Synthetic and Systems Biotechnology 2 (2017) 243–252

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: http://www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology/

Review Article

A review of computational tools for design and reconstruction of metabolic pathways

Lin Wang, Satyakam Dash, Chiam Yu Ng, Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA

ARTICLE INFO

Article history:
Received 21 August 2017
Received in revised form
6 November 2017
Accepted 6 November 2017

ABSTRACT

Metabolic pathways reflect an organism's chemical repertoire and hence their elucidation and design have been a primary goal in metabolic engineering. Various computational methods have been developed to design novel metabolic pathways while taking into account several prerequisites such as pathway stoichiometry, thermodynamics, host compatibility, and enzyme availability. The choice of the method is often determined by the nature of the metabolites of interest and preferred host organism, along with computational complexity and availability of software tools. In this paper, we review different computational approaches used to design metabolic pathways based on the reaction network representation of the database (i.e., graph or stoichiometric matrix) and the search algorithm (i.e., graph search, flux balance analysis, or retrosynthetic search). We also put forth a systematic workflow that can be implemented in projects requiring pathway design and highlight current limitations and obstacles in computational pathway design.

Contents

1. Introduction

Nature has endowed specific biochemical capabilities to many organisms spanning diverse metabolic pathways ranging from carbon dioxide fixation by *Clostridium ljungdahlii* using Wood-Ljungdahl pathway [1] to ammonia assimilation by cyanobacteria using the glutamine synthase cycle (GS-GOGAT) [2]. Advancements in metabolic engineering have enabled us to engineer and express enzymes and construct novel pathways for various applications including drug discovery [3,4] and value-added biochemical production [5]. Notably, Galanie et al. recently engineered the complete opioids biosynthesis pathways constituting of 21 and 23

* Corresponding author. Department of Chemical Engineering, The Pennsylvania State University, 126 Land and Water Research Building, University Park, PA 16802, USA.
E-mail address: costas@psu.edu (C.D. Maranas).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.synbio.2017.11.002

native and heterologous enzymes to produce thebaine and hydrocodone, respectively in yeast [4]. In addition, multi-enzymatic steps can nowadays be engineered in a cell-free system for *in vitro* synthesis [6,7]. The pathway search involves finding the right combination of enzymes to form the pathway connecting a given source molecule (e.g., carbon substrate or any native metabolites in a cell) to a target molecule. Computational pathway design algorithms enumerate potential routes linking the two molecules, while often taking into consideration a multitude of criteria such as shortest route, minimal number of heterologous reactions, thermodynamic feasibility, and enzyme availability. While most methods capitalize on the large number of enzymatic reactions available in nature, there is also an increasing number of tools that employ biotransformation rules derived from the existing reactions to design *de novo* pathways [8,9]. The latter relies on the remarkable malleability of enzymes [10–12] to accept a broad range of substrates as well as the potential of protein engineering [13,14] and *de novo* enzyme design [15]. As an example, Savile et al. carried out *in vitro* synthesis of enantiopure anti-diabetic sitagliptin by combining computational protein engineering and directed evolution to broaden the substrate range of transaminase enzyme [7].

Pathway discovery tools have successfully guided several metabolic engineering efforts. In particular, Yim et al. [5] demonstrated the production of up to 18 g/L of 1,4-butanediol (BDO) in *E. coli* by engineering the best pathways after surveying over 10,000 computationally designed pathways. The BDO titer was increased to 110 g/L with improved downstream enzymes [16]. Their success highlights the potential application of pathway design algorithms to a variety of projects [6]. Pathway design tools are not only applicable to pathway prospecting for biosynthesis of commodity chemicals, biofuels, or pharmaceuticals, but have also been applied to develop biosensing pathways for target molecules. For example, Libis et al. used the retrosynthetic approach (XTMS) to identify pathways from undetectable target molecules such as drugs, pollutants, and biomarkers to known inducer molecules, which could then activate transcription factors [17]. The activated transcription factors can be used to regulate an easily detectable metabolite, antibiotic marker or fluorescence protein, which can be subsequently used to screen for strains producing target molecules [17].

As a large number of pathway design tools have been published, identifying the best method depending on the overarching project goal and available computational tools is a non-trivial task. Although a number of review articles have been published to complete the task, they generally focus on specific aspects such as existing *de novo* pathway design tools [8,9,18], reconstructing metabolic pathways in organisms of interest [19], or identifying/refactoring parts and circuit designs beyond pathway prediction [20]. In this review, we discuss in detail all the steps involved in implementing pathway design algorithms (e.g., database construction, pathway ranking, enzyme selection, etc.). There exist several classifications of pathway design tools based on different aspects of the implementation procedures. For example, Koreta et al. classified these tools into reference-based, reaction-filling, and compound-filling frameworks [19]; Nakamura et al. classified them as fingerprint-based, maximum common substructure-based, and rule-based method [21]; Cho et al. classified them as chemical structural changes-based, enzymatic information-based, and reaction mechanism-based methods [22]. In this review, we choose to classify the tools based on their algorithmic choices such as graph theory, integer optimization, and retrosynthetic organic synthesis. In particular, the algorithms are classified into graph-based [23] (i.e., reactions and metabolites represented as a graph), stoichiometric-based [24] (i.e., reactions and metabolites represented using a stoichiometric matrix) and retrosynthesis-based [8] (i.e., iteratively identify reaction rules that can transform a reactant molecule) approaches. We also compare the pathway design algorithms based on their database curation, reaction network representation and its pruning, search algorithm, and pathway ranking methods used to prioritize the often-expansive list of possible pathways. As the next step for pathway design, we discuss the possibilities to apply protein engineering and *de novo* enzyme design tools to aid in protein design and discovery for the designed pathways. Finally, we highlight current limitations and explore potential applications of these tools.

2. Generalized *in silico* pathway design workflow

A generalized pathway design workflow highlighting the five steps is presented in Fig. 1: (1) database construction, (2) metabolic network representation, (3) network pruning, (4) search algorithm implementation, and (5) pathway ranking to select the best pathways of interest. In the following section, we discuss a number of pathway design algorithms that follow this design workflow (see Table 1) and highlight the challenges that potential new tools can be developed to tackle.

