a portion of the starting compound is used to produce the target compound, which may filter out many biologically infeasible pathways. In previous years, atom mapping information was not as readily available; however, as new methods have been developed to computationally predict atom mapping, more and more pathfinding algorithms have included atom tracking in the search. Tracking individual atoms can be computationally expensive, especially if every possible combination of atoms conserved from compound to compound is considered [55]. Even so, the fact that many recent pathfinding approaches incorporate atom tracking suggests it is an important parameter for the pathfinding problem.

3.2 Chemical Similarity

If two compounds have similar chemical structures, there is a decent chance that these compounds can be connected by a common reaction. Several approaches have used different representations of chemical structure as a way of guiding, constraining, and ranking the search.

3.2.1 Chemical Fingerprint

Several approaches use chemical fingerprints and Tanimoto coefficients [56] to measure compound similarity. A chemical fingerprint is a binary vector consisting of a string of ones and zeros. Each bit represents whether the compound contains a certain structural feature, such as the number of single carbon, carbon bonds present in the compound and the presence of chemical functional groups or ring structures. There are many available compound fingerprints that include different numbers and types of structural features. The Tanimoto coefficient is used to measure the similarity between two different compounds and is calculated by dividing the total number of structural features shared between the two compounds by the total number of structural features contained in both compounds. In Pathway Hunter Tool (PHT) [23], chemical fingerprints are included in the metabolite mapping scoring function, which is calculated by summing the calculated chemical similarity score and percentage atomic mass contribution. The algorithm uses this score to determine which reactants and products will be connected by edges in their search graph.

3.2.2 Graph-Based Comparison

Other approaches rely on the graph representation of chemical compounds. Metabolic Tinker [28] uses a heuristic based on similarity of functional groups of atoms and bonds between the current compound and the target compound identified using a graph comparison technique similar to the one described in [58]. In this technique, each compound is represented as a graph, where atoms are the nodes and bonds are the edges. Common structural features between compounds are then identified by finding the maximal common subgraph(s). SIMCOMP [59], an algorithm that identifies the maximum common substructure between the graph representations of two compounds, was used for building the KEGG RPAIR database utilized by many pathfinding algorithms (Figure 3A). SIMCOMP uses a variant of the Bron-Kerbosch maximum clique algorithm [60] to identify the maximum common substructure of two compounds.

