Abstract

Exhaustive identification of all alternate possible pathways that exist within metabolic networks can provide valuable insights into cellular metabolism. With the growing number of metabolic reconstructions, there is a need for an efficient method to enumerate pathways, which can also scale well to large metabolic networks, such as those corresponding to microbial communities.

We developed MetQuest, an efficient graph-theoretic algorithm to enumerate all possible pathways of a particular length between a given set of source and target molecules. Our algorithm employs a guided breadth-first search to identify all feasible reactions based on the availability of the precursor molecules, followed by a novel dynamic-programming based enumeration, which assembles these reactions into pathways producing the target from the source. We demonstrate several interesting applications of our algorithm, ranging from predicting amino acid biosynthesis pathways to identifying the most diverse pathways involved in degradation of complex molecules. We also illustrate the scalability of our algorithm, by studying larger graphs such as those corresponding to microbial communities, and identify several metabolic interactions happening therein.

A Python-based implementation of MetQuest is available at https://github.com/RamanLab/MetQuest

1 Introduction

Genome-scale metabolic networks are very useful to understand the complex network of metabolic reactions happening inside cells [37]. Typically, a genome-scale metabolic network consists of thousands of reactions and metabolites, which capture vital metabolic pathways such as the biosynthesis of amino acids and lipids, ATP synthesis, as well as transport of molecules inside the cells. Over the recent years, the construction of such
networks has increased tremendously, in part also due to the growing number of genomes sequenced [34]. There are several well-established constraint-based methods to analyse metabolic networks [41]; however, these methods require well-curated genome-scale metabolic models for making reliable predictions. The number of such well-curated models are very few [13, 34, 42], in comparison to the genome sequences available. Alternative approaches to analyse genome-scale models are based on network “topology” [4, 5, 18] or Boolean rules [26, 27]. In the former, the metabolic networks are abstracted as networks or graphs, and graph-theoretic algorithms are employed to predict or infer metabolic pathways. The latter set of methods are based on definitions of Boolean functions for identifying if particular reactions can proceed or not. Both methods generate qualitative predictions and do not entail the requirement of well-curated genome-scale models.

Several graph-based methods, using naïve network expansion [17, 18], atom–atom mapping, subgraph matching or stoichiometry [4, 22, 35, 39] have been developed previously. These methods, regardless of the approach, aim to determine the route(s) of conversion between sets of source and target molecules. The algorithms based on network expansion traverse the graph depending on the availability of precursor compounds and determine different routes of conversions from the metabolic network. The input graph to these algorithms can be represented in different ways, such as a substrate graph [28] or a hypergraph [32], each of which captures information at different levels of complexity. Substrate graphs are one of the simplest forms of representation, where all metabolites participating in a reaction are connected to each other. Traversing these graphs using breadth-first or depth-first search to identify different routes of conversion often leads to erroneous paths such as a 2-step glycolysis, converting glucose to pyruvate (via ADP), since these could proceed using the connections between “side” metabolites. To avoid such spurious results, several tools, such as FMM (From Metabolite to Metabolite; [7]) and Metabolic Tinker [30] use substrate graphs after excluding the connections from (high degree) “currency metabolites” such as H\(^+\), Water and NADPH. However, the substrate graph representation fails to capture the requirement of more than one reactant for any given reaction to occur.

To overcome these problems, algorithms were developed to operate on more informative bi-partite and hypergraph representations of metabolic networks [10, 11, 17, 18, 32]. These representations consider the participation of multiple compounds in a reaction to produce product(s), which is crucial while depicting metabolic networks. One such hypergraph-based algorithm, Rahnuma [32], performs a depth-first search to identify pathways that lead to the target metabolite from the source. In a few other studies [17, 18], the metabolic networks are represented as bipartite graphs to first determine the scope of the starting seed metabolites. The synthesis pathways were then assembled by backtracking from the target to the source, which becomes computationally challenging due to the presence of branched and cyclic pathways in metabolic networks.