2.1. Databases

All pathway search tools rely on a database from which biochemical reactions and molecules can be recruited to constitute the pathway of interest. Currently, a number of databases have been constructed for known biochemical reactions and pathways (i.e., BIGG [25], KEGG [26], MetaCyc [27], BRENDA [28], ModelSEED [29], MetRxn [30], Rhea [31], UM-BBD [32], MOS [33], and Beilste Crossfile [34]), as well as hypothetical metabolites and reactions (such as ATLAS of Biochemistry [35], and MINE [36]) (see Fig. 1A). The current version of BIGG database consists of 80 manually curated organism-specific genome-scale metabolic models (GSMs) [25], while KEGG and MetaCyc catalog a more comprehensive array of organisms and their metabolic pathways [37]. KEGG contains 4102 more metabolites than MetaCyc while MetaCyc contains 3695 more reactions (as of August 20, 2017) [37]. BRENDA includes detailed enzyme information such as measured kinetic parameters [28], whereas ModelSEED provides the reaction mapping between KEGG and curated GSMs [29]. Existing databases sometimes have incorrect stoichiometries, imbalanced charges, redundancies due to molecule and reaction synonyms, as well as the lack of chemical structures. Since these databases are ultimately used for pathway design and hypothetical reaction rules construction, manual curation is often a necessary step in order to unify the metabolite and reaction names (to ensure network connectivity) and remove stoichiometrically imbalanced and redundant reactions [38,39]. Attempts to standardize reaction and metabolite name (e.g., MetRxn [30] and Rhea [31]) were previously made, however, automation of the process remains elusive and is further compounded by the continued discovery of new reactions and metabolites [40]. In a recent effort, MetaNetX [41] employed a reconciliation algorithm MNXref [42] to resolve the discrepancies in reaction and metabolite naming between GSMs. BKM-react matched metabolites and reactions by comparing InChI structures and compound synonyms [43]. RxnFinder used PubChem database to unify compound synonyms and added more than 50,000 reactions curated from literature to its in-house database [44]. Alternatively, the Chemical Translation Service [45] and UniChem [46] provide a simple web application that can interconvert metabolite IDs across different databases. Organism-specific GSMs or knowledge-bases (e.g., EcoCyc [47], AraCyc [48], and HumanCyc [49]) are often constructed or extracted from the larger databases, to ensure the design or identification of alternative pathways within an organism's native network. While certain tools limit the

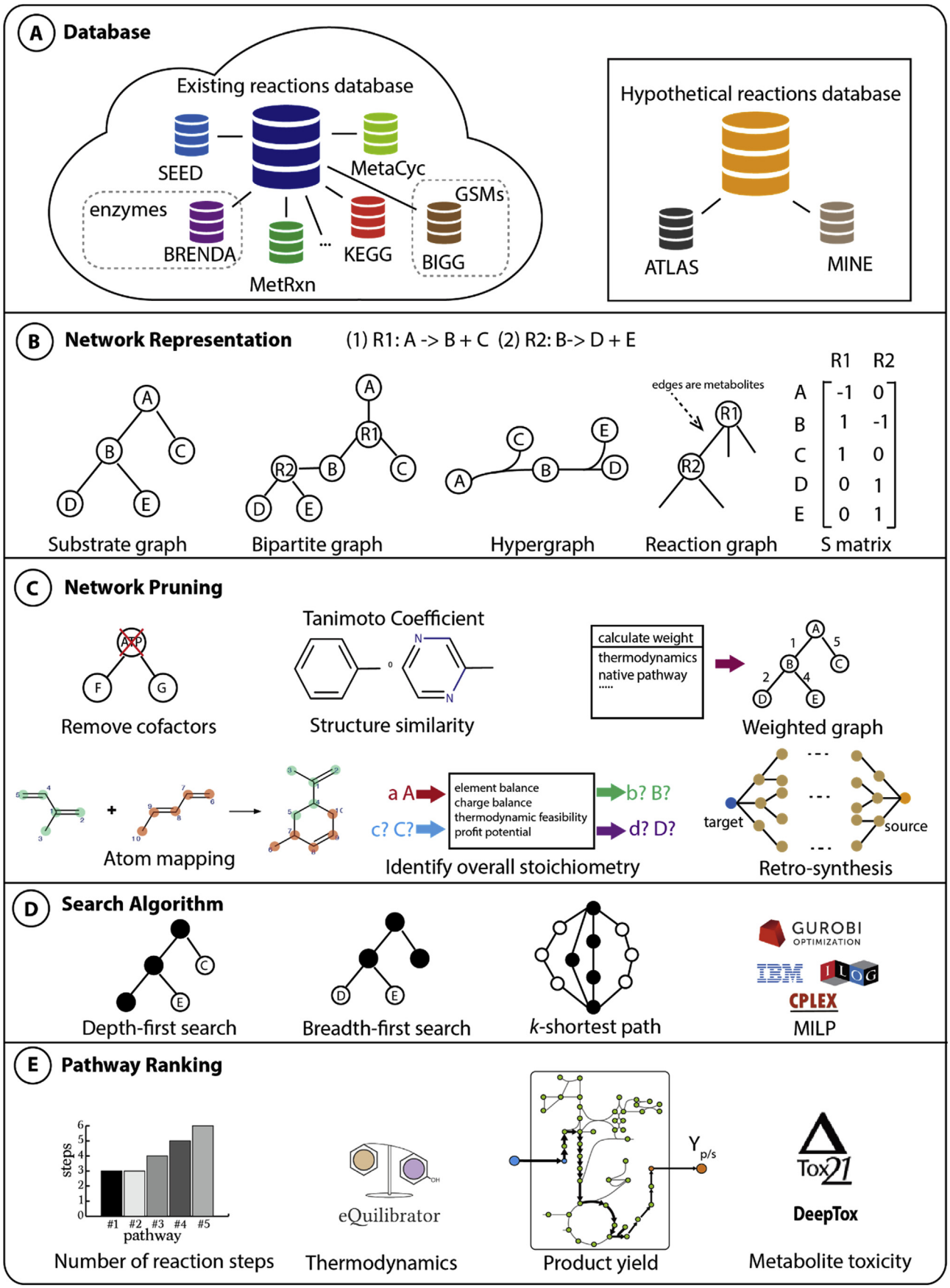

Fig. 1. A conceptualized pathway design workflow.

Table 1
Graph-based, Stoichiometry-based, and retrosynthesis-based pathway design tools and their characteristics.