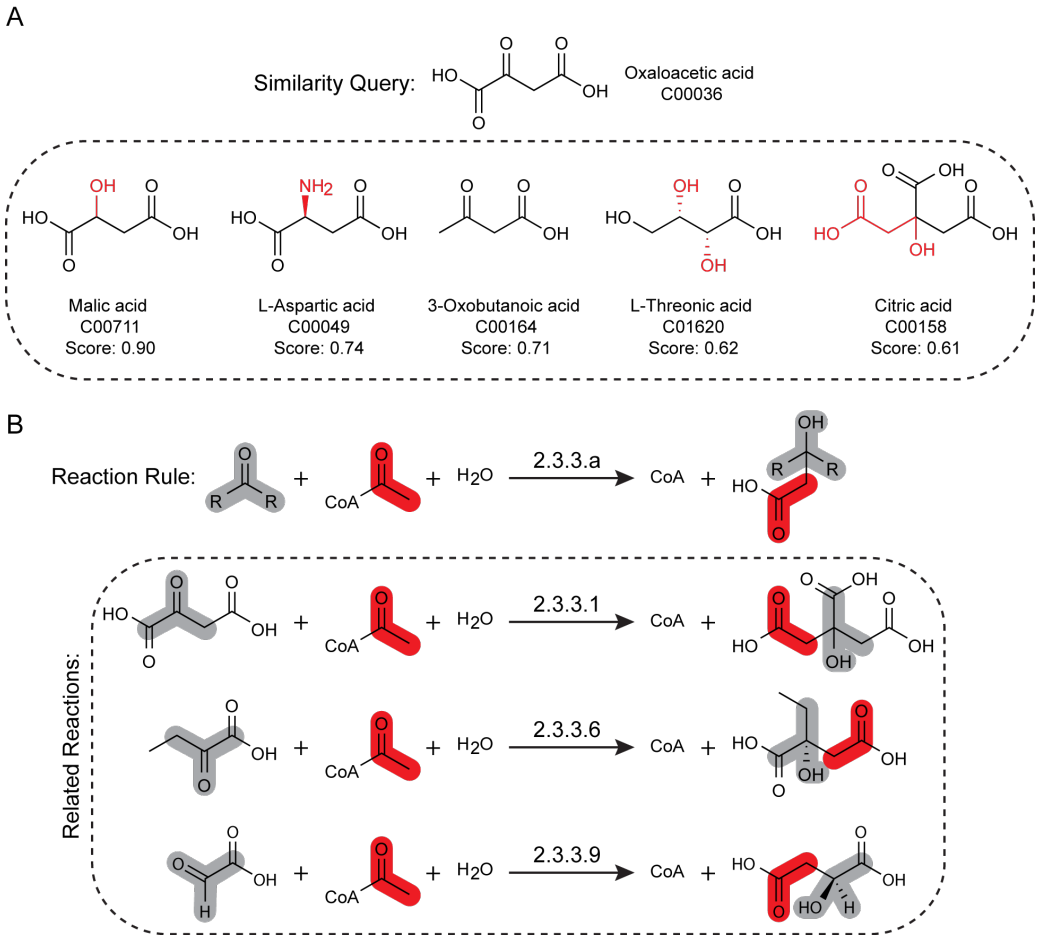


Figure 3 **A** Chemical similarity can be quantified by comparing common structural features. For example, SIMCOMP is an algorithm that calculates a similarity score by determining the maximum common substructure between two compounds. SIMCOMP can be used to rank all compounds in the KEGG database against a single queried compound, in this case oxaloacetic acid, by similarity score. The structural dissimilarities (substitutions and additions marked in red) of five familiar compounds relative to oxaloacetic acid are highlighted to provide context to the calculated similarity score. **B** Generalized reaction rules describe enzymatic reactions where the reactants share structural motifs and undergo related transformations within an EC class. Reaction rule 2.3.3.a is one of 86 BNICE generalized reaction rules described by Henry et al. [57]. Examples of enzymatic reactions described by this reaction rule include citrate (Si)-synthase (2.3.3.1) that forms citrate from oxaloacetate, 2-ethylmalate synthase (2.3.3.6) that forms 2-ethylmalate from 2-oxobutyrate, and malate synthase (2.3.3.9) that forms malate from glyoxylate. Motifs of the constituent reactants are highlighted in grey (aldehyde or ketone group) and red (acetyl group), and the corresponding atoms are identified in the products. Databases can be expanded, as is the case with MINEs and ATLAS, by applying reaction rules to metabolites that share a common motif.

Unlike chemical fingerprints, where a pre-determined set of chemical characteristics are used to compare two compounds, the graph comparison approach directly compares the chemical structure of two compounds against each other. The graph comparison approach tends to be more accurate in calculating structural similarity but is more computationally expensive [61]. In GEM-Path [47], both chemical fingerprints with Tanimoto coefficients and the subgraph matching of chemical structure are used to measure chemical similarity.

Calculating compound similarity is not as computationally expensive as atom mapping and serves as a check that the reactions included in pathways are biochemically feasible. However, compound similarity falls short in the cases where two compounds share many common structural components but are not biochemically related.

4 REACTIONS

In addition to information about the compounds involved in the network, graph-based searches also include information on reactions. This information can be used to both constrain and expand the search to find novel pathways.

4.1 Reaction rules

Building off the idea of structural similarity, some algorithms introduced reaction rules, or more general transitions

between compounds based on changes in chemical structure. Two enzymatic reactions may involve different reactants and products; however, if the same structural change occurs between reactants and products in these reactions (i.e., functional group A is replaced by functional group B), these reactions may both fall under the same reaction rule. Reaction rules allow new, potentially feasible pathways to be found by introducing reactions that may not yet have been added to metabolic databases. These rules can both be used to (1) create a metabolic network without directly requiring information on enzymatic reactions from metabolic databases and (2) help expand an existing metabolic network created based on a metabolic database.

