A pathway for every product? Tools to discover and design plant metabolism

James G. Jeffryes, Samuel MD Seaver, José P. Faria, Christopher S. Henry*

Argonne National Laboratory, Mathematics and Computer Science Division, Argonne Illinois, United States

Abstract

The vast diversity of plant natural products is a powerful indication of the biosynthetic capacity of plant metabolism. Synthetic biology seeks to capitalize on this ability by understanding and reconfiguring the biosynthetic pathways that generate this diversity to produce novel products with improved efficiency. Here we review the algorithms and databases that presently support the design and manipulation of metabolic pathways in plants, starting from metabolic models of native biosynthetic pathways, progressing to novel combinations of known reactions, and finally proposing new reactions that may be carried out by existing enzymes. We show how these tools are useful for proposing new pathways as well as identifying side reactions that may affect engineering goals.

Keywords: Cheminformatics, Metabolic modeling, Pathway design, Plant specialized metabolism

1. Introduction

Synthetic biology is a diverse field that seeks to redesign biological systems using a range of engineering principles. To date, much of the synthetic biology efforts in plants have focused either on the introduction of heterologous metabolic pathways into a plant host (such as the beta-carotene synthesis pathway to produce Golden Rice)[1, 2] or the manipulation of existing pathway regulation[3, 4, 5]. A number of plant pathways have also been transferred into microbial hosts to produce complex natural products[6, 7, 8]

*chenry@mcs.anl.gov

such as Artemisinin[9] and Hydrocodone[10]. Microbial hosts are well suited to the production of chemicals because of their rapid development cycle and robust engineering toolkits. However, plants cells have advantages expressing metabolic pathways from other plants because they possess machinery to compartmentalize pathways and to process plant mRNA and proteins properly[11].

There have been a wide variety of metabolic pathways constructed *de novo* that: (i) use enzymes in novel combinations; (ii) take advantage of an enzymes' ability to synthesize multiple reactions; or (iii) engineer completely new enzymes[12, 13]. Enzymes have a reputation for specificity, but in fact, it has been shown that the majority of reactions in metabolism are carried out by promiscuous enzymes that catalyze multiple reactions[14]. Furthermore, these promiscuous enzymes are more likely to appear in specialized metabolism, perhaps because these enzymes may experience less selective pressure to narrow their substrate range[15, 16]. Furthermore, in a larger specialized metabolite, each functional site may be sufficiently separated on the chemical scaffold so as to not dictate the modification of the other sites. This flexibility in the sequence of reactions turns the ideal of a neat linear biosynthesis pathway into a metabolic mesh[17].

Fortunately, computational tools of many types (See Figure 1) exist to help researchers make sense of this complexity in order to design and evaluate metabolic pathways. We explore the tools available to those wishing to explore metabolic pathways in plants in four parts: (i) we first describe the databases of biochemistry, metabolic models and genomic clustering methods that catalog known biochemical pathways in plants; (ii) we then explore graph-based and and constraint-based algorithms which combine these reactions in novel ways to form new biosynthetic pathways; (iii) we discuss tools that have been developed to predict new plausible biochemical reactions based on known enzymatic activities; and (iv) we explore how these tools can be used to propose pathways and potential side reactions.

2. Resources for plant metabolism

2.1. Biochemistry databases

The diversity in metabolic products in plants is reflected through the number, size, and variety of databases that store plant biochemistry data. One type of database aims to act as a general archive for many known metabolites and their associated structures and curated properties. Some

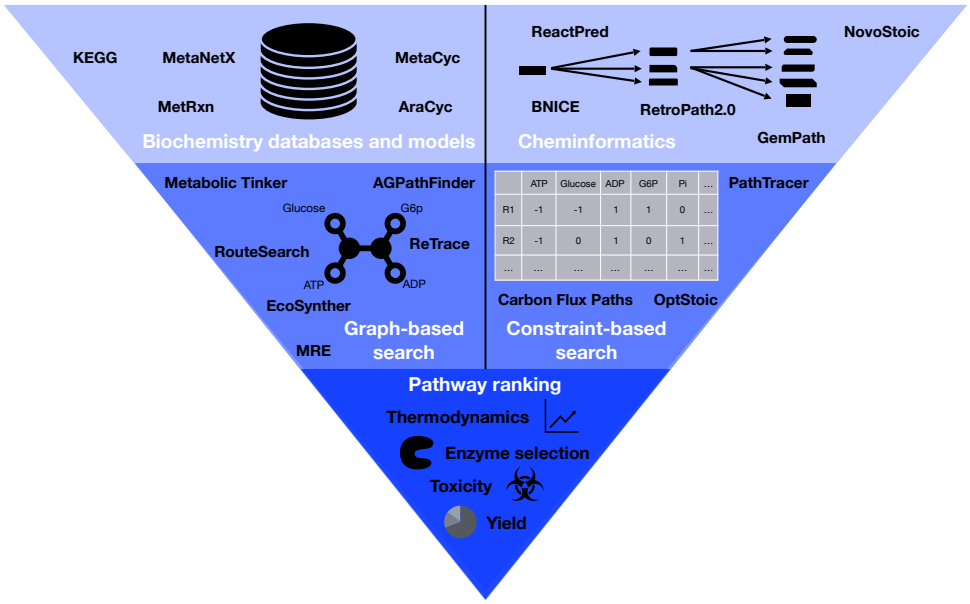

Figure 1: Overview of methods and data used to propose new metabolic pathways. Cheminformatics and Metabolite databases enumerate reaction possibilities. Graph-based and Constraint-based search algorithms assemble viable pathways which can then be evaluated on a number of criteria like thermodynamics and yield.

of these databases also include chemical compounds that do not naturally occur as metabolic intermediates, and as such, we list the total size of the database rather than the number of metabolites. This includes widely used public databases such as PubChem (94M compounds) [18], ChemSpider (62M compounds)[19], ChEMBL (1.7M compounds)[20], and ChEBI (53K compounds)[21]. Another type of database focuses exclusively on those metabolites that are asserted to be present in a specific range of plant species, including AtMetExpress[22] and KNApSAcK[23]. Other databases provide metabolomics data collected in the form of either raw spectral data or spectral libraries, or both, such as the Golm Metabolome Database[24], PlantMetabolomics.org[25], GNPS[26], MetaboLights[27], Weizmass[28] and MassBank[29].

The biochemical resources listed above do not necessarily include information on the pathways and reactions in which the metabolites are involved. The development of such a database involves substantial literature mining and curation of derived biochemical reactions and their as-

sociated enzymes. Many databases of this type are derived from one of two pathway databases: MetaCyc[30] and KEGG[31]. Developed in parallel, these databases have continued to grow over the years. More recently, new databases have arisen that combine and extend the data in KEGG and MetaCyc, such as BKMReact[32], MetanetX[33] and MetRxn[34]. Other recent databases leverage KEGG and MetaCyc, but focus more specifically on plants, including PlantCyc[35], PlantSEED[36], and the Arabidopsis Reactome[37]. Finally, there are multiple smaller databases available for plants which are focused on specific parts of plant metabolism, such as the isoprenoid pathway database[38] and MetaCrop[39]. These pathway databases provide an essential parts-lists of reactions that might comprise a potential designed pathway.

2.2. Metabolic reconstructions

Ultimately, pathway design requires more than a parts-list of metabolites and reactions. Reaction data must be integrated with genomic data to provide insights into how species must be re-engineered to implement or improve a new pathway. Furthermore, support must be provided for simulation to enable prediction of potential yield, rates, titer of desired chemical products, and impact of pathway activity on host cell function. Today, this additional support is typically provided by constraint-based metabolic models [40, 41]. Since the first genome-scale metabolic reconstruction of *Arabidopsis* was published in 2009[42], a number of additional models have been released for *Arabidopsis*[43, 44, 45] and other species such as corn[46, 47] and tomato[48].

The approaches used to build and refine plant metabolic models depend on a variety of data inputs, including data on protein localization, biomass composition, and biochemistry from MetaCyc[30] or KEGG[31]. This biochemistry source data is a key limitation in building plant metabolic models, because much of the research and curation on plant metabolism performed to date has been focused on primary metabolism, for which the pathways in both MetaCyc and KEGG are well defined. Genome-scale metabolic models are so named because they are meant to encompass as much of the overall metabolism of a cell as possible, which should ideally include all key specialized metabolic pathways and their end-point products. However, due to the incomplete and sporadic information available for most specialized metabolic pathways in plants, extensive manual curation is required to identify pathway steps and intermediates that have been characterized in the literature

and reconstruct individual specialized pathways for a plant species.