Category	Name	Database	Network representation	Network pruning	Search algorithm	Pathway ranking	Reference
Graph-based	ReTrace	KEGG	Bipartite graph	Atom mapping	Heuristic search	Atom conservation and pathway length	[23]
	PathComp	KEGG	Substrate graph	–	Depth-first search (DFS)	–	[26]
	MetaRoute	KEGG	Reaction graph	Weighted graph and atom mapping	Eppstein's k-shortest path	Atom conservation and metabolite connectivity	[51]
	Pathway Hunter Tool	KEGG	Substrate graph	–	Breadth-first search (BFS) with (Higher-order horn logic) HOHL	Structure similarity and pathway length	[63]
	FMM	KEGG	Substrate graph	Manual cofactor removal	BFS	Compare pathway across organisms	[64]
	RouteSearch	MetaCyc	Substrate graph	Atom mapping	Branch and Bound	Atom conservation and pathway length	[65]
	MRE	KEGG	Substrate graph	Weighted graph	Yen's loopless k-shortest path	Thermodynamics and genes from host organism	[66]
	CMPF	KEGG, RPAIR	Bipartite graph	Weighted graph	Bounded depth path enumeration	Metabolite connectivity, reaction occurrence frequency, and pathway switching	[67]
	NeAT	MetaCyc	Bipartite graph	Weighted graph	Takahashi–Matsuyama, Pairwise K-shortest paths, and kWalks	Metabolite connectivity	[68]
	LPAT/BPAT	KEGG	Bipartite graph	Atom mapping	BPAT-M Search	Atom conservation and pathway length	[69,139]
	Rahnuma	KEGG	Hypergraph	Phylogeny or sub-network	DFS	–	[70]
	Metabolic Tinker	CHEBI, Rhea	Hypergraph	Weighted graph	Heuristic search	Pathway length, structure similarity, and thermodynamics	[71]
	FogLight	KEGG, MetaCyc	Hypergraph	And/Or graph	Brute-force search	Pathway length	[73]
	MRSD	KEGG	substrate graph	Weighted graph	Eppstein's k-shortest path	Reaction occurrence frequency	[78]
	DESHARKY	KEGG	–	Phylogeny	Monte Carlo	Metabolic burden	[88]
Stoichiometry-based	optStoic	KEGG, MetRxn	*S* matrix	Design overall stoichiometry	MILP	Pathway length or total metabolic flux	[24]
	PathTracer	BIGG, *i*JO1366	Substrate graph, *S* matrix	Atom mapping (MapMaker)	MILP	Pathway length or most active path	[50]
	CFP	BIGG	Substrate graph, *S* matrix	Atom mapping (carbon exchange network)	MILP	Pathway length	[75]
	METATOOL 5.0/k-shortest EFM	BIGG, *i*AF1260	*S* matrix	–	MILP	Pathway length	[76,89]
	OptStrain	KEGG	*S* matrix	–	MILP	Number of heterologous reactions	[99]
Retrosynthesis-based	Simpheny	BIGG	Substrate graph	molecule sizes	Retrosynthetic enumeration	Pathway length, thermodynamics, product yield, number of known metabolites/enzymes, and existence of reaction operators	[5]
	GEM-Path	BIGG, *i*JO1366	Substrate graph	Third level EC number and substrate similarity	Retrosynthetic enumeration	Thermodynamics and product yield	[57]
	XTMS/ RetroPath/ RetroPath 2.0	MetaCyc, BioCyc	*S* matrix	Molecular signature with predetermined distance	Retrosynthetic enumeration and MILP	Thermodynamics, gene prediction, pathway length, number of putative steps, and product yield	[52,82,84]
	BNICE	KEGG, ATLAS	Substrate graph	Qualitative/ Quantitative pruning	Retrosynthetic enumeration	Pruning criteria assessment (thermodynamics, pathway length, etc.)	[8,53]
	UM-PPS	UM-BBD	Substrate graph	Rule priority	Retrosynthetic enumeration	–	[56]
	PathPred	KEGG, RPAIR	Substrate graph	Structure similarity	Retrosynthetic enumeration	Compound similarity and pathway score	[54]
	Route Designer	MOS, Beilste Crossfile	Substrate graph	Heuristics and user defined limits	Retrosynthetic enumeration	Weighted function (wastage, example counts, and balanced disconnections.)	[55]
	SimIndex/ SimZyme	BRENDA	Substrate graph	Structure similarity	Byers–Waterman type pathway search	Pathway length	[83]
	Method by Cho et al.	KEGG	Substrate graph	–	Retrosynthetic enumeration	Combination of five priority factors	[22]

search space to only reactions within a particular organism (e.g., PathTracer [50] uses the *E. coli* GSM *i*JO1366), the search for a heterologous pathway would entail the use of a more comprehensive database encompassing multiple organisms, thereby ensuring that a desirable biotransformation (i.e., gene/enzyme) can be found (e.g., optStoic [24] and MetaRoute [51] use the curated KEGG database; XTMS [52] uses MetaCyc).

The potential of broad-substrate enzymes or synthetic enzymes to catalyze previously unknown or *de novo* reactions have also garnered the interest in using *de novo* pathway involving various non-natural molecules (e.g., pharmaceutical drugs). Such *de novo* pathways can be designed by exploiting the generalized reaction rules (e.g., ATLAS of Biochemistry [35]) which could act upon structurally similar metabolites in the databases (e.g., MINE [36]). Although hypothetical reactions have been generated for the implementation of most *de novo* pathway design algorithms, only two databases, namely ATLAS of Biochemistry and MINE, are currently available for public access. The database of hypothetical reactions can be developed using (but not limited to) five different reaction operators which encode chemical transformation mechanisms in a different manner as described here: (i) BNICE uses bond-electron matrix (BEM) to define non-bonded valence electrons and bond orders [53]; (ii) XTMS uses a molecular signature to generate reaction rules based on substructure of adjacent atoms [52]; (iii) PathPred uses the RDM pattern (developed by KEGG researchers) consisting of reaction center atom (R), atoms of different region (D), and atoms of the matched region (M) [54]; (iv) Route Designer applies a similar rule to that of RDM, by defining reaction core and extended reaction core with primary and secondary bonds and non-reacting neighborhood atoms [55]; (v) UM-PPS [56] and GEM-path [57] use SMIRKS and SMARTS, which exploit a feature string encoding the chemical properties of each atom. Reactions operators generally loose information while converting a known reaction to a rule due to the inherent assumptions in their method to encode chemical transformation mechanisms [58]. In particular, stereochemical changes are often overlooked, including BEM (in BNICE) and molecular signatures (in XTMS), as they do not contain chiral center information [58]. As a result, the predicted pathway would use stereoisomers (such as L-alanine and D-alanine) interchangeably. This would increase the number of potential pathways with several biologically incorrect predictions (i.e., subsequent reaction steps may use different stereoisomers). However, stereochemical changes have already been captured by many computational tools such as EC-Blast [59] and Reaction Decoder Tool [60], which use the Chemistry Development Kit (CDK) [61], and CLCA [62] to model stereochemical changes by appending stereochemical descriptors to the canonical labeling of each atom. It is therefore timely for reaction rule-based methods to incorporate such advanced descriptors to generate descriptions that are more detailed.

2.2. Representation of the database (metabolic network)

The curated reaction and metabolite database used for pathway search are henceforth denoted as a metabolic network in the text. Metabolic network (with and without hypothetical reactions) can be represented by a graph (i.e., substrate graph, bipartite graph, hypergraph, or reaction graph) or a stoichiometric matrix (*S* matrix) (see Fig. 1B). The vertices of substrate graph are metabolites and edges represent reactions, while the bipartite graph uses both metabolites and reactions as vertices with the edges connecting either a substrate to a reaction or a reaction to a product. Most of the pathway design tools are based on substrate graphs, namely Pathway Hunter Tools [63], FMM [64], RouteSearch [65], and MRE [66]. In contrast to the substrate graph, bipartite graph accounts for enzyme information in the vertices of reactions. Bipartite graph-based tools (such as CMPF [67] and NeAT [68]) enable tracking of reactions identified during pathway search thereby avoiding any post-processing step to link the identified edges to reactions as implemented in substrate graphs [23,67–69]. On the other hand, a hypergraph is a more direct representation of biochemical reaction wherein a hyper-edge (representing the reaction) connects all of its participating metabolites (represented as vertices). However, due to the dearth of sophisticated algorithms which can be applied to search hypergraphs, only three methods, namely Rahnuma [70], Metabolic Tinker [71] and the method developed by Carbonell et al. [72], use hypergraph directly, whereas another tool, Foglight simplifies the hypergraph into matrices before performing pathway search [73].