The problem with determining branched-pathways has been addressed in few studies, which seek to use atom tracing. In one such method [19] BPAT-S (Branched Pathfinding using Atom Tracking and Seed pathways), seed pathways (linear) are first identified between the source and target compounds, based on the atom loss/gain. The branched pathways are then identified by finding the linear pathways between the metabolites contribut-
ing to this atom loss/gain. These pathways are then ranked based on the total number of reactions and the number of conserved atoms. ReTrace [40], another algorithm to find branched pathways, tries to combine linear shortest paths based on the transfer of a high fraction of atoms from source to target, thereby finding only the \( k \)-shortest paths.

Another class of methods use either information from atom-mapping, thermodynamics or structural transformation. For instance, the algorithm developed by [4] defines a set of atom-mapping rules based on the possible chemical conversions and seeks to identify paths where the atom loss is minimum. The predictions of this algorithm is restricted to the first \( k \)-shortest paths. PathPred [35], another tool to predict metabolic pathways, performs subgraph matching on the graph generated from KEGG RPAIR database. Similarly, RouteSearch [25], another algorithm based on branch-and-bound search, finds paths based on the mapped network generated from the atom mappings.

Despite the availability of several types of methods to understand and analyse metabolic networks, a simple and efficient method with minimum input for performing large-scale analyses is still lacking. Also, the current methods to identify pathways between the source and target molecules are restricted to smaller networks and do not enumerate long pathways, thereby restricting the scope of analyses. Although few of these algorithms handle cyclic and branched pathways, they require additional information such as atom transfer, which may not be readily available from the semi-curated metabolic networks. Further, we find that many of these tools are web-based, and do not lend easily to large-scale analyses; a large fraction of these tools are also currently inaccessible (Supplementary Table S1).

Thus, there is a need for a method to find pathways, which (a) requires only the topology of the reaction network (rather than stoichiometry and atom mapping), (b) is simple and scalable to metabolic networks (especially those comprising more than one organism), (c) efficiently handles cyclic and branched pathways, and (d) examines multiple alternate routes of conversion. To this end, we developed MetQuest, an efficient and scalable graph-theoretic algorithm, which exhaustively identifies all the pathways, or “sub-networks” between a given set of source and target metabolites using the input reaction network. Due to the nature of implementation, we are able to successfully identify branched and cyclic pathways. In comparison to the other algorithms, we observe better results, in terms of the completeness of the output pathways. Since we perform exhaustive enumeration, we not only recover the well-known pathways such as glycolysis and amino acid biosynthesis, but also identify several diverse pathways such as those involved in biodegradation of pollutants such as catechol. Further, because of the scalability of our algorithm to operate on larger metabolic networks, we are also able to demonstrate several metabolic exchanges happening in known natural and synthetic microbial communities.

2 Methods

In this section, we present a detailed overview of our algorithm, broadly divided into two phases. The first phase involves a guided breadth-first search (BFS) of the metabolic network, to identify all metabolites that can be reached from the source. In the second phase, we design a dynamic programming algorithm, which solves the non-trivial problem of assembling
reactions into pathways to produce the metabolite of interest.

The primary aim here, is to find different sets of reactions that would be involved in the conversion of source(s) to target(s). Specifically, we find the pathways that comprise at most $\beta$ reactions each, to enable this conversion. We define a pathway or sub-network as a set of reactions that are together necessary to convert a given set of source compounds to one or more target molecules. MetQuest seeks to efficiently traverse a metabolic network with branched and cyclic pathways and identify sub-networks for a given pair of source–target sets.

The input to MetQuest is a directed bipartite graph $G$ derived from a given metabolic network, a specified set of source ($Y$) and a set of target ($T$) metabolites, a set of seed metabolites ($S$), a number $\beta$ denoting the maximum number of reactions in any output sub-network. Any given metabolic network can be represented as a directed bipartite graph $G(M, R, E)$, where $M$ is the set of metabolites in the metabolic network, $R$ is the set of reactions and $E$ is the set of edges. Directed edges connect reactant metabolites ($m_i \in M$) to a reaction node ($r_j \in R$) or a reaction node to product metabolites. Such a representation disallows invalid conversions as may be interpreted from substrate graphs and helps in generating valid paths with biologically meaningful conversions [12].