The prediction of meaningful flux profiles from plant metabolic models poses additional challenges. The mapping of pathways to different organelles requires the addition of specific transport reactions in order to make these pathways active, a problem which can be solved with manual curation. Furthermore, many specialized metabolic reactions specified in existing pathway databases (described previously) are often not mass-balanced, or contain generic compounds that prevent accurate simulations within models. Some of these problems may be partially resolved by using an automated heuristic which will find and suggest alternative reactions that would improve model simulations, such as GapFilling [49]. However, when specialized metabolic pathways are added to a plant metabolic model, the simulation of these pathways requires additional work. We encountered this challenge when building a genome-scale metabolic model of *Arabidopsis* using PlantSEED and AraCyc[36, 50, 51]. The draft reconstruction contained 1,978 unique reactions, and 1,748 unique metabolites. However, 474 (27%) of the metabolites in the model were dead-ends, incapable of being either produced or consumed by the model, despite extensive use of gapfilling during reconstruction. This illustrates the extent to which information gaps still exist in our pathway databases, particularly for plants.

2.3. Genomic Clustering of Specialized Enzymes

This review primarily focuses on the means by which biochemical networks can be extended using reasoning from chemical motifs, but this work can be complemented by predictions produced from genomic evidence. The general mode of catalytic action of a protein-encoding gene can often be predicted by detecting patterns of conserved domains present in the protein sequence. For example, Phytozome[52] assigns and links Pfam[53] domains while Uniprot[54] does so similarly with domains from Prosite[55]. However, as previously described, specialized enzymes often can be promiscuous and utilize a range of similar substrates. As such, additional information is needed to link enzymes within the same pathway.

Work has been done to link and cluster the enzymes using two separate and complementary approaches. First, researchers have utilized the notion of physical clustering, popularized by bacterial operons[56]. Genes that are found close together along a chromosome often form a functional cluster wherein each member catalyzes a step in the same pathway[57, 58, 59]. The second approach involves temporal clustering, or the clustering of genes that

are expressed within similar time-frames. These gene may be transcriptionally co-regulated and often also catalyze steps in the same pathway[60, 61]. The result of these methods is a set of generally-annotated enzymes such as methyltransferases, dioxygenases, and cytochrome P450s from which the researcher must assemble the specific reaction steps. One example of the utility of this approach is the characterization of unknown enzymes in the podophyllotoxin pathway from Mayapple by the Sattely laboratory. Identifying the missing enzymatic steps allowed the entire pathway to be expressed in tobacco leaves, generating a new semi-synthetic option for the production of a chemotherapeutic etoposide[62]. These genomic approaches continue to uncover new biosynthetic gene clusters that may be utilized by the pathway generation algorithms described in the reminder of this review.

3. Algorithms for predicting new pathways from known reactions

Many algorithms have been developed that use the data from metabolic models and biochemistry databases to propose new metabolic pathways [63, 64, 65]. These pathways are often alternatives to native pathways that offer advantages for synthetic biology such as greater carbon and cofactor efficiency or avoiding certain kinds for feedback regulation. The methods for pathway design that we explore in this review fall into two major categories which are described in further detail below: (i) methods built upon constraint-based modeling; and (ii) methods that employ graph-based algorithms for path finding. We also note that, in this review, we are focusing primarily on recently published work and unique approaches to the pathway prediction problem.

3.1. Constraint-based methods

We start our survey of potential pathway design methods by exploring the methods that apply a constraint-based modeling approach. Generally, all constraint-based approaches use linear programing to search through a solution space to maximize a mathematical goal while meeting specified constraints. The most common example of this type of approach is Flux Balance Analysis, where the goal is to maximize the growth of an organism by changing the flow of metabolites though its reaction network under a variety of environmental and genetic constraints [66]. Carbon Flux Paths (CFPs) [67], optStoic [68], and PathTracer [69] similarly represent metabolism as a series of reaction equations in matrix form[70]. The goal of all of these methods

is to find the right combination of reactions from native and heterologous metabolism that allow for the production of a desired compound from specified starting molecules. These techniques allow researchers to prioritize the shortest and/or most efficient pathway (e.g. highest carbon yield) between two metabolites as well as add additional constraints like the thermodynamic feasibility which is discussed more fully below[71]. This approach to pathway design is robust and flexible, permitting the design of lengthy pathways that involve complex branching, produce numerous potential byproducts, and involve converging sub-paths from multiple intermediates or starting molecules. The disadvantage of these approaches is the computational complexity associated with solving the linear and mixed-integer optimization problems underlying these formulations. While excellent solvers are available[72, 73], these pathway searches can be time consuming, often requiring hours to complete depending on the number of candidate reactions and length of the pathway being designed. This is further complicated by the fact that typically multiple alternative pathways designs are desired.

3.2. Graph-based methods

An alternative approach to pathway design involves the use of graph-based methods. These methods [74, 75] represent metabolism as a network with nodes and edges representing the connections between reactions and metabolites. Abstracting the metabolic network in this form provides a well-defined and scalable structure for pathway prediction due to the multitude of algorithms and methods available to search and analyze graphs[76]. The search algorithms for graph-based methods generally allow them to scale more easily to larger metabolic networks and longer pathways at a smaller computational cost. However, this representation of the metabolic network disregards co-reactants and byproducts such as ATP or NADH. Neglecting reaction stoichiometry in metabolic pathway finding can lead to the prediction of biologically irrelevant pathways [77], so these algorithms must often introduce heuristics to further guide pathway search [63]. ReTrace [78], Route Search [79] and AGPathFinder [80] track the atoms in metabolites from reaction to reaction, while other methods such as MRSD [81], Metabolic Tinker [82] and MRE [83] apply penalties to cofactors to prevent pathways from "short-circuiting" by using a cofactor to link two disparate pathways. Several of these methods are available as web-services and are detailed in Table 1.

3.3. Pathway Selection

In practice, coming up with potential candidate pathways is not the most significant challenge associated with pathway design. The plethora of pathway-finding algorithms will easily propose thousands, or even ten of thousands of candidate pathways, particularly when the addition of heterologous reactions steps is considered. Thus, it is essential to have robust techniques for whittling down candidate pathways to a small number of top-choices to guide the metabolic engineering process.

The first criteria is the number of heterologous steps included in the pathway for a given host. Generally, pathways involving large numbers of heterologous steps will be difficult to engineer, as each new enzyme that must be added to an organism incurs a metabolic cost to express th protein as well as a potential for off-target activity. These considerations make shorter pathways more efficient and easier to control.

A second consideration in pathway selection is the thermodynamic feasibility of each reaction step within the pathway. Reaction steps with positive changes in standard Gibbs free energy are "uphill" and will require a favorable concentration gradient in order to proceed. If a pathway includes too many unfavorable reaction steps, it can quickly become intractable. A variety of methods[84, 85] can be used to predict and avoid these thermodynamic bottlenecks.

The efficient use of carbon and energy is also an important consideration for commercially relevant pathways. Pathways must have high yields in order to be economically viable, particularly when the end-products are lower-cost commodity chemicals like biofuels. In a plant context, pathways with high carbon efficiency are less important when the plant generates its own sugars. Additionally, precisely balancing reactions that produce and consume redox carriers like NADH and FADH is critical for anaerobic fermentation with microbes but not relevant for plants which can convert reduction potential to energy by aerobic ATP synthesis. Finally, pathways containing intermediates which are toxic could impact cell growth and should be avoided if possible[86].

Several recent publications have demonstrated the utility of pathway design and selection tools by proposing metabolic routes to a number of compounds of interest. For example, Carbon Flux Paths (CFPs)[67] were able to correctly predict the known long pathway to convert bicarbonate to cytidine-diphosphate in *E. coli*, instead of a biologically irrelevant short pathway via ADP. Additionally, while exploring all CFPs between pyruvate and oxaloacate the authors demonstrate that the method can accurately predict condi-

tions when the glyoxylate shunt of the TCA cycle will be active. OptStoic[68] proposed a highly efficient pathway for glucose conversion to acetate and identified methods to overcome the thermodynamic unfavorable conversion of methane to acetate. PathTracer[69] was applied to the genome-scale model of *E. coli* to uncover viable pathways for the conversion of putrescine to glutamate and to identify a possible CO_2 fixation pathway. ReTrace analysis was conducted to propose biosynthesis of inosine 5'-monophosphate (IMP) from glucose in *E. coli*. More interestingly, the authors also used ReTrace analysis to predict amino acid synthesis in the filamentous fungus *Trichoderma reesei*. Paired with manual curation of the predicted pathways, this example highlights how pathway prediction tools can be a powerful asset to analyze recently sequenced or less studied organisms.