Although bipartite graph and hypergraph can be interconverted, it has been shown that bipartite graphs fail to ensure co-reactant availability while predicting a pathway feasibility unlike hypergraphs [74]. In addition, stoichiometry matrix representation of a metabolic network is equivalent to hypergraphs. The stoichiometry matrix and hypergraph representations have also been shown to be superior to the substrate or bipartite graphs as they retain (co) metabolite information from the original network [74]. Graph-based methods require an additional post-processing step to balance (co)metabolites of the identified pathway due to missing stoichiometry information, which was resolved in CFP [75] and PathTracer [50] by combining graph search with additional stoichiometry constraints to ensure the steady state of the identified pathways. Thus, with careful addition of co-reactant/co-products availability and their stoichiometry information, graph-based methods can make prediction with similar accuracy as stoichiometry-based methods. Moreover, stoichiometry-based pathway search methods can operate on the *S* matrix alone (such as METATOOL 5.0 [76], optStoic [24], and XTMS [52]) to identify cofactor-balanced pathways. However, the reversibility of a reaction has to be defined as constraints alongside the *S* matrix while a directed graph inherently contains the information.

2.3. Network pruning

Graph-based methods search for pathways from a given source metabolite to a target metabolite by looking for adjacent reactions which share metabolites as done in PathComp [26]. However, this procedure often arrives at irrelevant biological transitions due to the overwhelming participation of cofactors in metabolic networks as highlighted by Rahman et al. [63]. These procedures rely on substrate graph based metabolic network representation which maps all reactants to all products of a reaction in the graph. However, this representation also connects metabolites that do not exchange any carbon atoms but are participants of the same reaction (e.g., ADP to pyruvate). A simple resolution to this problem is the exclusion of cofactors and other highly connected metabolites (hub metabolites) from the search space (see Fig. 1C), but this option could miss pathways such as nucleotide biosynthesis which involve hub metabolites such as ADP as major intermediates [51]. A more systematic approach involves incorporation of structural similarity between the intermediate metabolites to guide pathway search by using a 1-D chemical fingerprint of the metabolites [63]. Alternatively, this can be achieved by using weighted edges where the network hub metabolites (such as ATP, NAD, etc. that have high participation) can be penalized [77]. The pathway searches can also be made biologically relevant by weighing the reactions edges in the graphs with more information. This was achieved in MRSD using reaction occurrence frequency across multiple organisms as reaction weights to account for biochemical transformations which are conserved across multiple species [78]. Similarly, reaction with more negative Gibbs-free energy can be assigned larger weight

according to the thermodynamic favorability-based weighting scheme used in MRE [66]. However, all these approaches do not track the atoms from the substrate which are lost during the transformation to the target metabolites. By including the atom conservation criteria, Route Search [65] enables us to measure the fraction of the carbon atoms from substrates which are lost by the pathway while producing the target metabolite, thus capturing the efficiency of the discovered pathway. The network pruning steps can also take advantage of more systematic information of chemical structure such as atom mapping and KEGG RPAIR database [79]. Atom mapping methods map the transfer of C, O, N, P, S atoms between metabolites in a given reaction. Thus, atom-mapping rules for reactions can also be incorporated to ensure the chemical feasibility of the identified pathways as done by MetaRoute [51], CFP [75], PathTracer [50], AGPathFinder [80], and RouteSearch [65]. KEGG RPAIR data offers a manually curated catalog of main and side metabolites thus avoiding irrelevant biological transitions [79].

Stoichiometry-based methods employ flux balance analysis (FBA) during pathway search which only identifies mass balanced pathways [23]. However, the mass balance restrictions can be relaxed by allowing for cofactors, co-reactants or co-products to be exchanged with the environment which reflects biological reality where pathways do not exist in isolation and exchange metabolites with their surrounding or other pathways. Moreover, the stoichiometry can also be predefined as employed in the first step of the optStoic algorithm [24] to establish an overall stoichiometry design goal that is necessary for the selection of cofactors and co-reactants (see Fig. 1C). Upon identifying the overall stoichiometry equation, the stoichiometric coefficients of the reactants and products are fixed as “uptake” and “secretion” flux of the network. The mixed-integer linear programming (MILP)-based minFlux [24] (i.e., minimize total flux through the network) or minRxn [24] (i.e., minimize the total number of reactions) formulation can then be used to identify an internal network of reactions that could convert the reactants to the products in a mass-balanced manner. Unlike the CFP [75] or the PathTracer [50] approach which generates carbon exchange networks *a priori*, the minFlux/minRxn [24] formulation uses a metabolic network identical to that of a typical FBA analysis and can be easily extended to any currently available GSMs. Moreover, CFP-related approaches might predict pathways with biologically irrelevant carbon (or other elemental) exchanges [75,81] due to inaccuracies in the carbon exchange network, which can be resolved using more sophisticated atom-mapping algorithms [81].

In addition to the network pruning steps of existing metabolic networks, retrosynthesis-based approaches for designing *de novo* pathway require the generation of a metabolic network based on all the hypothetical reaction rules derived using reaction operators (see section 2.1 and Fig. 1C). This hypothetical network can be appended to the network of known reactions, thereby allowing the search of both putative and verified reactions. However, the extended metabolic network is often too large for exhaustive pathway exploration. For example, the initial BNICE computational framework generates an exponentially growing range of hypothetical molecules [53]. In order to reduce the search space, XTMS/RetroPath uses a diameter which defines graph distance of atoms within the radius to control the network size [52,82]. UM-PPS applies reactions rules based on an ‘absolute aerobic likelihood’ to prune unlikely biotransformation thus avoiding exploration of redundant reaction network [56]. The number of reaction rules that can act upon a metabolite can also be culled based on the availability of (broad-substrate or promiscuous) reactions/enzymes (e.g., GEM-Path uses third level EC number [57]; SimZyme/SimIndex quantifies a molecule's similarity to the typical substrate of an enzyme [83]; RetroPath 2.0 uses enzyme score [84]), molecule sizes (e.g. SimPheny), and expert knowledge such as THERESA [85].

2.4. Search algorithms

The selection of pathway search algorithm is inherently dependent on the underlying representation of the metabolic network (see Fig. 1D). Breadth-first search (BFS) is a widely used algorithm to find the k-shortest paths in a loopless unweighted graph as applied in the Pathway Hunter Tool [63]. As many pre-processing steps assign weights to a graph based on thermodynamics or other criteria, the search requires algorithms such as Yen's k-shortest path algorithm [86] that works on loopless graphs, and Eppstein's k-shortest path algorithm and its modifications [87] that do not require the graph to be loopless. In addition to the weighted graph with a fixed cost at each edge, RouteSearch defined an additional criterion at each edge for non-static atom lost and applied branch-and-bound search to find the best paths to minimize the loss [65]. On the other hand, DESHARKY [88] applied Monte Carlo method to search for reaction combinations. To find the shortest path by *S* matrix representation, stoichiometry-based pathway design algorithms that use MILP is the common approach (e.g., k-shortest Elementary Flux Modes (k-shortest EFMs) [89] and optStoic [24]). k-shortest EFM has been shown to provide more accurate pathway designs from fatty-acids to glucose than graph-based methods such as Pathway Hunter Tools [90–93]. Stoichiometry-based methods are better than graph-based methods as they account for mass balance constraints by directly incorporating the stoichiometry information. Currently, MILP can be solved by many open-source solvers (e.g., SCIP [94]) and commercial solvers (e.g., CPLEX [95] and GUROBI [96]), which employ algorithms such as Branch-and-Bound and Branch-and-Cut along with customized heuristic searches. Alternative pathways can be also identified by adding integer cut constraints.