Reaction rules are based heavily on structural representations of compounds. In BNICE [13, 14], compounds are represented as an atom-bond matrix, and the reactions are represented as the difference between the matrices of the substrate and product compounds (Figure 3B). With this more generalized representation of reactions, BNICE reduces the existing database of 43,000 enzymes to 250 generalized enzymatic reactions by grouping together enzymes that catalyze reactions which follow the same reaction rules. In PathMiner [39], each compound is similarly described as a set of 145 chemical descriptors (based on atoms/bond information), and reactions are represented as vector differences. The reactions are used as a heuristic to guide an A* search [62]. In M-path [37], compounds are represented by chemical feature vectors that account for 318 atom and bond feature types. Atom types include primary, secondary, and tertiary carbons, and each covalent bond in a compound is counted as a pair of atom types. Reactions are again represented as reaction feature vectors that describe difference in number of atom/bond feature types between substrates and products. In Cho et al. [44], there is a reaction rules database containing constructed reaction rules. PathPred [45] uses so-called RDM patterns from RPAIRs, which take into account the reaction center, the difference regions, and the matched regions between the reactants and products. PathPred also uses Jaccard coefficient [63] to compare compounds, and it weights the atoms closer to the reaction center more greatly compared to more distant atoms. A reaction score is calculated based on the Jaccard coefficient for each reaction, and the overall pathway score is the average of the reaction scores of all its reactions. In Faust et al. [26], RPAIR mappings are used without atom tracking to show the connectivity of compounds without annotations of atoms. In FMM [24], reactions are represented as a $16{,}884 \times 16{,}884$ matrix, where each row and column represents a compound and having a '1' represents that there exists a forward reaction between the compounds. In RetroPath by Carbonell et al. [15], the molecular signature of any given compound is defined by a subset of neighboring atoms and chemical bonds surrounding each individual atom in the compound. The reaction rules are defined as the differences in molecular signatures between the reactant compounds and product compounds in a reaction. Only the atoms and bonds within a given number of bonds away from each atom are considered as part of the molecular signature. This distance, referred to as the diameter by Carbonell et al., could be increased to include more surrounding atoms and bonds in the molecular signature and in turn, make each reaction rule include more detailed differences in molecular structure between reactants and products. Or, the diameter could be decreased to include less of the surrounding atoms and bonds in the molecular signature, causing each reaction rule to be more general and applicable to more groups of compounds. Thus, by changing the diameter, the strictness of reaction rules can be adjusted to prevent an exponential explosion of potential reactions. Reaction rules allow the search to find novel pathways not present in existing metabolic databases. However, the issue with using reaction rules to find new paths is that there is a potential for an exponential explosion of results.

4.2 Thermodynamics

Another common feature taken into account by pathfinding algorithms is thermodynamic feasibility of the reactions in pathways. Almost all algorithms that include thermodynamics use the component contribution method [64] for calculating ΔG. In MetabolicTinker [28], missing directional information is inferred from ΔG. If it is not possible to calculate the ΔG, the edge is treated as a bidirectional edge. The search heuristic is based partially on thermodynamics, and paths are ranked based on thermodynamic feasibility. In BNICE [14], the ΔG value is used to analyze enzymatic reactions in different groups (profiling) and suggest feasibility of reactions. In Cho et al. [44], enzymes are ranked based on thermodynamic favorability, among other factors (such as binding site covalence and chemical similarity). The XTMS webserver [12] uses a scoring function to rank pathway results found by the RetroPath search algorithm. The XTMS scoring function incorporates the thermodynamic favorability of a pathway by both including the sum of all the ΔG values (taken from MetaCyc) of each reaction in a pathway and including the number of unfavorable reactions (any reactions with a ΔG value greater than zero) for each pathway. AGPathFinder [52] uses ΔGs (in addition to compound similarity) to guide the search as weights.

4.3 Stoichiometry

Graph-based pathfinding methods can incorporate reaction stoichiometry to limit the number of biologically irrelevant pathway results. The carbon flux paths algorithm proposed by Pey et al. [53, 54] introduces steady-state constraints. Pey et al. demonstrate that using carbon flux paths significantly reduces the connectivity of certain compounds, such as oxaloacetate in *E. coli*, compared to a graph-based search without stoichiometric constraints. Introducing stoichiometric constraints allows carbon flux paths to distinguish between oxic and anoxic conditions in E. coli, which was not

possible in previous graph-based algorithms. However, this pathfinding method was only tested within the metabolic network of a single well-studied organism (E. coli) and, like constraint-based methods, is not easily scalable to large multi-organism networks.