The creators of Metabolic Tinker examined the mevalonate pathway, which is the pathway used by eukaryotes to make 3-isopentenyl pyrophosphate (IPP), the monomer precursor of all terpenoid natural products[87]. They propose multiple alternatives to the natural pathway, several of which have favorable thermodynamics. RouteSearch showed success in predicting multiple pathways described in the literature including production of the flavonoid umbelliferone from L-tyrosine[88] and production of ethylene glycol from Aldehydo-D-Xylose[89]. Similarly, MRE was used to propose pathways for the biosynthesis of the plant metabolite and antioxidant naringenin starting from L-tyrosine[90].

While the preceding examples demonstrate the breadth and potential of these tools, a direct comparison of their effectiveness is difficult. Not all methods have a software implementation publicly available and the community has not settled on a set of challenges and standards for evaluation. While constraint-based methods, like OptStoic and PathTracer, excelled at finding pathways with high carbon efficiency and favorable redox balance, these factors are less important for an pathway in an aerobic, autotrophic context and therefore these tool are less useful in examining plants. On the other hand, several tools like MRE and RouteSearch allow researchers to explore plant metabolic models.

4. Algorithms for proposing potential novel reactions

The pathway-finding algorithms described above do have great utility but are all subject to a fundamental limitation. These algorithms can only find pathways comprised of previously known reactions that are provided as

inputs to these algorithms. Yet, it is well-known that enzymes are capable of performing chemistry that has not yet been captured in existing biochemistry databases. This includes: (i) new reactions that are catalyzed by enzymes that have not yet been characterized or annotated; (ii) promiscuous activities of existing enzymes; and (iii) uncharacterized spontaneous chemistry. All of this new potential chemistry could enable the design of new pathways that are presently not possible using only known reactions. These new pathways can also include routes for the production of non-natural chemical products. Thus, algorithms with the potential to explore the chemical space to propose novel potential reactions are critical to support the pathway design process.

4.1. Statistical models

Some of the first techniques for novel reaction prediction come from toxicology and the pharmaceutical industry, which predict the metabolism of drug compounds within human cells[91, 92]. Given the narrow range of enzyme classes of interest (subtypes of cytochrome p450, for example), individual statistical models were developed for each enzyme with a focus on the site of metabolism[93, 94]. However, this approach was also applied broadly to multiple classes of enzyme chemistry[95] and a similar approach was used to propose new reactions between a known set of compounds.[96]. These algorithms are subject to the limitations of their training data; not many enzyme classes have been tested on a broad range of chemical substrates. These enzyme classes with extensive training data are biased towards those that are easier to purify and catalyze reactions of commercial or medical interest. Furthermore, the substrates on which these enzymes are tested are constrained by the price, stability, and availability of the test compounds[97]. While these approaches can identify the likelihood of a predicted reaction, they are not applied iteratively to build up a novel reaction pathway. Despite these limitations, these approaches remain powerful for predicting the likelihood of a particular reaction-enzyme pair and may prove useful in the selection of a specific enzyme for a reaction step[98].

4.2. Reaction rule sets

An alternative approach is the use of rule-based methods. Unlike the statistical methods, rule-based methods focus on describing only the local neighborhood of the chemical bonds which are breaking and forming in a chemical reaction. As figure 2A demonstrates, reactions are proposed by detecting the same substructure in new compounds and breaking or forming the

bonds within this substructure as specified by the reaction rules. While these rules do not act upon the atoms neighboring the reactive site of a molecule, they often include constraints that require an active site to exist within a specified broader chemical context (e.g. when carbonyl groups are needed to activate a reaction site). Rules with more comprehensive constraints on the context of an active reaction site result in more specific predictions, while rules with little or no context will produce a greater number of novel reaction predictions.

The flexibility of these parameters allows a limited set of rules to replicate a large fraction of known metabolism as well as predict new putative reactions. Table 2 contains a collection of tools using rule sets to predict novel reactions. Recently, new techniques have been developed for algorithmically generating reaction rules from sets of example reactions such as ReactPred[99], Retropath2.0[100] and NovoStoich[101]. These methods are able to rapidly expand the range of reaction types predicted by a rule set and easily incorporate new evidence by removing the bottleneck of human curation. While many of these tools are not publicly available, several offer web interfaces including PathPred[102] and XTMS[86]. Other algorithms are available as open-source applications, including ReactPred[99] and Retropath2.0[100].

As this review is meant to focus specifically on the design of novel plant pathways, it is important to note that plant metabolism, particularly specialized metabolism, poses a number of significant challenges for these novel reaction prediction algorithms. The most basic of these challenges is that the average metabolic intermediate in a specialized pathway may be 2 to 5 times larger than intermediates in energy metabolism. The number of putative reactions predicted by these algorithms depends on the number of reaction sites in each substrate, which typically scales with the size of the starting molecule. This difference may be modest initially but it is compounded every time the rules are applied to predict a new step. For example, applying the enzymatic reaction rules from the Pickaxe tool (see Table 2) to pyruvate with three iterations results in over 5600 reactions. Application of the same rules to rosmarinate with three iterations results in nearly 300,000 reactions. Constraining the branching of these networks by reaction rule specificity[103], similarity of predicted compounds to a target metabolite[104], presence in biochemical databases[105], or reaction thermodynamics[86, 106] helps to address this problem, but the combinatorial explosion of reactive possibilities still constrains the maximum number of novel steps that may be integrated

into a pathway design.

Additionally, these algorithms struggle to generalize the complicated reactions that dictate formation of several specialized plant metabolites. One such example is Squalene-hopene cyclase (Figure 2B), which catalyzes the formation of seven new bonds across the length of the molecule. Reaction rules that are specific enough to capture this reaction accurately will also be too specific to represent the documented ability of this enzyme to catalyze a variety of alternative reactions[107]. This challenge may be overcome by integration of a diverse range of plant metabolites as the starting points for pathway prediction, but most freely-available tools are currently focused on *E. coli* metabolism. Finally, many algorithms do not discriminate between chiral forms of a metabolite, either due to intrinsic limitations in the way molecules are computationally represented in the methods or due to chiral inconsistencies in reference reactions used to generate reaction rules. This limitation may lead to the prediction of biologically infeasible pathways as enzymes may have preferences for one chiral form of a molecule or the predicted starting point (a D-amino acid for example) may not exist in an organism's native metabolism.

5. Applications of novel reaction predictions

5.1. Pathway predictions

There are numerous examples in the literature where tools for automated reaction prediction and novel pathway design were merged to design new pathways. These include the prediction of pathways to produce small molecule commodity chemicals (chemical building blocks that are produced and consumed in large volumes by the chemical industry) such as 3-hydroxy propionate[108], methyl ethyl ketone precursors[109], terephthalic acid[100] and various other short chain alcohols and acids[106]. These examples also include more challenging targets (see Figure 2C) such as rosmarinate, a natural phenol antioxidant[103], and drug molecules such as pyrazinamide, tenofovir[110] phenylephrine, and naproxin[101].

These publications apply many of the previously described metrics to evaluate each of the candidate pathways they propose. However, when pathways integrate novel reaction steps proposed by cheminformatics algorithms, the evaluation of these pathways must also include an analysis of the likelihood that each proposed novel reaction can be catalyzed by an existing

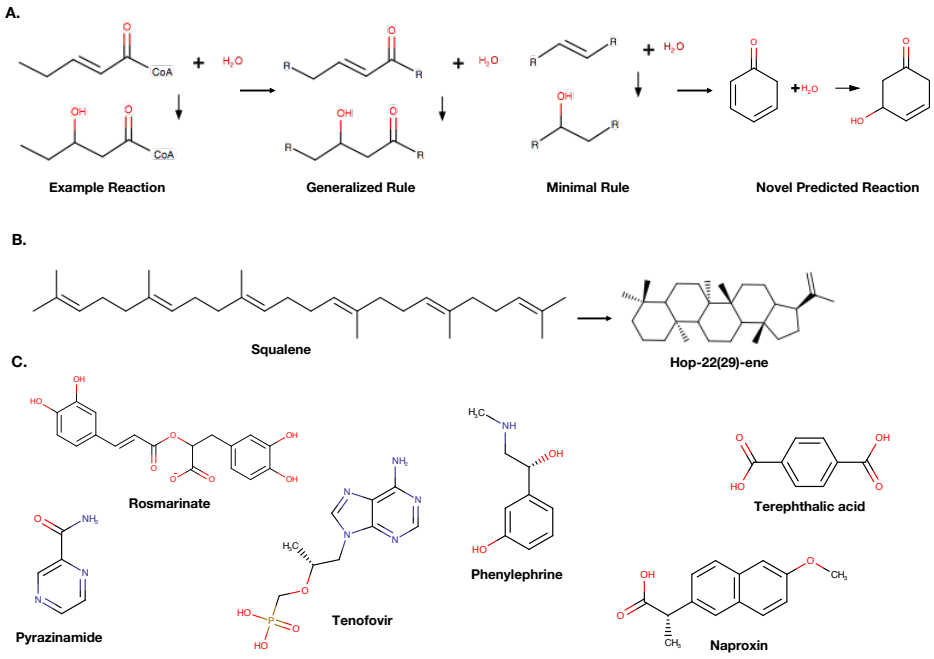

Figure 2: Rule based Reaction prediction: A. Process of generalizing a reaction rule from an example reaction and applying to a novel substrate. B. Squalene-hopene cyclase is an example of a complicated, concerted reaction which is difficult to predict. C. A sample of the various products that have had novel reaction pathways proposed with rule-based methods.

enzyme. Existing cheminformatics approaches propose a variety of mechanisms and metrics for computing this likelihood. Some tools consider the specificity of the reaction rule that created the novel reaction(s) in question, reasoning that more specific reaction rules are more likely to produce reliable predictions. For example Retropath2.0[100] gives precedence to larger(more specific) reaction rules while others like enviPath[111] have manually-assigned likelihood scores. Others tools, like GemPath[106] and the method of Cho et al.[112] compare the chemical similarity of the predicted substrates to the known substrates of an enzyme class to determine the likelihood of a proposed reaction.