2.1 Phase 1: Guided BFS

BFS is a classic graph traversal technique that visits all nodes of a given graph, starting at a source node, in a breadth-first fashion. BFS employs a queue of vertices, where newly discovered vertices are enqueued, to be processed at a later stage. A complete description of the BFS algorithm can be found elsewhere [8]. We modify the standard BFS by guiding it, based on the availability of precursor metabolites. Starting with the set of seed metabolites $S$, the algorithm first finds all the reactions from set $R$, whose precursor metabolites are in $S$. Such reactions are marked “visited” and added to the visited reaction set $R_v$. The metabolites produced by these reactions, $m_c$, are then added to $S$. The traversal continues in a breadth-first manner, incrementally adding triggerable reactions to the BFS queue. The expansion stops when there are no further reactions that can be visited. During the expansion, a reaction node is labelled as stuck, if it does not (yet) have the necessary precursors in $S$. Such reactions are automatically triggered if the precursor metabolites are produced at any later stage. A formal description of the algorithm can be found in Supplementary Algorithm S1. The traversed graph consists of all reactions that can be visited.

At the end of the traversal, we obtain the scope $M_s \subseteq M$ of $S$ and the set of visited reaction nodes, $R_v$. We also obtain the minimum number of steps to reach any metabolite $m$ and reaction $r$, starting at the seed metabolite set $S$, denoted as $\ell_m$ and $\ell_r$, respectively. We note that the value $\ell_m$ may not indicate the exact minimum number of reactions required to produce a metabolite. Instead, we only leverage it for algorithm optimisation (see Supplementary Methods §4.1). The scope $M_s$ comprises all metabolites that can be produced from the seed set $S$, in the given metabolic network. This process resembles the ideas of network expansion [18] and forward propagation [1] reported earlier. The former method seeks to identify the synthesising capacity of metabolic network using the input seed set of metabolites. The latter aims to identify the minimal precursor sets of seed metabolites that are required to produce a given set of target metabolites.
This step in our algorithm also derives the information about the scope of metabolites from the given metabolic network; it however distinguishes itself by making a systematic note of the visited and stuck reaction nodes, which are exploited at a later stage for efficient and exhaustive enumeration of biosynthetic pathways.

### 2.2 Phase 2: Generation of Sub-Networks

MetQuest uses a recursive dynamic programming formulation for computation of sub-networks. However, it avoids repeated recursive calls by memoising the precomputed values. The algorithm steps are listed below and the pseudo code can be found in Algorithm 1. Our algorithm maintains a table of size $M_s \times \beta$ and initialises all the entries to $\bot$, indicating that we do not (yet) know about the sub-networks for every metabolite. Recall that $M_s$ and $\beta$ denote the scope of seed metabolites and the maximum number of reactions we allow in a sub-network ("size cut-off") respectively.

We start filling the table entries by first considering the seed metabolite set $S$. For every seed metabolite $m \in S$, the entry in corresponding cell $\text{Table}(m, 0) = \emptyset$, indicating that no reaction is required to produce it; entry against other metabolites remain as $\bot$ (Lines 1–5).