4.4 Enzyme efficiency and promiscuity

Enzymes can have different reaction rates, depending on how efficient an enzyme is in converting the substrate to product. On the other hand, promiscuous enzymes can catalyze reactions which may not be found in existing databases and may be used to expand the metabolic pathfinding search. In Cho et al. [44], binding site covalence was factored into ranking enzymes, where the highest ranked enzyme candidates were included in the final pathway solutions. In MRSD [46], edges between compounds are weighted based on the frequency of reactions that use the specified substrate to produce the specified product. This approach does not filter out species duplicates. The XTMS webserver scoring function [12] takes into account a gene score in ranking pathway results found by the RetroPath algorithm. The gene score is calculated for each pathway based on the average of the pathway's individual reaction scores, which is determined by the estimated promiscuity of the putative enzyme assigned to the given reaction based on the tensor product technique.

5 ORGANISM

Many algorithms give the user the ability to select an organism of interest. Arita et al. [22] mention that their search algorithm can find pathways specific to one organism if the user specifies a weighting scheme that heavily penalizes reactions taken from all other organisms. In RouteSearch [48], the user can specify weights for reactions taken from organism vs. reactions taken from a larger library including all organisms. Many others require the user to select which organism or group of organisms to look at [32, 46]. Other methods do not require user input. In Cho et al. [44], enzymes are ranked based on organism specificity. DESHARKY [41] limits the number of compounds that are not organism-specific to only one non-specific reactant and one non-specific product. In GEM-Path [47], there is an association between reactions and organisms. One of the more interesting of these algorithms is MRE [51], where the search takes into account endogenous competition of reactions. By considering which reactions happen more frequently in an organism, pathways can be optimized to include the most common reactions to maximize the production of the target compound and exclude reactions that may only occur at very low rates in the organism.

6 DISCUSSION

Pathfinding is a critical and preliminary step in the development of novel biosynthetic pathways. Pathfinding is often done manually, though there are many existing tools that can enumerate putative pathways with minimal input from the user. After a pathway has been identified, much time and effort goes into building, testing, troubleshooting, and optimizing the biological system, and not the initial pathway discovery [65]. This is acknowledged by metabolic engineers and synthetic biologists alike. Assisted pathfinding, for now, is typically restricted to providing and suggesting a series of enzymatic conversions through the aforementioned algorithms and ranking heuristics. It is up to the user to determine what organisms the genes should be sourced from based on limited enzyme kinetic data, which genetic system to use to regulate expression, and which organism to use as an appropriate host. Each step of this process is a challenge, and widespread adoption of assisted pathway discovery algorithms will depend on improved integration with the pathway engineering workflow. For this reason, future directions of assisted pathfinding must include the following: 1) maximizing the utility of existing but limited databases to find paths to non-native or other diverse commodity compounds, 2) facilitating the interpretation of the generated pathway solutions through visualizations and other methods, 3) assisting in gene selection based on known enzyme kinetics and other parameters of enzyme activity, and 4) identifying solutions with specific network topologies such as branched pathways and or cycles.

6.1 Non-native compounds

There has been a recent push to expand searches to non-native compounds using reaction rules, building on BNICE [13, 14], because it is appreciated that the single greatest limiting factor to pathfinding is the completeness of the referenced databases. The ability to find paths to a non-native compound is severely limited when restricted to metabolic databases consisting of almost entirely of native compounds. General reaction rules can substitute for predicted enzyme promiscuity where specific enzyme reactions for a structurally similar but a non-native substrate are needed as either the target or an intermediate in a pathway. Reaction rules can serve as an acceptable best guess or a lead when a pathway cannot be found in its absence. This need has recently lead to the generation of expanded databases (e.g., MINEs [66] and ATLAS [11]) that apply reaction rules to existing databases (e.g., KEGG [67]) to augment them and expand their reach. More work is needed in this area, as our research has identified a number of compounds of interest that still remain outside the reach of these expanded databases.