Each of these methods has strengths and weaknesses. Manual evaluation, for example, can be accurate but cannot be applied to the thousands of rules

generated by modern methods. The favoring of more specific reaction rules is a criteria that can be applied uniformly but ignores the variability in substrate specificity between a generalist enzyme and a selective specialist enzyme. Evaluating likelihood based on known substrates from published literature is subject to the same biases discussed in the statistical models section, namely, that an enzyme may appear to be highly specific simply because it is poorly studied and has not been tested on a range of substrates.

Once a pathway containing a novel reaction step is selected, tools such as Selenzyme[113], Uniprot[54], and BRENDA[114] can be used to find enzymes which may be able to carry out the novel reaction. The first case where a computationally-designed pathway was actually engineered into a host involved the engineering of *E. coli* to produce the commonly used solvent 1,4-butanediol[56]. After screening through the pathway alternatives and engineering the pathway enzymes, the researchers were ultimately able to achieve a commercially-viable concentration of 110 g/L of product[115]. Reaction prediction tools have also been developed and utilized by companies such as Arzeda[116] and Genomatica[117]. Novel pathway design for plant natural products has received less attention so far because the higher sale price of specialized metabolites makes optimizing the synthesis pathway less critical for commercial success. These methods can still add value however, either by improving the synthesis efficiency of a precursor to the specialized pathway (such as the previously discussed example with IPP and terpenoids) or extending the synthesis pathway to form a derivate of the natural product with improved properties.

5.2. Off-target enzyme activity

Novel pathway design is not the sole practical application for novel reaction prediction tools. These tools can also predict potential side-activities for an enzyme. This is useful, because broad substrate specificity can be a double-edged sword when attempting to apply synthetic biology to construct a new pathway[118]. In particular, these enzymes may interact with the host's metabolism in unpredictable ways when they are overexpressed or heterologously introduced, potentially wasting the carbon supplied or producing toxic intermediates[119]. One example of novel functionality of host enzymes impacting an engineered pathway was uncovered in a yeast strain engineered to overproduce terpenoids. Two glycerol 3-phosphate phosphatases were diverting flux from the intended pathway by cleaving acetyl-phosphate to acetate. Knocking out these enzymes enabled the doubling of cell growth

and production of the desired product[120]. In another case, the transfer of a biosynthetic gene cluster for platencin (a microbial inhibitor) from one host to another resulted in the unintended production of novel "shunt metabolites" (metabolic dead-ends) branching off the biosynthetic pathway[121]. In a final example, a synthetic carbon fixation cycle was initially limited by the formation of malyl-CoA from glyoxalate and acetyl-CoA by a side reaction of one of it's novel cycle enzymes[122].

The side-reactions caused by broad-substrate specificity, like in the examples above, are seldom found on traditional maps of biochemistry. Enzyme specificity databases like BRENDA[114] present known side activities of enzymes extracted from the literature, but such data sources are inevitably incomplete and retrospective. Recently, this problem was addressed more systematically by databases that apply generalized reaction rules more broadly to all known metabolites and collect the novel predicted compounds and reactions. Resources like the MINE database[123] and the ATLAS of Biochemistry[124] organize potential off-target enzyme activities by compound, allowing researchers to formulate hypotheses about potential sources of unwanted byproducts. The generation of putative products of these reactions is particularly helpful because the predominate way of detecting metabolites, Liquid-Chromatography Mass-Spectrometry, requires a list of potential candidate structures to evaluate[125]. The MINE was recently used in this manner to annotate a set of six novel metabolites from a diverse set of biological contexts including green algae *Chlamydomonas reinhardtii* and herb *Artemisia douglasiana*[126].

5.3. Spontaneous chemical damage

The implementation of novel metabolic pathways can be further hampered by the intrinsic reactivity of the metabolites themselves. In particular, the addition of heterologous pathways to a cell can introduce a set of chemical intermediates at elevated concentrations which may react spontaneously with themselves or other proteins and small molecules[127]. For example, in the previously mentioned production pathway for 1,4-butanediol, the spontaneous cyclization of the intermediate 4-hydroxybutyryl-CoA limited productivity until a repair lactonase was introduced[56]. It can be challenging to identify metabolites which will be susceptible to spontaneous side reactions *a priori* because the regulation of native metabolism has an evolutionary pressure to ensure that these side reactions are symptom-less, either by keeping the metabolite concentrations low or through additional enzymes that repair

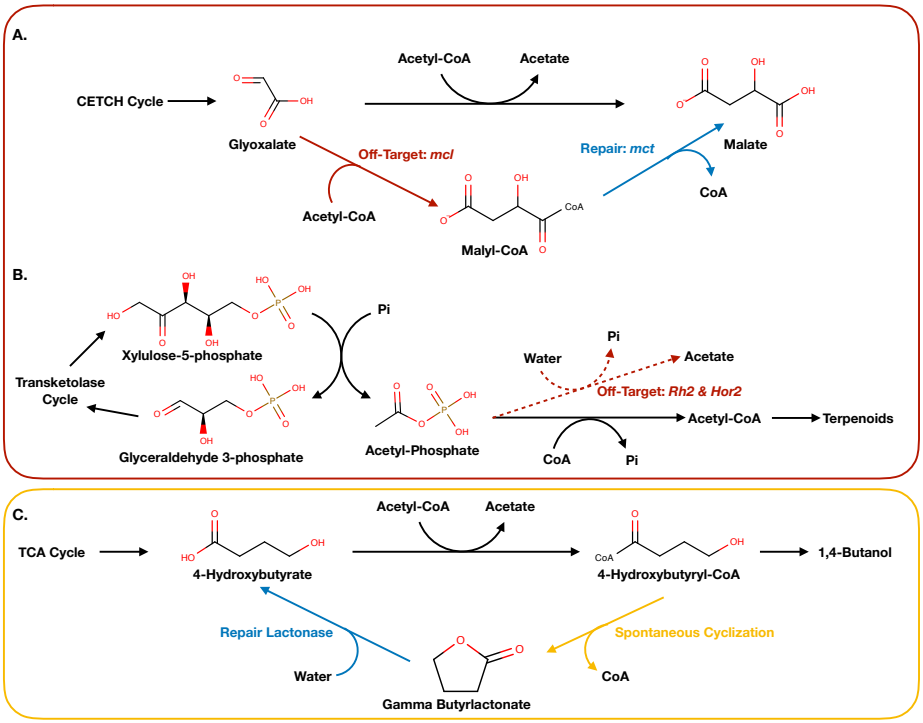

Figure 3: Impact of off-target reactions on metabolic pathways: A. Addition of a repair enzyme compensates for the undesired accumulation of malyl-CoA. B. Undesired side activity of two glycerol 3-phosphate phosphatases was abolished by knocking out the enzymes C. The spontaneous cyclization of 4-hydroxybutyryl-CoA is mitigated by adding a lactonase.

the side activity[128, 129]. Additionally, these labile metabolites may be degraded in the extraction process, complicating efforts to observe them when diagnosing problems with a synthetic pathway[130].

The Chemical Damage Metabolic *In silico* Network Expansion (or CD-MINE) [131] was developed to aid in the identification of damage-prone metabolites. Specifically, new spontaneous reaction rules were created by generalizing known spontaneous chemical transformations in biological systems. These rules were derived from known spontaneous reactions compiled from KEGG, MetaCyc, and numerous literature sources. The rules were then applied to a database of known metabolites to generate predicted spontaneous reactions and products. It should be noted that the mere presence of a pathway intermediate as a reactant in a spontaneous reaction predicted in

the CD-MINE database is not a guarantee that the reaction will occur at a materially-significant rate. However, this database collects hypotheses of side reactions that may become significant if the concentrations of the metabolite are elevated in the engineered system. Once these spontaneous side reactions are uncovered, they can be addressed through a number of mechanisms such as scaffolding pathway enzymes[132, 133], sequestering the reactive pathway in cellular compartments[134, 135], proper balancing of enzyme expression levels[136] and expression of damage-control enzymes[127].