The selection of search algorithms also relies on the desired type of pathways, namely linear or branching pathways. The above-mentioned graph-based algorithms only search for linear pathways with one source molecule and one target molecule. In order to identify branching pathways, ReTrace [23] combines shortest paths into branched pathways to reach a higher fraction of atom transfer from source to target metabolite. LPAT [69] developed another linear pathway merging algorithm BPAT-M using atom tracking information. In addition, graph-based algorithms that can efficiently find paths from two nodes (e.g. A*) can be adapted to recursively search the relevant branched sub-paths without atom mapping information. In contrast to graph-based algorithms, MILP algorithms can find branched or even cyclic pathways [24].

2.5. Pathway ranking

Pathway design tools often identify multiple pathways for a given substrate and metabolite pair which can be distinguished based on several factors such as host compatibility, availability of natural enzymes (or protein engineering), and protein solubility (see Fig. 1E). The most common method used to rank pathways is by the number of reaction steps, as this can be easily translated into an objective function in a number of methods (e.g., optStoic [24], CFP [75], and k-shortest EFM [89], FindPath [97]) to find the shortest pathway or pathway with the least total flux. The shortest pathway also implies fewest reaction steps or minimal enzyme requirement, thereby reducing the metabolic/genetic burden on the host cells. This is based on the assumption that each reaction is catalyzed by a single gene, which however is not always true [98]. Reduced number of genetic modifications also enables a faster and simpler experimental implementation. Alternately, one could directly aim for the minimal number of genes as the objective (e.g.,

SimOptStrain [98] identifies the minimal number of genetic interventions). Likewise, if the host organism is predefined, then it is also possible to minimize the number of heterologous reactions that need to be added (e.g., in OptStrain [99] and MRE [66]). This is a plausible objective as dealing with heterologous enzymes in metabolic engineering project often poses a different set of challenges including that of enzyme activity, protein solubility, codon optimization, and foreign cofactor utilization. In particular, for rule-based approaches, if an existing enzyme that can perform the biotransformation is not known, then it is often required to search for a natural promiscuous enzyme for the substrate of interest or even design a *de novo* protein [100]. Despite various successful cases (e.g., Merck & Co.'s *in vitro* sitagliptin synthesis [7] and deep learning to rank the most suitable reaction rules [101]), protein engineering to confer a novel enzyme activity is a time-consuming effort with an uncertain outcome. Therefore, when using a rule-based pathway design approach, it is common to rank pathways based on the number of known enzymes [5].

Thermodynamic feasibility is another commonly used method for pathway ranking. Similar to the preprocessing step that assigns weight to a graph to prune infeasible pathways and select pathway with more negative ΔG, the designed pathways can be sorted based on their most negative overall ΔG, which sums up the ΔG of each reaction step. Group Contribution Method [102], and a recently developed and publicly available Component Contributions Method [103] or eQuilibrator [104] can be used to estimate the standard transformed Gibbs free energy of reaction under the host cell environment (e.g., cellular compartment pH, growth temperature). Knowledge of the intracellular metabolite concentrations can be used to further refine the estimation of the actual Gibbs free energy of reaction, but it is generally not used due to the lack of metabolome data. Instead, the Max-min Driving Force [105] approach can be used to optimize the concentrations of metabolites within a pathway given the physiological concentration ranges and quantify the thermodynamic feasibility of the pathway. Although the availability of intracellular metabolite concentrations is often limited, this could be overcome by an approach that uses the support vector machine (SVM) model to infer the theoretical intracellular metabolite concentration [106].

In order to rank pathway based on product yield, a designed pathway can be introduced into the GSM of the host strain and FBA can be performed to identify the maximum achievable yield from the pathway [3,57,107]. This is particularly important as many pathway design tools often only target a short pathway from any precursor metabolite that can be produced by a host cell to the target product. However, in addition to simulating whether a carbon source (e.g., glucose) could drive flux towards the target product, FBA also ensures all cofactors/co-substrates and biomass of the cell could be produced and every reaction (native or heterologous) used in the identified pathway must carry non-zero flux [50,108]. Another possible method is by constructing kinetic models of the designed pathway and calculating the flux through the pathway. This method is used recently to evaluate a large number of trunk glycolytic pathways [109]. However, the paucity of kinetic parameters could hamper such an approach. Alternately, an ensemble of kinetic parameters can be sampled and used to determine the stability of the pathway (EMRA) [110]. These approaches are however more computationally demanding and may not be suitable for initial filtering of a large number of the designed pathways [111,112]. A simpler more tractable approach based on modular kinetic rate law [113–115] can be applied to evaluate the protein cost of a pathway as an alternate pathway ranking criteria.

Other possible approaches include a scoring system based on assessing the toxicity of intermediate metabolites of a pathway in a host cell [116,117]. For example, a database like Tox21 [118] and a deep learning based algorithm (DeepTox) [119] have been applied to identify potentially toxic effects of chemical compounds. The selection of pathway ranking method(s) depends on the design goal. For example, in order to design a pathway for the production of a certain biomolecule, one may consider thermodynamics, theoretical yield from FBA simulation, the minimal number of reactions, and toxicity of intermediate metabolite on host cell as primary ranking criteria. Often, a combination of these filtering, ranking or scoring systems can be used as demonstrated by Yim et al. [5].

3. DNA sequence selection, protein engineering, and *de novo* enzyme design

The selection and design of DNA sequence still remain elusive for pathway designs. Most of the pathway design tools identify the list of reactions that are needed to fill the gap between source and target metabolites. However, there is currently a large natural catalog of enzyme sequences from which the user has to select to express the pathway. A number of pathway design tools prioritize the selection based on binding site covalence, chemical similarity, and organism specificity [22]. Additional screening criteria such as protein solubility in the host system can also be employed to refine the sequence selection. Generally, the host organism is known *a priori*, and native enzymes are assigned higher priority and the minimal number of heterologous reactions are selected from closely related species. However, in certain projects where the target metabolite is not common (e.g., xenobiotics), it is necessary to select a host organism based on the pathways that are identified. In addition, promiscuous enzymes are required to perform the biotransformation in the reactions without natural enzymes. The *in vivo* discovery of such enzymes is a daunting task and requires the assistance of computational prediction tools. For example, Carbonell et al. [120] performed a machine learning-based promiscuity analysis to predict if a reaction rule can be catalyzed by a natural enzyme. However, the prediction relies on manually defined promiscuity instead of actual *in vivo* data. Supervised machine learning algorithms can give better predictions with more correctly labeled “big data”. Hence, a database of DNA sequences and the corresponding enzyme substrate and enzyme activity (such as BRENDA [28] and SABIO-RK [121]) should be included in machine learning workflow to predict enzyme-substrate pairs with high likelihood for interaction.