6.2 Databases

Although the cumulative information that is available across all metabolic databases is extensive, manually searching, gathering, and compiling information from different databases is a challenging task. Each database often has its

own representation and set of ID numbers for identifying components like compounds and reactions, in addition to its own organization schema, suited specifically for the intended purposes of the database. These differences make it challenging to determine the exact links and relationships between information in different databases. There have been a few recent efforts to integrate different metabolic databases and create a less redundant, more comprehensive, and more accessible resource for metabolic information (e.g., BKMreact [68], MetRXN [69], and MNXref [70]). The effort to make a more comprehensive, unified metabolic resource could be a great asset to developing new metabolic pathfinding algorithms, as the metabolic representations, heuristics, and constraints used in these algorithms rely heavily on the breadth and completeness of the used metabolic database(s). In addition to this, it would be very helpful for databases to adopt an open distribution model when fiscally reasonable. Restrictions on data distribution hinder further development of pathfinding tools, and licensing barriers make it harder to adopt a single framework.

6.3 Interface and Visualization

As the pathfinding capabilities improve, so do the number of solutions that can potentially be generated, and with it the challenge of providing the user with tools to explore the solutions that can number in the thousands and identify pathways of interest. Because of this, there is an increasing amount of user interaction built into pathfinding webservers (see MRSD [46], BioSynther [50], ATLAS [11], and XTMS [12]). By having a more interactive webserver interface, users can quickly modify their queries or filter the results to find the solutions they want. This filtering may be achieved either by ranking as has been previously discussed, clustering of results based on pathway similarity or overlap [71], allowing the user to exclude pathways based on the presence or absence of specific intermediates that the user chooses to avoid, or some mixture of all of these. Improved visualization solutions will provide users with a balance between an abundance of options and ease of identifying promising pathways.

6.4 Gene selection

In addition to visualizations, a well-developed interface could integrate suggestions for genes based on enzyme activity and evidence of heterologous gene expression so that the user can seamlessly transition from pathway discovery to the initial build phase. Databases, such as BRENDA [5], have experimentally determined values for many enzymatic characteristics that could be used in determining the gene of choice for each reaction step. However, this information has yet to be implemented in a pathway discovery and selection webserver.

6.5 Topology

Almost all pathfinding algorithms are limited to producing linear pathways with a few exceptions [43, 42]. Branched pathways and cycles represent different topologies of metabolic networks that are of interest to metabolic engineering because the resulting condensation or recycling of constituent material can potentially improve the theoretical yield for a pathway. Though linear pathways are sufficient in most cases, the capability of identifying more complex and efficient pathways would be desirable.

6.6 Conclusion

Ultimately, the best pathfinding algorithm is the one that suits the user's needs and is paired with an interface that facilitates pathway discovery. Pathfinding webservers can assist with the design of novel, feasible, and hopefully improved pathways, but as discussed, pathfinding needs to become more highly integrated with the entire process of metabolic engineering. This survey of the available features and future directions aims to increase adoption of existing pathfinding tools while advocating for advancements that will increase their utility.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Sarah Kim and Matthew Peña co-wrote section 1. Sarah Kim wrote sections 2–5, while Matthew Peña wrote section 6 and created all the figures. Table 1 was jointly created by Sarah Kim, Matthew Peña, and Mark Moll. Mark Moll, George N. Bennett, and Lydia E. Kavraki advised and provided expert feedback on the writing of the paper.

Funding

This work was supported in part by NSF DBI-1262491 and the NSF Graduate Fellowship awarded to Sarah Kim.

Acknowledgements

We would like to acknowledge Allison Heath for introducing Sarah Kim to the topic of metabolic pathfinding.

Author details

[1]Department of Computer Science, Rice University, 6100 Main St., Houston, TX, 77005 USA. [2]Department of BioSciences, Rice University, 6100 Main St., Houston, TX, 77005 USA.

References

1. Paddon CJ, Westfall P, Pitera DJ, Benjamin K, Fisher K, McPhee D, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature. 2013;496:528–532.
2. Galanie S, Thodey K, Trenchard IJ, Interrante MF, Smolke CD. Complete biosynthesis of opioids in yeast. Science. 2015;349:1095–1100.
3. Petzold CJ, Chan LJG, Nhan M, Adams PD. Analytics for metabolic engineering. Frontiers in Bioengineering and Biotechnology. 2015;3:135.
4. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research. 2016;44:D471–D480.
5. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Research. 2017;45:D380–D388.
6. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. 2003;31:3784–3788.

7. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research. 2008;36:D344–D350.
8. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, Boldt J, et al. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nature Chemical Biology. 2011;7:445–452.
9. Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Current opinion in chemical biology. 2015;28:99–104.
10. Carbonell P, Planson AG, Faulon JL. Retrosynthetic design of heterologous pathways. Systems Metabolic Engineering: Methods and Protocols. 2013;p. 149–173.
11. Hadadi N, Hafner J, Shajkofci A, Zisaki A, Hatzimanikatis V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. ACS Synthetic Biology. 2016;5:1155–1166.
12. Carbonell P, Parutto P, Herisson J, Pandit SB, Faulon JL. XTMS: Pathway design in an eXTended metabolic space. Nucleic Acids Research. 2014;42:W389–W394.
13. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. Chemical engineering science. 2004;59:5051–5060.
14. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. Bioinformatics. 2005;21:1603–1609.
15. Carbonell P, Parutto P, Baudier C, Junot C, Faulon JL. Retropath: Automated pipeline for embedded metabolic circuits. ACS Synthetic Biology. 2013;3(8):565–577.
16. Chowdhury A, Maranas CD. Designing overall stoichiometric conversions and intervening metabolic reactions. Scientific Reports. 2015;5.
17. Klamt S, Stelling J. Two approaches for metabolic pathway analysis? Trends in biotechnology. 2003;21(2):64–69.
18. Trinh CT, Wlaschin A, Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Applied microbiology and biotechnology. 2009;81(5):813.
19. Acuna V, Chierichetti F, Lacroix V, Marchetti-Spaccamela A, Sagot MF, Stougie L. Modes and cuts in metabolic networks: Complexity and algorithms. Biosystems. 2009;95(1):51–60.
20. Carbonell P, Fichera D, Pandit SB, Faulon JL. Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC systems biology. 2012;6(1):10.
21. Vieira G, Carnicer M, Portais JC, Heux S. FindPath: a Matlab solution for in silico design of synthetic metabolic pathways. Bioinformatics. 2014;30(20):2986–2988.
22. Arita M. In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism. Genome Research. 2003;13:2455–2466.
23. Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics. 2005;21:1189–1193.
24. Chou CH, Chang WC, Chiu CM, Huang CC, Huang HD. FMM: A web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Research. 2009;37:W129–W134.
25. Blum T, Kohlbacher O. MetaRoute: Fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics. 2008;24:2108–2109.
26. Faust K, Croes D, van Helden J. Metabolic Pathfinding Using RPAIR Annotation. Journal of Molecular Biology. 2009;388:390–414.
27. Heath AP, Bennett GN, Kavraki LE. Finding metabolic pathways using atom tracking. Bioinformatics. 2010;26:1548–1555.
28. McClymont K, Soyer OS. Metabolic tinker: An online tool for guiding the design of synthetic metabolic pathways. Nucleic Acids Research. 2013;41:e113–e113.
29. Fehér T, Planson AG, Carbonell P, Fernández-Castané A, Grigoras I, Dariy E, et al. Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering. Biotechnology Journal. 2014;9:1446–1457.
30. Khosraviani M, Zamani MS, Bidkhori G. FogLight: An efficient matrix-based approach to construct metabolic pathways by search space reduction. Bioinformatics. 2015;32:398–408.
31. Croes D, Couche F, Wodak SJ, Van Helden J. Inferring meaningful pathways in weighted metabolic networks. Journal of Molecular Biology. 2006;356:222–236.
32. Mithani A, Preston GM, Hein J. Rahnuma: Hypergraph-based tool for metabolic pathway prediction and network comparison. Bioinformatics. 2009;25:1831–1832.
33. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings of the Royal Society of London B: Biological Sciences. 2001;268:1803–1810.
34. Arita M. The metabolic world of Escherichia coli is not small. Proceedings of the National Academy of Sciences of the United States of America. 2004;101:1543–7.
35. Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics. 2003;19:270–277.
36. van Helden J, Wernisch L, Gilbert D, Wodak SJ. Graph-based analysis of metabolic networks. Ernst Schering Research Foundation workshop. 2002;38:245–74.
37. Araki M, Cox RS, Makiguchi H, Ogawa T, Taniguchi T, Miyaoku K, et al. M-path: A compass for navigating potential metabolic pathways. Bioinformatics. 2015;31:905–911.
38. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature reviews genetics. 2004;5:101–113.
39. McShan DC, Rao S, Shah I. PathMiner: Predicting metabolic pathways by heuristic search. Bioinformatics. 2003;19:1692–1698.
40. Boyer F, Viari A. Ab initio reconstruction of metabolic pathways. In: Bioinformatics. vol. 19; 2003. p. ii26–ii34.
41. Rodrigo G, Carrera J, Prather KJ, Jaramillo A. DESHARKY: Automatic design of metabolic pathways for optimal cell growth. Bioinformatics. 2008;24:2554–2556.
42. Pitkänen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Systems Biology. 2009;3:103.
43. Heath AP, Bennett GN, Kavraki LE. An Algorithm for Efficient Identification of Branched Metabolic Pathways. Journal of Computational Biology. 2011;18:1575–1597.
44. Cho A, Yun H, Park JH, Lee SY, Park S. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Systems Biology. 2010;4:35.
45. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, et al. PathPred: An enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Research. 2010;38:W138–W143.
46. Xia D, Zheng H, Liu Z, Li G, Li J, Hong J, et al. MRSD: A web server for Metabolic Route Search and Design. Bioinformatics. 2011;27:1581–1582.
47. Campodonico MA, Andrews BA, Asenjo JA, Palsson BO, Feist AM. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. Metabolic Engineering. 2014;25:140–158.
48. Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. Bioinformatics. 2014;30:2043–2050.
49. Gerard MF, Stegmayer G, Milone DH. EvoMS: An evolutionary tool to find de novo metabolic pathways. BioSystems. 2015;134:43–47.
50. Tu W, Zhang H, Liu J, Hu QN. BioSynther: A customized biosynthetic potential explorer. Bioinformatics. 2015;32:472–473.
51. Kuwahara H, Alazmi M, Cui X, Gao X. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. Nucleic Acids Research. 2016;.
52. Huang Y, Zhong C, Lin HX, Wang J. A Method for Finding Metabolic Pathways Using Atomic Group Tracking. PLoS ONE. 2017;12:e0168725.
53. Pey J, Prada J, Beasley JE, Planes FJ. Path finding methods accounting for stoichiometry in metabolic networks. Genome biology. 2011;12(5):R49.
54. Pey J, Planes FJ, Beasley JE. Refining carbon flux paths using atomic trace data. Bioinformatics. 2013;p. btt653.
55. Heath AP. Computational discovery and analysis of metabolic pathways [dissertation]. Rice University. 6100 Main St., Houston, TX 77005, USA; 2010.
56. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics. 2015;7:20.
57. Henry CS, Broadbelt LJ, Hatzimanikatis V. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. Biotechnology and bioengineering. 2010;106(3):462–473.
58. Hattori M, Okuno Y, Goto S, Kanehisa M. Development of a chemical

structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. Journal of the American Chemical Society. 2003;125:11853–11865.
59. Hattori M, Okuno Y, Goto S, Kanehisa M. Heuristics for chemical compound matching. Genome Informatics. 2003;14:144–153.
60. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM. 1973;16(9):575–577.
61. Öztürk H, Ozkirimli E, Özgür A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. BMC bioinformatics. 2016;17(1):128.
62. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall; 2009.
63. Jaccard P. The distribution of the flora in the alpine zone. New Phytologist. 1912;11:37–50.
64. Mavrovouniotis ML. Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. Biotechnology and Bioengineering. 1990;36:1070–1082.
65. Keasling JD. Synthetic biology and the development of tools for metabolic engineering. Metabolic Engineering. 2012;14:189–195.
66. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of Cheminformatics. 2015;7:44.
67. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45:D353–D361.
68. Lang M, Stelzer M, Schomburg D. BKM-react, an integrated biochemical reaction database. BMC Biochem. 2011;12:42.
69. Kumar A, Suthers PF, Maranas CD. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. BMC Bioinformatics. 2012;13:6.
70. Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M. Reconciliation of metabolites and biochemical reactions for metabolic networks. Brief Bioinform. 2014;15:123–35.
71. Kim SM, Peña MI, Moll M, Giannakopoulos G, Bennett GN, Kavraki LE. An Evaluation of Different Clustering Methods and Distance Measures Used for Grouping Metabolic Pathways. In: 2016 International Conference on Bioinformatics and Computational Biology. ISCA; 2016. p. 115–122.