**Algorithm 1 MetQuest**

1. for each metabolite $m \in M_s$ do
2.  if $m \in S$ then
3.   $\text{Table}[m][0] = \emptyset$;
4.  else
5.   $\text{Table}[m][0] = \bot$;
6. for each reaction $r \in R_v$ whose inputs are seed metabolites do
7.  for each output $m$ of $r$ do
8.   $\text{Table}[m][1] = \text{Table}[m][1] \cup \{r\}$;
9. for $k = 2 \ldots \beta$ do
10.  for each reaction $r \in R_v$ do
11.   Let $n$ denote the number of inputs to $r$.
12.   Let $m_1, m_2 \ldots m_r$ denote the inputs to $r$.
13.   for $\ell = k - 1 \ldots n \times (k - 1)$ do
14.     if $l \leq n \times (k - 2)$ then
15.       $t = \left\lfloor \frac{l}{k-1} \right\rfloor$.
16.     for $\ell' = 1 \ldots t$ do
17.       Fix $\ell'$ inputs that take a value of $k - 1$.
18.       This can be done in $\binom{n}{\ell'}$ ways.
19.       $\mathcal{P} = \text{generatePartitions}(\ell - (k - 1)\ell', n - \ell', k - 1)$;
20.     else
21.       $\mathcal{P} = \text{generatePartitions}(\ell, n, k - 1)$;
22.     for each $p \in \mathcal{P}$ do
23.       $\text{Table} = \text{populateTable}(p)$

**Note:** $k$ denotes the maximum size of the sub-network that we intend to find. $\text{Table}[m][k]$ contains exactly $k$ reactions to generate $m$ from the seed and source nodes. At the end of all iterations, any entry $\text{Table}[m][k] = \bot$ implies that $m$ cannot be generated using exactly $k$ reactions.
2.2.1 Algorithm Steps

1. In the first iteration, we find all the reactions \( r \in R_v \), whose inputs are only seed metabolites. For every product metabolite \( m \) produced by \( r \), we fill Table \((m, 1)\) with the reaction that produced \( m \) (Lines 6–8).

2. We now propose that a pathway whose size \( k = 2, \ldots, \beta \), consisting of a triggerable reaction \( r \in R_v \), can be constructed by merging the pathways that generated the reactants of \( r \). Let \( n \) denote the number of inputs to reaction \( r \) and \( m_1, m_2, \ldots, m_r \) denote the metabolites participating in \( r \). In order to account for the branched pathways, it is necessary to generate all possible combinations of pathways producing these metabolites. However, evaluating all these combinations becomes computationally challenging since the metabolic networks are branched and there may be multiple combinations to be evaluated. To achieve this efficiently, we generate partitions using \( n \) non-negative integers, which sum from \( k - 1 \) to \( n \times (k - 1) \) (see \texttt{generatePartitions}, as detailed in Supplementary Methods §4.2). By assessing the partitions for all the values till \( n \times (k - 1) \), we guarantee the generation of all the pathways for the metabolites in the scope \( M_s \) (Supplementary Methods §4.3). In order to achieve this sum, we choose from the set of integers whose value is between \( \ell_m \) and \( k - 1 \) (and not 0 and \( k - 1 \), as a part of optimization).

3. We generate such partition of numbers for all the reactions \( r \in R_v \), whose \( \ell_r \leq k \).

4. Using these partitions of numbers, we fetch the sub-networks of a particular length (dictated by the numbers in that partition) for the input metabolites \( m \) from the corresponding cell entry in the table.

5. At a time, we consider one sub-network \( SN \) of every metabolite \( m \) in the input of \( r \).

6. We then perform a union of these sub-networks for all the input metabolites \( m \), and finally add the reaction \( r \) to it. We denote this resultant sub-network as \( RS \). These operations are carried out by the function \texttt{populateTable} (Algorithm 2).

7. Against every output metabolite \( m \) of \( r \), based on the value obtained in \( RS \), we add the entry to corresponding cell.

8. Since there can be several sub-networks producing these metabolites, we repeat the steps 5, 6 and 7 for all the entries.

9. The final reaction set will be a union of all the reaction sets that have been found for this metabolite.

10. We divide the process of generating partitions for numbers into two major stages, in order to avoid repetitions and optimise on running times (Lines 13–19 and Lines 20–21). More details can be found in Supplementary Methods §4.1.