6. Conclusion

Here we describe much of the data and many of the tools needed to support novel pathway design in plants (and many other host types). In particular, we discuss the strengths and weaknesses of the various tools and data-sources as well as provide examples of additional applications of these tools beyond proposing new pathways. As is often the case, the best approach will vary based on the complexity of the task at hand. Existing metabolic models of plants can be powerful in supporting the pathway design process, but these models remain incomplete. Novel reaction prediction algorithms can propose reactions to fill many pathway gaps, but they still often fail to predict many of the complex and lengthy synthesis pathways for specialized metabolism in plants. Additionally, comparing these approaches can be challenging because no standard set of metrics to evaluate tools exists. Ideally, we should introduce metrics for evaluation of the type used in DREAM challenges for systems biology[137]

In spite of these difficulties, several promising tools have recently emerged, making this field more relevant to plant science than ever before. New databases of predicted metabolism like ATLAS[124] and MINE [123] allow researchers to hypothesize the off-target effects of overexpressing or introducing heterologous enzymes on cellular metabolism. Techniques that integrate novel reactions and constraint-based solutions such as NovoStoic[101] are opening new ways of searching the space of potential metabolism. Finally, open-source reaction prediction software like Retropath2.0[100] now allows researchers to apply these techniques to new products of interest. While the currently-released rules are based on *E. coli*, an open code base allows plant scientists to extend these rules with new reactions and metabolites unique to their organism of interest.

Just like the first constraint-based models were focused on core metabolism and microbes, these tools so far have been primarily focused on producing simpler products in fermentation. However, with new methods for generating and curating reactions, the groundwork is being laid for extending this to more complex plant metabolism. These approaches for novel reaction generation complement existing efforts to determine biosynthesis pathways for plant natural products from genomic evidence that currently rely on manual inspection of chemical scaffolds. These computational approaches will not replace human intuition but rather spur the researcher's creativity to consider new reaction possibilities.

Funding: This work was supported by the National Science Foundation grants MCB 1611952 (JGJ, JPF CSH) and IOS 1444202 (SMDS) and the U.S. Department of Energy, Office of Biological and Environmental Research; under contract DE-AC02-06CH11357 (CSH)

[1] P. Beyer, S. Al-Babili, X. Ye, P. Lucca, P. Schaub, R. Welsch, I. Potrykus, Golden Rice: introducing the beta-carotene biosynthesis pathway into rice endosperm by genetic engineering to defeat vitamin A deficiency., The Journal of nutrition 132 (2002) 506S–510S.

[2] C. Martin, J. Li, Medicine is not health care, food is health care: plant metabolic engineering, diet and human health, New Phytologist 216 (2017) 699–719.

[3] T. Capell, P. Christou, Progress in plant metabolic engineering, Current Opinion in Biotechnology 15 (2004) 148–154.

[4] W. Lau, M. A. Fischbach, A. Osbourn, E. S. Sattely, Key Applications of Plant Metabolic Engineering, PLoS Biology 12 (2014) e1001879.

[5] E. C. Tatsis, S. E. OConnor, New developments in engineering plant metabolic pathways, Current Opinion in Biotechnology 42 (2016) 126–132.

[6] M. Ibdah, S. Martens, D. R. Gang, Biosynthetic Pathway and Metabolic Engineering of Plant Dihydrochalcones, Journal of Agricultural and Food Chemistry (2017) acs.jafc.7b04445.

[7] L. Narcross, E. Fossati, L. Bourgeois, J. E. Dueber, V. J. Martin, Microbial Factories for the Production of Benzylisoquinoline Alkaloids, Trends in Biotechnology 34 (2016) 228–241.

[8] M. Furubayashi, M. Ikezumi, S. Takaichi, T. Maoka, H. Hemmi, T. Ogawa, K. Saito, A. V. Tobias, D. Umeno, A highly selective biosynthetic pathway to non-natural C50 carotenoids assembled from moderately selective enzymes, Nature Communications 6 (2015) 7534.

[9] D.-K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. a. Ho, R. a. Eachus, T. S. Ham, J. Kirby, M. C. Y. Chang, S. T. Withers, Y. Shiba, R. Sarpong, J. D. Keasling, Production of the antimalarial drug precursor artemisinic acid in engineered yeast., Nature 440 (2006) 940–3.

[10] S. Galanie, K. Thodey, I. J. Trenchard, M. Filsinger Interrante, C. D. Smolke, Complete biosynthesis of opioids in yeast, Science 349 (2015) 1095–1100.

[11] C. Owen, N. J. Patron, A. Huang, A. Osbourn, Harnessing plant metabolic diversity, Current Opinion in Chemical Biology 40 (2017) 24–30.

[12] J. Sun, H. S. Alper, Metabolic engineering of strains: from industrial-scale to lab-scale chemical production, Journal of Industrial Microbiology & Biotechnology 42 (2015) 423–436.

[13] T. J. Erb, P. R. Jones, A. Bar-Even, B. Hauer, S. Lutz, Synthetic metabolism: metabolic engineering meets enzyme design, Current Opinion in Chemical Biology 37 (2017) 56–62.

[14] H. Nam, N. E. Lewis, J. a. Lerman, D.-H. Lee, R. L. Chang, D. Kim, B. O. Palsson, Network context and selection in the evolution to enzyme specificity., Science (New York, N.Y.) 337 (2012) 1101–4.

[15] A. Babtie, N. Tokuriki, F. Hollfelder, What makes an enzyme promiscuous?, Current opinion in chemical biology 14 (2010) 200–7.

[16] A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, R. Milo, The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters, Biochemistry 50 (2011) 4402–4410.

[17] J.-K. Weng, R. N. Philippe, J. P. Noel, The rise of chemodiversity in plants., Science (New York, N.Y.) 336 (2012) 1667–70.

[18] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, PubChem Substance and Compound databases., Nucleic acids research 44 (2016) 1202–13.

[19] A. Williams, V. Tkachenko, The Royal Society of Chemistry and the delivery of chemistry data repositories for the community, Journal of Computer-Aided Molecular Design 28 (2014) 1023–1030.

[20] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Research 40 (2012) D1100–D1107.

[21] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013, Nucleic Acids Research 41 (2013) D456–D463.

[22] F. Matsuda, M. Y. Hirai, E. Sasaki, K. Akiyama, K. Yonekura-Sakakibara, N. J. Provart, T. Sakurai, Y. Shimada, K. Saito, AtMetExpress Development: A Phytochemical Atlas of Arabidopsis Development, PLANT PHYSIOLOGY 152 (2010) 566–578.

[23] F. M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, L. K. Darusman, K. Saito, S. Kanaya, KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research., Plant & cell physiology 53 (2012) e1.

[24] J. Hummel, J. Selbig, D. Walther, J. Kopka, The Golm Metabolome Database: adatabase for GC-MS based metabolite profiling, Springer, Berlin, Heidelberg, 2007, pp. 75–95.

[25] P. Bais, S. M. Moon, K. He, R. Leitao, K. Dreher, T. Walk, Y. Sucaet, L. Barkan, G. Wohlgemuth, M. R. Roth, E. S. Wurtele, P. Dixon, O. Fiehn, B. M. Lange, V. Shulaev, L. W. Sumner, R. Welti, B. J. Nikolau, S. Y. Rhee, J. A. Dickerson, PlantMetabolomics.org: a web portal for plant metabolomics experiments., Plant physiology 152 (2010) 1807–16.

[26] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. . Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, N. Bandeira, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, Nature Biotechnology 34 (2016) 828–837.

[27] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, C. Steinbeck, MetaboLightsan open-access general-purpose repository for metabolomics studies and associated meta-data, Nucleic Acids Research 41 (2013) D781–D786.

[28] N. Shahaf, I. Rogachev, U. Heinig, S. Meir, S. Malitsky, M. Battat, H. Wyner, S. Zheng, R. Wehrens, A. Aharoni, The WEIZMASS spec-

tral library for high-confidence metabolite identification, Nature Communications 7 (2016) 12423.

[29] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences., Journal of mass spectrometry : JMS 45 (2010) 703–14.

[30] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, Nucleic Acids Research 44 (2016) D471–D480.

[31] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: back to metabolism in KEGG., Nucleic acids research 42 (2014) 199–205.

[32] M. Lang, M. Stelzer, D. Schomburg, BKM-react, an integrated biochemical reaction database., BMC biochemistry 12 (2011) 42.

[33] M. Ganter, T. Bernard, S. Moretti, J. Stelling, M. Pagni, MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks, Bioinformatics 29 (2013) 815–816.