When natural enzymes fail to perform the predicted reaction steps, protein engineering fills the void by altering existing enzyme activity and specificity [100] and ultimately designing *de novo* enzymes. Computational protein engineering tools can guide rational protein design and facilitate the efforts involving random mutagenesis-based directed evolution. Two widely applied strategies are used to predict protein designs: (i) statistical methods to compare modified sequences with sequence database (e.g. GenBank [122]), and (ii) molecular modeling methods to take advantage of the atom-level structural information to predict its biochemical properties. Taking advantage of both these methods, Pantazes et al. [123–125] developed an Iterative Protein Redesign and Optimization (IPRO) suite which applies the workflow of alternating protein backbone perturbations and amino acid sequence mutations to design proteins with desired catalytic activity. In addition, a number of computational tools (e.g. DEZYMER [126,127], ORBIT [128], ROSETTA [129], CCBuilder [130,131], and Protein WISDOM [132]) are aimed at *de novo* enzyme design by exploiting the underlying protein biochemistry and biophysics. However certain enzyme properties such as configurational entropy changes are beyond the scope of computational tools [15], thus in practice, the computational predictions are often

complemented with directed evolution methods to provide a starting design for *in vitro* improvement. For example, Rothlisberger et al. [15] used directed evolution after the molecular modeling method to further improve the enzyme's catalytic efficiencies. In spite of the numerous successes in engineering and designing new proteins, protein engineering and design tools have not been integrated with traditional pathway design tools as the latter focus on naturally occurring enzymes. However, retrosynthesis-based pathway design tools often propose novel biotransformation by natural enzymes based on the structural similarities between the enzyme's native substrate and the new substrate (e.g. SimZyme [83]). The identified enzyme can serve as a starting point for the protein engineering and design tools to further its catalytic activity towards the novel biotransformation. For detailed tools and applications of protein engineering for pathways, we refer the reader to the following resources [133–136].

4. Perspective

In this paper, we have compared different computational approaches used to design metabolic pathways in terms of the database used, its representation and pruning, search algorithm and pathway ranking. The graph-based methods often need additional post-processing steps to balance co-metabolites of predicted pathways that may be unbalanced. Stoichiometry-based methods avoid the preprocessing steps to remove no-carbon transfer connections because the *S* matrix representation accounts for all participating metabolites of a reaction similar to hypergraph representation. In addition, stoichiometry-based methods can incorporate pathway ranking criteria, such as thermodynamics, relative cost, and pathway length into the MILP optimization framework as an objective function thus homing at first at the most desirable designs avoiding exhaustive enumeration of pathways and ranking them *a posteriori*. Retrosynthesis-based methods employ similar search methods as used by graph-based or stoichiometry-based methods to design pathways by searching through an extended graph or stoichiometric network. Out of the reviewed retrosynthesis-based pathway design tools, only XTMS uses stoichiometry-based EFM tools to search for pathways while other methods rely on graph-based tools [52]. The performance of *de novo* pathway tools can be improved by switching to stoichiometry-based tools in order to circumvent unbalanced pathways and pathway post-processing that are inherent in graph-based methods. Although XTMS [52] enumerates pathways based on EFMs, it does not directly consider the stoichiometry of the source and target metabolites. A retrosynthetic tool that can design stoichiometry *a priori* based on the first step of optStoic [24] and identify cofactor balanced pathways by design while limiting the number of novel reaction steps would be an important advancement for *de novo* pathway design.

Nevertheless, it is worth mentioning that stoichiometry-based methods are not without their challenges such as longer computational time and the possible presence of thermodynamically infeasible cycles within designed pathways. The computational time depends highly on the objective function and constraints (or integer cuts) that are imposed. For example, the minRxn formulation of optStoic requires significantly higher computational cost than the minFlux formulation [24]. Furthermore, the computational time also scales with the search space, which can be resolved by first removing blocked reactions (i.e., reaction that could not carry any flux under a specific condition) from the *S* matrix. Alongside formulating the strongest MILP problem, a number of heuristics based on branch-and-bound and branch-and-cut methods are available and are under-development in open-source, academic, and commercial software packages to improve the solving time [137]. Stoichiometry-based methods sometimes identify futile cycles to balance cofactors. In CFP [75] and PathTracer [50], this is remedied by preventing flux from re-visiting a metabolite (node), whereas an updated version of optStoic is currently in development to resolve this issue in a systematic manner (Ng, Chowdhury, Maranas, *manuscript in preparation*).

Despite the success of current computational design tools to identify pathways, challenges still remain on the selection of genes, discovery of promiscuous enzymes, engineering proteins, and designing *de novo* enzymes to catalyze putative reactions. Overall, a completely automated pipeline that goes from selection of source and target molecule to the final output of DNA sequences of a pathway would significantly facilitate the discovery of new metabolic pathways for various applications. Experimental validation of multiple pathway designs has already been accelerated through an automated digital-to-biological DNA manufacturing system [138]. Although there remain several limitations that need to be addressed, exemplary efforts such as RetroPath 2.0 [84] and ATLAS [35] provide a benchmark that can be improved upon to realize the ultimate goal.

Acknowledgment

The authors gratefully acknowledge funding from the DOE (http://www.energy.gov/) grant no. DE-SC0008091 and NSF (http://www.nsf.gov/) award no. EEC-0813570 and no. NSF/MCB-1546840.

References

[1] Wood HG. Life with Co or Co2 and H2 as a source of carbon and energy. Faseb J 1991;5(2):156–63.
[2] Chavez S, et al. The presence of glutamate dehydrogenase is a selective advantage for the Cyanobacterium synechocystis sp. strain PCC 6803 under nonexponential growth conditions. J Bacteriol 1999;181(3):808–13.
[3] Moura M, et al. Evaluating enzymatic synthesis of small molecule drugs. Metab Eng 2016;33:138–47.
[4] Galanie S, et al. Complete biosynthesis of opioids in yeast. Science 2015;349(6252):1095–100.
[5] Yim H, et al. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat Chem Biol 2011;7(7):445–52.
[6] Blass LK, Weyler C, Heinzle E. Network design and analysis for multi-enzyme biocatalysis. BMC Bioinforma 2017;18(1):366.
[7] Savile CK, et al. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science 2010;329(5989):305–9.
[8] Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Curr Opin Chem Biol 2015;28:99–104.
[9] Prather KL, Martin CH. De novo biosynthetic pathways: rational design of microbial chemical factories. Curr Opin Biotechnol 2008;19(5):468–74.
[10] Nam H, et al. Network context and selection in the evolution to enzyme specificity. Science 2012;337(6098):1101–4.
[11] Khare SD, et al. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat Chem Biol 2012;8(3):294–300.
[12] Ekroos M, Sjogren T. Structural basis for ligand promiscuity in cytochrome P450 3A4. Proc Natl Acad Sci U. S. A 2006;103(37):13682–7.
[13] Coelho PS, et al. Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome P450 enzymes. Science 2013;339(6117):307–10.
[14] Young EM, et al. Rewiring yeast sugar transporter preference through modifying a conserved protein motif. Proc Natl Acad Sci U. S. A 2014;111(1):131–6.
[15] Rothlisberger D, et al. Kemp elimination catalysts by computational enzyme design. Nature 2008;453(7192):190–5.
[16] Burgard A, et al. Development of a commercial scale process for production of 1,4-butanediol from sugar. Curr Opin Biotechnol 2016;42:118–25.
[17] Libis V, Delepine B, Faulon JL. Expanding biosensing abilities through computer-aided design of metabolic pathways. ACS Synth Biol 2016;5(10):1076–85.
[18] Fernandez-Castane A, et al. Computer-aided design for metabolic engineering. J Biotechnol 2014:302–13. 192 Pt B.
[19] Kotera M, Goto S. Metabolic pathway reconstruction strategies for central metabolism and natural product biosynthesis. Biophys Physicobiol 2016;13:195–205.