We note that the output pathways for every metabolite are generated by considering the sub-networks of all the reactants of a reaction. Due to this, any pathway generated by our algorithm is complete, i.e., it comprises of all the reactants required by the reactions constituting the pathway. Further,
Algorithm 2 populateTable(p)

1: function populateTable(p)
2: // Table, r, number n and inputs to r are global
3: if Table[m_k][p_k] = ⊥ for any k ∈ 1...n then
4:   continue;
5:   RS = \{ r \} ∪ Table[m_1][p_1] ∪ Table[m_2][p_2] ∪ ... ∪ Table[m_s][p_s]; //
          // Here, the ∪ is overloaded
6:   for each output y of R do
7:     Table[y][|RS|] = Table[y][|RS|] ∪ R
8: return Table
9: end function

since for any metabolite, we perform the union of sets of reactions, we avoid repeated generation of the same reaction sets. Due to this, we automatically report only the first occurrence of cyclic pathways. The demonstration of our algorithm on a toy-network and a cyclic pathway can be found in Supplementary Methods §4.4 and §4.5 respectively. Further implementation details and the formal proof of correctness for our algorithm can be found in Supplementary Methods §4.6 and §4.7, respectively.

3 Results

In this section, we show that our algorithm can recover sub-networks of different types, including the well-known glycolysis and amino acid synthesis pathways. We also compare our output sub-networks with those generated by other algorithms, and demonstrate that MetQuest produces better results since it generates complete pathways. Further, we showcase the ability of MetQuest to identify diverse pathways involved in biodegradation of an important industrial pollutant. Finally, we show that MetQuest scales well to large networks, and is therefore a powerful tool to predict and understand metabolic exchanges in microbial communities.

3.1 MetQuest exhaustively identifies multiple pathways in metabolic networks

Genome-scale metabolic networks catalogue numerous metabolic pathways happening inside a cell. Identifying these pathways helps us to better understand the level of redundancy in the cells for producing key metabolites. To this end, we applied MetQuest on well-curated networks to identify and understand pathways between different substrates and products.

3.1.1 Central carbon metabolism

We identify well-known biochemical pathways such as glycolysis using the genome-scale metabolic model of \textit{E. coli} iJO1366 [38]. We constructed the corresponding directed bipartite graph, consisting of 12,974 edges and 5,659 nodes, and specified glucose as the source node and pyruvate as the target node. We included the co-factors, co-enzymes and the energy currencies in the seed metabolite set \( S \) (Supplementary Table S2) and found all the sub-networks within a size cut-off \( β \) of 15.
We successfully recovered the well-known glycolysis pathway and, in addition, due to the exhaustive enumeration, we could also identify 12,686 paths from the seed and the source set of metabolites to the target of varying lengths. Of all these paths, 6,175 of them use glucose as the source, while the rest of the paths comprise compounds formed by seed metabolites reacting amongst themselves.

These observations raise an important question: “Why should so many pathways potentially exist inside the cell in addition to glycolysis, which is the most preferred pathway?” To answer this question and understand the differences between each of these pathways, we studied the similarity between the obtained sub-networks, using the Jaccard index. Of \( \approx 1.9 \times 10^7 \) combinations of sub-networks analysed we find only 2,858 (0.0149%) sub-networks are similar with Jaccard values between 0.93 and 0.98, while 8,992,712 (47.17%) sub-networks are much different with Jaccard values between 0.03 and 0.13 (Supplementary Figure S1).

The former set of sub-network pairs, with higher Jaccard values can be understood in light of the presence of multiple transporters, which are either used to transport compounds from the environment or between different compartments of the cells. These transporters function under varied environmental conditions, where they employ different mechanisms to transport glucose inside the cell [20]. The latter set, with lower Jaccard values, point towards the existence of alternate pathways that produce identical precursor metabolites to synthesise pyruvate. These predominantly include methyl glyoxal pathway, Entner-Doudoroff pathway and pentose-phosphate pathway. A few of these alternate pathways were also shown to outperform the canonical pathways under different sets of physiological conditions [9].