[34] A. Kumar, P. F. Suthers, C. D. Maranas, MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases., BMC bioinformatics 13 (2012) 6.

[35] K. Dreher, Putting the Plant Metabolic Network Pathway Databases to Work: Going Offline to Gain New Capabilities, Humana Press, Totowa, NJ, 2014, pp. 151–171.

[36] S. M. D. Seaver, S. Gerdes, O. Frelin, C. Lerma-Ortiz, L. M. T. Bradbury, R. Zallot, G. Hasnain, T. D. Niehaus, B. El Yacoubi, S. Pasternak, R. Olson, G. Pusch, R. Overbeek, R. Stevens, V. de Crécy-Lagard,

D. Ware, A. D. Hanson, C. S. Henry, High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource., Proceedings of the National Academy of Sciences of the United States of America 111 (2014) 9645–50.

[37] N. Tsesmetzis, M. Couchman, J. Higgins, A. Smith, J. H. Doonan, G. J. Seifert, E. E. Schmidt, I. Vastrik, E. Birney, G. Wu, P. D'Eustachio, L. D. Stein, R. J. Morris, M. W. Bevan, S. V. Walsh, Arabidopsis reactome: a foundation knowledgebase for plant systems biology., The Plant cell 20 (2008) 1426–36.

[38] E. Vranová, M. Hirsch-Hoffmann, W. Gruissem, AtIPD: a curated database of Arabidopsis isoprenoid pathway models and genes for isoprenoid network analysis., Plant physiology 156 (2011) 1655–60.

[39] F. Schreiber, C. Colmsee, T. Czauderna, E. Grafahrend-Belau, A. Hartmann, A. Junker, B. H. Junker, M. Klapperstuck, U. Scholz, S. Weise, MetaCrop 2.0: managing and exploring information about crop plant metabolism, Nucleic Acids Research 40 (2012) D1173–D1177.

[40] S. M. D. Seaver, C. S. Henry, A. D. Hanson, Frontiers in metabolic reconstruction and modeling of plant genomes, Journal of Experimental Botany 63 (2012) 2247–2258.

[41] E. OBrien, J. Monk, B. Palsson, Using Genome-scale Models to Predict Biological Capabilities, Cell 161 (2015) 971–987.

[42] M. G. Poolman, L. Miguet, L. J. Sweetlove, D. A. Fell, A genome-scale metabolic model of Arabidopsis and some of its properties., Plant physiology 151 (2009) 1570–81.

[43] C. G. de Oliveira Dal'Molin, L.-E. Quek, R. W. Palfreyman, S. M. Brumbley, L. K. Nielsen, AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis., Plant physiology 152 (2010) 579–89.

[44] S. Mintz-Oron, S. Meir, S. Malitsky, E. Ruppin, A. Aharoni, T. Shlomi, Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity., Proceedings of the National Academy of Sciences of the United States of America 109 (2012) 339–44.

[45] A. Arnold, Z. Nikoloski, Bottom-up Metabolic Reconstruction of Arabidopsis and Its Application to Determining the Metabolic Costs of Enzyme Production., Plant physiology 165 (2014) 1380–1391.

[46] C. G. d. O. Dal'Molin, L.-E. Quek, R. W. Palfreyman, S. M. Brumbley, L. K. Nielsen, C4GEM, a genome-scale metabolic model to study C4 plant metabolism., Plant physiology 154 (2010) 1871–85.

[47] M. Simons, R. Saha, N. Amiour, A. Kumar, L. Guillard, G. Clément, M. Miquel, Z. Li, G. Mouille, P. J. Lea, B. Hirel, C. D. Maranas, Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model., Plant physiology 166 (2014) 1659–74.

[48] H. Yuan, C. M. Cheung, M. G. Poolman, P. A. J. Hilbers, N. A. W. van Riel, A genome-scale metabolic network reconstruction of tomato (¡i¿Solanum lycopersicum¡/i¿ L.) and its application to photorespiratory metabolism, The Plant Journal 85 (2016) 289–304.

[49] V. Satish Kumar, M. S. Dasika, C. D. Maranas, Optimization based automated curation of metabolic reconstructions., BMC bioinformatics 8 (2007) 212.

[50] P. Schläpfer, P. Zhang, C. Wang, T. Kim, M. Banf, L. Chae, K. Dreher, A. K. Chavali, R. Nilo-Poyanco, T. Bernard, D. Kahn, S. Y. Rhee, Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants., Plant physiology 173 (2017) 2041–2059.

[51] S. M. D. Seaver, L. M. T. Bradbury, O. Frelin, R. Zarecki, E. Ruppin, A. D. Hanson, C. S. Henry, Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm., Frontiers in plant science 6 (2015) 142.

[52] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, D. S. Rokhsar, Phytozome: a comparative platform for green plant genomics, Nucleic Acids Research 40 (2012) D1178–D1186.

[53] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L.

Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Research 42 (2014) D222–D230.

[54] UniProt: the universal protein knowledgebase, Nucleic Acids Research 45 (2017) D158–D169.

[55] C. J. A. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Research 41 (2012) D344–D347.

[56] H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J. D. Trawick, R. E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H. B. Schreyer, S. Andrae, T. H. Yang, S. Y. Lee, M. J. Burk, S. Van Dien, Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol, Nature Chemical Biology 7 (2011) 445–452.

[57] B. Field, A.-S. Fiston-Lavier, A. Kemen, K. Geisler, H. Quesneville, A. E. Osbourn, Formation of plant metabolic gene clusters within dynamic chromosomal regions., Proceedings of the National Academy of Sciences of the United States of America 108 (2011) 16116–21.

[58] N. Töpfer, L.-M. Fuchs, A. Aharoni, The PhytoClust tool for metabolic gene clusters discovery in plant genomes, Nucleic Acids Research 45 (2017) 7049–7063.

[59] S. A. Kautsar, H. G. SuarezDuran, K. Blin, A. Osbourn, M. H. Medema, plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters, Nucleic Acids Research 45 (2017) W55–W63.

[60] Q. Lv, R. Cheng, T. Shi, Regulatory network rewiring for secondary metabolism in Arabidopsis thaliana under various conditions., BMC plant biology 14 (2014) 180.

[61] C. Ruprecht, A. Mendrinna, T. Tohge, A. Sampathkumar, S. Klie, A. R. Fernie, Z. Nikoloski, S. Persson, M. Mutwil, FamNet: A framework to identify multiplied modules driving pathway diversification in plants., Plant Physiology 170 (2016) pp.01281.2015.

[62] W. Lau, E. S. Sattely, Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone, Science 349 (2015) 1224–1228.

[63] S. M. Kim, M. I. Peña, M. Moll, G. N. Bennett, L. E. Kavraki, A review of parameters and heuristics for guiding metabolic pathfinding, 2017.

[64] M. Kotera, S. Goto, Metabolic pathway reconstruction strategies for central metabolism and natural product biosynthesis., Biophysics and physicobiology 13 (2016) 195–205.

[65] M. H. Medema, R. van Raaphorst, E. Takano, R. Breitling, Computational tools for the synthetic design of biochemical pathways., Nature reviews. Microbiology 10 (2012) 191–202.

[66] J. D. Orth, I. Thiele, B. . Palsson, What is flux balance analysis?, Nature biotechnology 28 (2010) 245–8.

[67] J. Pey, J. Prada, J. E. Beasley, F. J. Planes, Path finding methods accounting for stoichiometry in metabolic networks, Genome Biology 12 (2011) R49.

[68] A. Chowdhury, C. D. Maranas, Designing overall stoichiometric conversions and intervening metabolic reactions, Scientific Reports 5 (2015).

[69] C. J. Tervo, J. L. Reed, MapMaker and PathTracer for tracking carbon in genome-scale metabolic models, Biotechnology Journal 11 (2016) 648–661.

[70] O. Resendis-Antonio, Stoichiometric Matrix, Springer New York, New York, NY, p. 2014.

[71] X. Zhang, C. J. Tervo, J. L. Reed, Metabolic Assessment of E. coli as a Biofactory for Commercial Products, Metabolic Engineering (2016) 1–11.

[72] S. J. Maher, T. Fischer, T. Gally, G. Gamrath, A. Gleixner, R. L. Gottwald, G. Hendel, T. Koch, M. E. Lübbecke, M. Miltenberger, B. Müller, M. E. Pfetsch, C. Puchert, D. Rehfeldt, S. Schenker, R. Schwarz, F. Serrano, Y. Shinano, D. Weninger, J. T. Witt, J. Witzig, The SCIP Optimization Suite 4.0, Technical Report, 2017.

[73] G. O. Inc., Gurobi Optimizer reference manual, Www.Gurobi.Com 6 (2014) 572.