[20] Medema MH, et al. Computational tools for the synthetic design of biochemical pathways. Nat Rev Microbiol 2012;10(3):191–202.
[21] Nakamura M, et al. An efficient algorithm for de novo predictions of biochemical pathways between chemical compounds. Bmc Bioinforma 2012;13.
[22] Cho A, et al. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Syst Biol 2010;4:35.
[23] Pitkanen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol 2009;3:103.
[24] Chowdhury A, Maranas CD. Designing overall stoichiometric conversions and intervening metabolic reactions. Sci Rep 2015:5.
[25] King ZA, et al. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res 2016;44(D1):D515–22.
[26] Goto S, et al. Organizing and computing metabolic pathway data in terms of binary relations. Pac Symp Biocomput 1997:175–86.
[27] Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2016;44(D1):D471–80.
[28] Scheer M, et al. BRENDA, the enzyme information system in 2011. Nucleic Acids Res 2011;39:D670–6.
[29] Henry CS, et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 2010;28(9). 977-U22.
[30] Kumar A, Suthers PF, Maranas CD. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. BMC Bioinforma 2012;13:6.
[31] Alcantara R, et al. Rhea-a manually curated resource of biochemical reactions. Nucleic Acids Res 2012;40(D1):D754–60.
[32] Gao J, Ellis LB, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. Nucleic Acids Res 2010;38(Database issue):D488–91.
[33] Sutter J, et al. New features that improve the pharmacophore tools from Accelrys. Curr Comput Aided Drug Des 2011;7(3):173–80.
[34] Vanco J. The beilstein CrossFire information system and its use in pharmaceutical chemistry. Ceska Slov Farm 2003;52(2):68–72.
[35] Hadadi N, et al. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. Acs Synth Biol 2016;5(10):1155–66.
[36] Jeffryes JG, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Cheminformatics 2015;7.
[37] Altman T, et al. A systematic comparison of the MetaCyc and KEGG pathway databases. Bmc Bioinforma 2013;14.
[38] Chan, S.H., et al., Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics.
[39] Dash S, Ng CY, Maranas CD. Metabolic modeling of clostridia: current developments and applications. FEMS Microbiol Lett 2016;363(4).
[40] Mukherjee S, et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol 2017;35(7):676–83.
[41] Moretti S, et al. MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res 2016;44(D1):D523–6.
[42] Bernard T, et al. Reconciliation of metabolites and biochemical reactions for metabolic networks. Brief Bioinform 2014;15(1):123–35.
[43] Lang M, Stelzer M, Schomburg D. BKM-react, an integrated biochemical reaction database. BMC Biochem 2011;12:42.
[44] Hu QN, et al. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. Bioinformatics 2011;27(17):2465–7.
[45] Wohlgemuth G, et al. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. Bioinformatics 2010;26(20):2647–8.
[46] Chambers J, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform 2013;5(1):3.
[47] Keseler IM, et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res 2017;45(D1):D543–50.
[48] Mueller LA, Zhang P, Rhee SY. AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol 2003;132(2):453–60.
[49] Romero P, et al. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 2005;6(1):R2.
[50] Tervo CJ, Reed JL. MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. Biotechnol J 2016;11(5):648–61.
[51] Blum T, Kohlbacher O. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics 2008;24(18):2108–9.
[52] Carbonell P, et al. XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res 2014;42(Web Server issue):W389–94.
[53] Hatzimanikatis V, et al. Exploring the diversity of complex metabolic networks. Bioinformatics 2005;21(8):1603–9.
[54] Moriya Y, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res 2010;38(Web Server issue):W138–43.
[55] Law J, et al. Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. J Chem Inf Model 2009;49(3):593–602.
[56] Gao J, Ellis LB, Wackett LP. The university of Minnesota pathway prediction system: multi-level prediction and visualization. Nucleic Acids Res 2011;39(Web Server issue):W406–11.
[57] Campodonico MA, et al. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. Metab Eng 2014;25:140–58.
[58] Shin JH, et al. Production of bulk chemicals via novel metabolic pathways in microorganisms. Biotechnol Adv 2013;31(6):925–35.
[59] Rahman SA, et al. EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods 2014;11(2):171–4.
[60] Rahman SA, et al. Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics 2016;32(13):2065–6.
[61] Steinbeck C, et al. The chemistry development Kit (CDK): an open-source java library for chemo- and bioinformatics. J Chem Inf Comput Sci 2003;43(2):493–500.
[62] Kumar A, Maranas CD. CLCA: maximum common molecular substructure queries within the MetRxn database. J Chem Inf Model 2014;54(12):3417–38.
[63] Rahman SA, et al. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics 2005;21(7):1189–93.
[64] Chou CH, et al. FMM: a web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res 2009;37(Web Server issue):W129–34.
[65] Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. Bioinformatics 2014;30(14):2043–50.
[66] Kuwahara H, et al. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. Nucleic Acids Res 2016;44(W1):W217–25.
[67] Lim K, Wong L. CMPF: class-switching minimized pathfinding in metabolic networks. BMC Bioinforma 2012;13(Suppl 17):S17.
[68] Faust K, et al. Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 2010;26(9):1211–8.
[69] Heath AP, Bennett GN, Kavraki LE. An algorithm for efficient identification of branched metabolic pathways. J Comput Biol 2011;18(11):1575–97.
[70] Mithani A, Preston GM, Hein J. Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. Bioinformatics 2009;25(14):1831–2.
[71] McClymont K, Soyer OS. Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways. Nucleic Acids Res 2013;41(11):e113.
[72] Carbonell P, et al. Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC Syst Biol 2012;6:10.
[73] Khosraviani M, Saheb Zamani M, Bidkhori G. FogLight: an efficient matrix-based approach to construct metabolic pathways by search space reduction. Bioinformatics 2016;32(3):398–408.
[74] Klamt S, Haus UU, Theis F. Hypergraphs and cellular networks. PLoS Comput Biol 2009;5(5), e1000385.
[75] Pey J, et al. Path finding methods accounting for stoichiometry in metabolic networks. Genome Biol 2011;12(5):R49.
[76] von Kamp A, Schuster S. Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics 2006;22(15):1930–1.
[77] Croes D, et al. Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res 2005;33(Web Server issue):W326–30.
[78] Xia D, et al. MRSD: a web server for metabolic route search and design. Bioinformatics 2011;27(11):1581–2.
[79] Faust K, Croes D, van Helden J. Metabolic pathfinding using RPAIR annotation. J Mol Biol 2009;388(2):390–414.
[80] Huang YR, et al. A method for finding metabolic pathways using atomic group tracking. Plos One 2017;12(1).
[81] Pey J, Planes FJ, Beasley JE. Refining carbon flux paths using atomic trace data. Bioinformatics 2014;30(7):975–80.
[82] Carbonell P, et al. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst Biol 2011;5:122.
[83] Pertusi DA, et al. Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics 2015;31(7):1016–24.
[84] Delépine B, et al. RetroPath2.0: a retrosynthesis workflow for metabolic engineers. bioRxiv; 2017. p. 141721.
[85] Liu M, et al. Combining chemoinformatics with bioinformatics: in silico prediction of bacterial flavor-forming pathways by a chemical systems biology approach “reverse pathway engineering”. PLoS One 2014;9(1):e84769.
[86] Yen JY. Finding K shortest loopless paths in a network. Manag Sci Ser a-Theory 1971;17(11):712–6.
[87] Eppstein D. Finding the k shortest paths. Siam J Comput 1998;28(2):652–73.
[88] Rodrigo G, et al. DESHARKY: automatic design of metabolic pathways for optimal cell growth. Bioinformatics 2008;24(21):2554–6.
[89] de Figueiredo LF, et al. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics 2009;25(23):3158–65.
[90] de Figueiredo LF, et al. Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'. Bioinformatics 2009;25(24):3330–1.
[91] Faust K, Croes D, van Helden J. In response to 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'. Bioinformatics 2009;25(23):3202–5.
[92] de Figueiredo LF, et al. Can sugars be produced from fatty acids? A test case