### 3.1.2 Amino acid biosynthesis

We also determined sub-networks between different metabolites in a compartmentalised model of *S. cerevisiae* iMM904 [33]. As there are additional transport reactions between the compartments, we used a higher cut-off \( \beta = 30 \) and found the sub-networks between D-Glucose and amino acids such as L-phenylalanine. We assumed a seed metabolite set with 42 metabolites (Supplementary Table S3), consisting of the co-factors and co-enzymes in different compartments, and found sub-networks to 638 metabolites (\( |M_s| \) was 681). Specifically, we identified 1,293 sub-networks from the source to target of varying lengths. From the sub-networks, we find that many of these amino acids are derived from the intermediates of the central carbon metabolism (Supplementary Results §5.1). Also, the sub-networks producing aromatic amino acids involve metabolites produced in different compartments of the cell, thereby requiring longer steps for conversion (Figure 1).

### 3.2 MetQuest also uncovers metabolic dependencies

We also investigated if there were specific centrally located amino acids in the metabolic network, which play a crucial role in synthesising other amino acids. Towards this end, we constructed a bipartite graph using the genome-scale metabolic model of *E. coli* iJO1366 [38]. We initialized our algorithm with a set of seed metabolites consisting of essential cofactors and D-glucose as the source node (Supplementary Table S4), and generated all the sub-networks for a cut-off \( \beta = 20 \). We analysed the sub-networks for synthesis
Figure 1. Sub-network with a size cut-off $\beta$ of 28 between D-glucose and L-phenylalanine in *S. cerevisiae* iMM904. Dark gray rectangles represent reactions, light gray circles represent metabolites. The nomenclature of reaction and metabolite names are consistent with iMM904 genome-scale metabolic model. The source and the target metabolite is gluc-D$_{c}$ and phe-L$_{c}$ respectively (shown in boldface and italics). We note that the synthesis of L-phenylalanine involves several metabolites from different compartments. The reactions in this pathway and the seed metabolites can be found in Supplementary Results §5.1.
of different amino acids, to determine if there were considerable differences in their sizes. We studied this by progressively adding amino acids (one at a time) to the seed set, and determining if there were significant reductions in the size of sub-networks producing other amino acids. In each iteration, the amino acid that is produced via the fewest steps is added to the seed set. At the end of all iterations, we found amino acids such as L-Glutamate to be the key players in amino-acid biosynthesis. This is because these amino acids, when added to the seed metabolite set, considerably reduce the size of sub-networks producing amino acids such as L-serine and L-alanine (Supplementary Results §5.2). These results reiterate the central role of L-glutamate in the metabolic network as shown previously [3].

Further, we determined the sub-networks to these amino acids, and found one of our sub-networks producing L-histidine, matches the one reported in LumpGEM [2], which uses a Mixed-Integer Linear Programming based approach to identify reaction sub-networks. We could, however, not find any sub-network producing L-cysteine, L-methionine and L-tryptophan. This is because the former two amino acids need special co-substrates such as S-Adenosyl-L-methionine for their synthesis. These co-substrates, could neither be synthesised by this network from the given seed metabolites, nor were a part of the input seed provided. The latter amino acid, L-tryptophan is derived from the reaction of L-serine and other compounds such as indole, C’-(3-Indolyl)-glycerol 3-phosphate. These examples showcase the ability of MetQuest to exhaustively identify not only long pathways, but also uncover the metabolic dependencies in the metabolic network.