[74] T. Aittokallio, B. Schwikowski, Graph-based methods for analysing networks in cell biology, 2006.

[75] Y. Deville, D. Gilbert, J. van Helden, S. Wodak, An overview of data models for the analysis of biochemical pathways, Briefings in Bioinformatics 4 (2003) 246–259.

[76] L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, Search in power-law networks, Physical Review E 64 (2001) 046135.

[77] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, D. Schomburg, Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), Bioinformatics 21 (2005) 1189–1193.

[78] E. Pitkänen, P. Jouhten, J. Rousu, Inferring branching pathways in genome-scale metabolic networks, BMC Systems Biology 3 (2009) 103.

[79] M. Latendresse, M. Krummenacker, P. D. Karp, Optimal metabolic route search based on atom mappings, Bioinformatics 30 (2014) 2043–2050.

[80] Y. Huang, C. Zhong, H. X. Lin, J. Wang, A Method for Finding Metabolic Pathways Using Atomic Group Tracking, PLOS ONE 12 (2017) e0168725.

[81] D. Xia, H. Zheng, Z. Liu, G. Li, J. Li, J. Hong, K. Zhao, MRSD: A web server for Metabolic Route Search and Design, Bioinformatics 27 (2011) 1581–1582.

[82] K. McClymont, O. Soyer, Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways, Nucleic acids research 41 (2013).

[83] H. Kuwahara, M. Alazmi, X. Cui, X. Gao, MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind, Nucleic acids research 44 (2016) W217–W225.

[84] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, V. Hatzimanikatis, Group contribution method for thermodynamic analysis of complex metabolic networks., Biophysical journal 95 (2008) 1487–99.

[85] E. Noor, H. Haraldsdóttir, Consistent Estimation of Gibbs Energy Using Component Contributions, PLoS computational ... 9 (2013).

[86] P. Carbonell, P. Parutto, J. Herisson, S. B. Pandit, J.-L. Faulon, XTMS: pathway design in an eXTended metabolic space., Nucleic acids research (2014) gku362–.

[87] I. Buhaescu, H. Izzedine, Mevalonate pathway: A review of clinical and therapeutical implications, 2007.

[88] C. N. S. Santos, M. Koffas, G. Stephanopoulos, Optimization of a heterologous pathway for the production of flavonoids from glucose, Metabolic Engineering 13 (2011) 392–400.

[89] H. Liu, K. R. M. Ramos, K. N. G. Valdehuesa, G. M. Nisola, W. K. Lee, W. J. Chung, Biosynthesis of ethylene glycol in Escherichia coli, Applied Microbiology and Biotechnology 97 (2013) 3409–3417.

[90] M. Cavia-Saiz, M. D. Busto, M. C. Pilar-Izquierdo, N. Ortega, M. Perez-Mateos, P. Muñiz, Antioxidant properties, radical scavenging activity and biomolecule protection capacity of flavonoid naringenin and its glycoside naringin: A comparative study, Journal of the Science of Food and Agriculture 90 (2010) 1238–1244.

[91] M. L. Peach, A. V. Zakharov, R. Liu, A. Pugliese, G. Tawa, A. Wallqvist, M. C. Nicklaus, Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software., Future medicinal chemistry 4 (2012) 1907–32.

[92] J. Kirchmair, A. H. Göller, D. Lang, J. Kunze, B. Testa, I. D. Wilson, R. C. Glen, G. Schneider, Predicting drug metabolism: experiment and/or computation?, Nature Reviews Drug Discovery (2015).

[93] N. K. Mishra, S. Agarwal, G. P. Raghava, Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule, BMC Pharmacology 10 (2010) 8.

[94] C. A. Marchant, E. M. Rosser, J. D. Vessey, A k -Nearest Neighbours Approach Using Metabolism-related Fingerprints to Improve In Silico Metabolite Ranking, Molecular Informatics 36 (2017) 1600105.

[95] F. Mu, C. J. Unkefer, P. J. Unkefer, W. S. Hlavacek, Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds., Bioinformatics (Oxford, England) 27 (2011) 1537–45.

[96] M. Kotera, Y. Tabei, Y. Yamanishi, T. Tokimatsu, S. Goto, Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets., Bioinformatics (Oxford, England) 29 (2013) 135–44.

[97] D. Pertusi, M. Moura, J. Jeffryes, S. Prabhu, B. Walters Biggs, K. Tyo, Predicting novel substrates for enzymes with minimal experimental effort with active learning, Metabolic Engineering 44 (2017).

[98] M. Moura, D. Pertusi, S. Lenzini, N. Bhan, L. J. Broadbelt, K. E. J. Tyo, Characterizing and predicting carboxylic acid reductase activity for diversifying bioaldehyde production, Biotechnology and Bioengineering 113 (2016) 944–952.

[99] T. V. Sivakumar, V. Giri, J. H. Park, T. Y. Kim, A. Bhaduri, REACT-PRED: a tool to predict and analyze biochemical reactions, Bioinformatics 32 (2016) btw491.

[100] B. Delépine, T. Duigou, P. Carbonell, J.-l. Faulon, RetroPath2.0: A retrosynthesis workflow for metabolic engineers, doi.org (2017) 141721.

[101] A. Kumar, Elucidation and Synthetic Design of Biochemical Pathways using novoStoic, 2017.

[102] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, M. Kanehisa, PathPred: an enzyme-catalyzed metabolic pathway prediction server., Nucleic acids research 38 (2010) 138–43.

[103] P. Carbonell, P. Parutto, C. Baudier, C. Junot, J.-L. Faulon, Retropath: Automated Pipeline for Embedded Metabolic Circuits., ACS synthetic biology (2013).

[104] D. A. Pertusi, A. E. Stine, L. J. Broadbelt, K. E. J. Tyo, Efficient searching and annotation of metabolic networks using chemical similarity, Bioinformatics 31 (2015) 1016–1024.

[105] N. Hadadi, Computational Studies on Cellular Metabolism : From Biochemical Pathways to Complex Metabolic Networks 6667 (2015).

[106] M. a. Campodonico, B. a. Andrews, J. a. Asenjo, B. O. Palsson, A. M. Feist, Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path, Metabolic Engineering (2014) 1–18.

[107] S. C. Hammer, A. Marjanovic, J. M. Dominicus, B. M. Nestl, B. Hauer, Squalene hopene cyclases are protonases for stereoselective Brønsted acid catalysis, Nature Chemical Biology 11 (2014) 121–126.

[108] C. S. Henry, L. J. Broadbelt, V. Hatzimanikatis, Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate., Biotechnology and bioengineering 106 (2010) 462–73.

[109] M. Tokic, N. Hadadi, M. Ataman, D. S. Neves, B. E. Ebert, L. M. Blank, L. Miskovic, V. Hatzimanikatis, Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors, doi.org (2017) 209569.

[110] M. Moura, J. Finkle, S. Stainbrook, J. Greene, L. J. Broadbelt, K. E. Tyo, Evaluating enzymatic synthesis of small molecule drugs, Metabolic Engineering 33 (2016) 138–147.

[111] J. Gao, L. B. M. Ellis, L. P. Wackett, The University of Minnesota Pathway Prediction System: multi-level prediction and visualization., Nucleic acids research 39 (2011) 406–11.

[112] A. Cho, H. Yun, J. H. Park, S. Y. Lee, S. Park, Prediction of novel synthetic pathways for the production of desired chemicals, BMC Systems Biology 4 (2010).

[113] P. Carbonell, J. Wong, N. Swainston, E. Takano, N. J. Turner, N. S. Scrutton, D. B. Kell, R. Breitling, J.-L. Faulon, Selenzyme: Enzyme selection tool for pathway design, bioRxiv (2017) 188979.

[114] A. Chang, I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. W. Sensen, D. Schomburg, BRENDA in 2015: exciting developments in its 25th year of existence, Nucleic Acids Research 43 (2015) D439–D446.

[115] A. Burgard, M. J. Burk, R. Osterhout, S. Van Dien, H. Yim, Development of a commercial scale process for production of 1,4-butanediol from sugar, Current Opinion in Biotechnology 42 (2016) 118–125.

[116] K. Medley, R. Toofanny, M. Galdzicki, Y.-E. A. Ban, H. Sauro, A. Zanghellini, Synthetic Enzymes for Synthetic Biology, in: Symposium on Biotechnology for Fuels and Chemicals.

[117] C. H. Schilling, R. Thakar, E. Travnik, S. Van Dien, S. Wiback, SimPheny: A Computational Infrastructure for Systems Biology (????).

[118] D. S. Tawfik, Messy biology and the origins of evolutionary innovations., Nature chemical biology 6 (2010) 692–6.