for pathway analysis tools. Bioinformatics 2009;25(1):152–8.
[93] de Figueiredo LF, et al. Can sugars be produced from fatty acids? A test case for pathway analysis tools. Bioinformatics 2008;24(22):2615–21.
[94] Maher SJ, et al. The SCIP optimization suite 4.0. 2017.
[95] Optimizer IIC. 12.6. 2. IBM ILOG. 2015.
[96] Gurobi Optimization I. Gurobi optimizer reference manual. 2016.
[97] Vieira G, et al. FindPath: a Matlab solution for in silico design of synthetic metabolic pathways. Bioinformatics 2014;30(20):2986–8.
[98] Kim J, Reed JL, Maravelias CT. Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. PLoS One 2011;6(9):e24162.
[99] Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. Genome Res 2004;14(11):2367–76.
[100] Siegel JB, et al. Computational protein design enables a novel one-carbon assimilation pathway. Proc Natl Acad Sci U. S. A 2015;112(12):3704–9.
[101] Segler MHS, Waller MP. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry 2017;23(25):5966–71.
[102] Jankowski MD, et al. Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J 2008;95(3):1487–99.
[103] Noor E, et al. Consistent estimation of Gibbs energy using component contributions. PLoS Comput Biol 2013;9(7):e1003098.
[104] Flamholz A, et al. eQuilibrator–the biochemical thermodynamics calculator. Nucleic Acids Res 2012;40(Database issue):D770–5.
[105] Noor E, et al. Pathway thermodynamics highlights kinetic obstacles in central metabolism. PLoS Comput Biol 2014;10(2):e1003483.
[106] Yang HF, et al. Theoretical studies of intracellular concentration of microorganisms' metabolites. Sci Rep 2017;7(1):9048.
[107] Wu D, et al. A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. J Chem Inf Model 2011;51(7):1634–47.
[108] Zhang XL, Tervo CJ, Reed JL. Metabolic assessment of E. coli as a Biofactory for commercial products. Metab Eng 2016;35:64–74.
[109] Court SJ, Waclaw B, Allen RJ. Lower glycolysis carries a higher flux than any biochemically possible alternative. Nat Commun 2015;6:8427.
[110] Lee Y, Lafontaine Rivera JG, Liao JC. Ensemble Modeling for Robustness Analysis in engineering non-native metabolic pathways. Metab Eng 2014;25:63–71.
[111] Khodayari A, et al. A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data. Metab Eng 2014;25:50–62.
[112] Khodayari A, Maranas CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat Commun 2016;7:13806.
[113] Liebermeister W, Uhlendorf J, Klipp E. Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. Bioinformatics 2010;26(12):1528–34.
[114] Flamholz A, et al. Glycolytic strategy as a tradeoff between energy yield and protein cost. Proc Natl Acad Sci U. S. A 2013;110(24):10039–44.
[115] Noor E, et al. The protein cost of metabolic fluxes: prediction from enzymatic rate laws and cost minimization. Plos Comput Biol 2016;12(11).
[116] Planson AG, et al. Compound toxicity screening and structure-activity relationship modeling in Escherichia coli. Biotechnol Bioeng 2012;109(3):846–50.
[117] Planson AG, et al. A retrosynthetic biology approach to therapeutics: from conception to delivery. Curr Opin Biotechnol 2012;23(6):948–56.
[118] Tice RR, et al. Improving the human hazard characterization of chemicals: a Tox21 update. Environ Health Perspect 2013;121(7):756–65.
[119] Mayr A, et al. DeepTox: toxicity prediction using deep learning. Front Environ Sci 2016;3:80.
[120] Carbonell P, Faulon JL. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics 2010;26(16):2012–9.
[121] Wittig U, et al. SABIO-RK–database for biochemical reaction kinetics. Nucleic Acids Res 2012;40(Database issue):D790–6.
[122] Benson DA, et al. GenBank. Nucleic Acids Res 2017;45(D1):D37–42.
[123] Pantazes RJ, et al. The iterative protein redesign and optimization (IPRO) suite of programs. J Comput Chem 2015;36(4):251–63.
[124] Fazelinia H, Cirino PC, Maranas CD. Extending Iterative Protein Redesign and Optimization (IPRO) in protein library design for ligand specificity. Biophys J 2007;92(6):2120–30.
[125] Saraf MC, et al. IPRO: an iterative computational protein library redesign and optimization procedure. Biophys J 2006;90(11):4167–80.
[126] Hellinga HW, Richards FM. Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with predefined geometry. J Mol Biol 1991;222(3):763–85.
[127] Hellinga HW, Caradonna JP, Richards FM. Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into Escherichia coli thioredoxin. J Mol Biol 1991;222(3):787–803.
[128] Dahiyat BI, Mayo SL. Protein design automation. Protein Sci 1996;5(5):895–903.
[129] Zanghellini A, et al. New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 2006;15(12):2785–94.
[130] Wood CW, Woolfson DN. CCBuilder 2.0: powerful and accessible coiled-coil modeling. Protein Sci 2017;00. 00–00.
[131] Wood CW, et al. CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. Bioinformatics 2014;30(21):3029–35.
[132] Smadbeck J, et al. Protein WISDOM: a workbench for in silico de novo design of biomolecules. J Vis Exp 2013;77:50476.
[133] Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature 2016;537(7620):320–7.
[134] Pleiss J. Protein design in metabolic engineering and synthetic biology. Curr Opin Biotechnol 2011;22(5):611–7.
[135] Damborsky J, Brezovsky J. Computational tools for designing and engineering biocatalysts. Curr Opin Chem Biol 2009;13(1):26–34.
[136] Eriksen DT, Lian J, Zhao H. Protein design for pathway engineering. J Struct Biol 2014;185(2):234–42.
[137] Klotz E, Newman AM. Practical guidelines for solving difficult mixed integer linear programs. Surv Operat. Res Manag Sci 2013;18(1):18–32.
[138] Boles KS, et al. Digital-to-biological converter for on-demand production of biologics. Nat Biotechnol 2017;35(7):672–5.
[139] Heath AP, Bennett GN, Kavraki LE. Finding metabolic pathways using atom tracking. Bioinformatics 2010;26(12):1548–55.