[119] R. a. Notebaart, B. Szappanos, B. Kintses, F. Pal, a. Gyorkei, B. Bogos, V. Lazar, R. Spohn, B. Csorg, a. Wagner, E. Ruppin, C. Pal, B. Papp, Network-level architecture and the evolutionary potential of underground metabolism, Proceedings of the National Academy of Sciences (2014).

[120] A. L. Meadows, K. M. Hawkins, Y. Tsegaye, E. Antipov, Y. Kim, L. Raetz, R. H. Dahl, A. Tai, T. Mahatdejkul-meadows, L. Xu, L. Zhao, M. S. Dasika, A. Murarka, H. Jiang, L. Chao, P. Westfall, J. Lenihan, D. Eng, J. S. Leng, C.-l. Liu, W. Jared, J. Lai, S. Ganesan, P. Jackson, R. Mans, D. Platt, C. D. Reeves, P. R. Saija, T. S. Gardner, A. E. Tsong, G. Wichmann, V. F. Holmes, K. Benjamin, W. Paul, Rewriting yeast central carbon metabolism for industrial isoprenoid production, Nature Publishing Group 537 (2016) 694–697.

[121] M. J. Smanski, J. Casper, R. M. Peterson, Z. Yu, S. R. Rajski, B. Shen, Expression of the Platencin Biosynthetic Gene Cluster in Heterologous Hosts Yielding New Platencin Congeners, Journal of Natural Products 75 (2012) 2158–2167.

[122] T. Schwander, S. Burgener, T. J. Erb, A synthetic pathway for the fixation of carbon dioxide in vitro, Science 354 (2016).

[123] J. G. Jeffryes, R. L. Colastani, M. Elbadawi-Sidhu, T. Kind, T. D. Niehaus, L. J. Broadbelt, A. D. Hanson, O. Fiehn, K. E. J. Tyo, C. S. Henry, MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics., Journal of cheminformatics 7 (2015) 44.

[124] N. Hadadi, J. Hafner, A. Shajkofci, K. Zisaki, V. Hatzimanikatis, ATLAS of Biochemistry A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies., ACS synthetic biology (2016).

[125] M. R. Showalter, T. Cajka, O. Fiehn, Epimetabolites: discovering metabolism beyond building and burning, Current Opinion in Chemical Biology 36 (2017) 70–76.

[126] Z. Lai, T. Kind, O. Fiehn, Using Accurate Mass Gas ChromatographyMass Spectrometry with the MINE Database for Epimetabolite Annotation, Analytical Chemistry 89 (2017) 10171–10180.

[127] J. Sun, J. G. Jeffryes, C. S. Henry, S. D. Bruner, A. D. Hanson, Metabolite damage and repair in metabolic engineering design, Metabolic Engineering 44 (2017) 150–159.

[128] E. Van Schaftingen, R. Rzem, A. Marbaix, F. Collard, M. Veiga-da Cunha, C. L. Linster, Metabolite proofreading, a neglected aspect of intermediary metabolism., Journal of inherited metabolic disease (2013).

[129] A. D. Hanson, C. S. Henry, O. Fiehn, V. de Crécy-Lagard, Metabolite Damage and Metabolite Damage Control in Plants., Annual review of plant biology 67 (2016) 131–52.

[130] W. P. Jones, A. D. Kinghorn, Extraction of Plant Secondary Metabolites, in: Natural Products Isolation, Humana Press, Totowa, NJ, 2006, pp. 323–351.

[131] C. Lerma-Ortiz, J. G. Jeffryes, A. J. L. Cooper, T. D. Niehaus, A. M. K. Thamm, O. Frelin, T. Aunins, O. Fiehn, V. de Crécy-Lagard, C. S. Henry, A. D. Hanson, 'Nothing of chemistry disappears in biology': the Top 30 damage-prone endogenous metabolites., Biochemical Society transactions 44 (2016) 961–71.

[132] H. Lee, W. C. DeLoache, J. E. Dueber, Spatial organization of enzymes for metabolic engineering, Metabolic Engineering 14 (2012) 242–251.

[133] G. Sachdeva, A. Garg, D. Godding, J. C. Way, P. A. Silver, In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner, Nucleic Acids Research 42 (2014) 9493–9503.

[134] S. Di Fiore, Q. Li, M. J. Leech, F. Schuster, N. Emans, R. Fischer, S. Schillberg, Targeting tryptophan decarboxylase to selected subcellular compartments of tobacco plants affects enzyme stability and in vivo function and leads to a lesion-mimic phenotype., Plant physiology 129 (2002) 1160–9.

[135] U. Heinig, M. Gutensohn, N. Dudareva, A. Aharoni, The challenges of cellular compartmentalization in plant metabolic engineering, 2013.

[136] J. A. Jones, T. D. Toparlak, M. A. G. Koffas, Metabolic pathway balancing and its role in the production of biofuels and chemicals, Current Opinion in Biotechnology 33 (2015) 52–59.

[137] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, G. Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference., Proceedings of the National Academy of Sciences of the United States of America 107 (2010) 6286–91.

[138] S. Ding, X. Liao, W. Tu, L. Wu, Y. Tian, Q. Sun, J. Chen, Q.-N. Hu, EcoSynther: A Customized Platform To Explore the Biosynthetic Potential in ¡i¿E. coli¡/i¿, ACS Chemical Biology 12 (2017) 2823–2829.

[139] M. Khosraviani, M. S. Zamani, G. Bidkhori, FogLight: An efficient matrix-based approach to construct metabolic pathways by search space reduction, Bioinformatics 32 (2015) 398–408.

[140] J. Wicker, T. Lorsbach, M. Gutlein, E. Schmid, D. Latino, S. Kramer, K. Fenner, enviPath - The environmental contaminant biotransformation pathway resource, Nucleic Acids Research (2015) 1–7.

Name	Description	Availability
AGPathFinder[80]	Tracks groups of atoms to search for the shortest thermodynamic feasible path between source and target metabolites.	Webservice not longer available
Carbon Flux Paths[67]	Applies carbon mappings to search for the shortest mass-balanced path between source and target metabolite.	
EcoSynther[138]	Stochastically searches for reaction pathways and uses flux balance analysis to evaluate feasibility	Webservice available: rxnfinder.org/-ecosynther/
FogLight[139]	Reduces pathway prediction search space by converting biochemical databases into Boolean graphs to identify meaningful enzyme-reaction sets	
Metabolic Tinker[82]	Given a source and target metabolite and using structure similarity, searchers for thermodynamic feasible paths across a database of all known chemistry.	Webservice no longer available
optStoic[68]	Optimizes the overall stoichiometry for converting a reactant to desired product.	Web application: narrative.kbase.us/#catalog-/modules/MaranasTools
PathTracer[69]	Uses carbon transfer maps to search for the shortest and/or most actives paths between two metabolites.	
MRE[83]	Examines competition with organisms endogenous reactions/enzymes when ranking potential pathways between source and target metabolite.	Webservice available: cbrc.kaust.edu.sa/mre/
MRSD[81]	Favors reaction conservation by accounting for reaction frequency across all organisms in KEGG.	
ReTrace[78]	Python tool searches for branching pathways that transfer as many atoms as feasible from source to target metabolite.	Under GNU license: github.com/epitkane/-ReTrace
RouteSearch[79]	Searches for optimal paths by minimizing the loss of atoms from source to target metabolite	Pathway Tools component: biocyc.org/download.html

Table 1: Constraint-based and graph-based reaction prediction algorithms

Name	Description	Availability
BNICE.ch[105]	Uses bond electron matrix addition to generate new compounds	
Cho et. al[112]	A set of 5 pathway scoring factors is used to evaluate reactions from a small set of manual curated rules	
enviPath[140]	Uses manually curated SMARTS-based reaction rules to propose metabolite degradation pathways	Webservice available: envipath.org
GemPath[106]	Manually curated SMARTS-based reaction rules are used to integrate novel reactions into metabolic models	
NovoStoic[101]	Combines reaction rules with known biochemistry and a constraint-based system for pathway selection	
PathPred[102]	A webserver using RDM patterns to propose novel reactions. Has a rule set designed for plant metabolism	Webservice available: genome.jp/tools/-pathpred
Pickaxe	Python application using manually curated SMARTS rules for enzymatic and spontaneous reaction predictions	Under MIT license: github.com/JamesJeffryes/-MINE-Database
ReactPred[99]	Java application that can generate SMARTS rules from example reactions and apply them for prediction	Under GNU license: sourceforge.net/projects/-reactpred
Retropath2.0[100]	KNIME workflow using a precalculated set of automatically generated SMARTS rules from *E. Coli* metabolism	Non-commercial use: www.myexperiment.org/-workflows/4987.html
XTMS[86]	A webserver containing precalculated paths to over 2000 products from *E. Coli*	Webservice available: xtms.issb.genopole.fr/

Table 2: Rule-based reaction prediction algorithms