3.3 MetQuest reveals diverse paths for catechol biodegradation

Phenols and catechols are among the primary pollutants in the effluent from several industries, such as the chemical, textile and steel. So far, many studies have used Pseudomonas putida for bioremediation of such pollutants [24]. We sought to identify the pathways that render P. putida unique and tailored for this application. To this end, we investigated pathways involved in the metabolism of such compounds, which are otherwise harm-
| Source                  | Target                    | Length | Output sub-network                          | Comments                                                                 |
|------------------------|---------------------------|--------|---------------------------------------------|--------------------------------------------------------------------------|
| L-Arginine (C00062)    | L-Citrulline (C00327)     | 2      | R00551, R00665                              | Matches with ATLAS and FMM                                              |
| Pyruvate (C00022)      | Itaconate (C00490)        | 4      | R02491, R00209, R00237, R02405              | Matches with FMM, FMM does not report R00209 which produces C00024 – required by R02405† |
| Pyruvate (C00022)      | Itaconate (C00490)        | 5      | R00351, R02243, R00209, R00217, R01325     | Matches with FMM, FMM does not report R00351 which produces C00036 – required by R00351† |
| L-Tyrosine (C00082)    | Naringenin (C00509)       | 5      | R02446, R00737, R01616, R01613, R06641     | Matches with FMM, FMM does not report R06641†                           |
| L-Phenylalanine (C00079)| Resveratrol (C03582)     | 5      | R01616, R00697, R02253, R06641, R01614     | Matches with FMM, FMM does not report R06641 which produces malonyl-CoA required by R01614† |
| Mevalonic acid (C00418) | Amorpha-4,11-diene (C16028)| 7      | R01658, R03245, R02245, R01121, R01123, R07630, R02003 | No paths found by FMM, ATLAS, however it is natively found in S. cerevisiae [29]. |
| D-Erythrose 4-phosphate (C00279) | 3-Amino-5-hydroxy benzoate (C12107) | 7      | —                                            | No paths reported by ATLAS, FMM and our algorithm                       |

Table 1. The sub-networks produced between different source and target molecules for a given length. These results were obtained on the graph constructed using reactions in the KEGG database. We denote reactions and compounds using their respective KEGG IDs. It is interesting to note that our algorithm is able to recover the well-known pathway reported in *S. cerevisiae*, while the other algorithms fail to do so. Also, there exists no pathways of length 7 between C00279 and C12107, which is being consistently reported by all the methods. More details about the sub-networks can be found in Supplementary Results §5.4.

†Note that we did not find any pathways in ATLAS, when queried for pathways using no novel reactions. These reactions, integrate KEGG metabolites into novel enzymatic steps and are present only in the ATLAS database.

...ful to many other micro-organisms. Further, we explored if the breakdown of catechols could directly lead to energy producing pathways. Hence, we identified all the pathways that start from catechols and lead to the intermediates of the tricarboxylic acid (TCA) cycle. Since catechols require complex mechanisms to be degraded to smaller intermediates, we used a higher cut-off $\beta$ of 25. Using this, we found 825 different paths of lengths up to 25 that degrade catechols to fumarate. Interestingly, our algorithm could identify the two most different mechanisms of ortho- and meta-cleavage used by *Pseudomonas putida* to degrade catechols to the intermediates of the TCA cycle (Figure 2), as reported by [15]. These sub-networks can be found in Supplementary Results §5.3.

### 3.4 MetQuest excels in comparison with other algorithms

To benchmark our algorithm, we compared our results with those obtained from some of the already existing path-finding methods such as FMM [7],
Table 2. Number of sub-networks found for each target of given cut-off β from the seed metabolite set S (which also includes the source metabolite). Note that there are many sub-networks of length 20 for pyruvate in all the models but there are no sub-networks of length 10 that produce oxaloacetate in most of the models.

| Models / Sub-network length | Targets     | Pyruvate | Oxaloacetate |
|----------------------------|-------------|----------|--------------|
|                             |             | 10 15 20 25 | 10 15 20 25 |
| *Escherichia coli* core model | 14 142 2358 226 | 6 111 603 251 |
| *Lactococcus lactis*         | 1 530 1673 1908 | No paths 81 No paths 543 |
| *Pichia stipitis iBB804*     | No paths 425 131 453 | No paths 65 267 790 |
| *Saccharomyces cerevisiae iMM904* | 3 1 1047 475 | No paths 65 1423 2514 |

We find that MetQuest performs better, in terms of the number and the completeness of the pathways generated, i.e., the output sub-networks are complete, in that they have all the reactions necessary to produce every reactant in that pathway. From the sub-networks (Table 1), we observe that the smaller pathways of length 2 completely match with those generated by the other algorithms. However, in many cases, we identify longer pathways, since these involve metabolites generated by branched pathways. It is interesting to note that our algorithm was able to correctly identify the already reported pathway between C00418 (Mevalonic acid) and C16028 (Amorpha-4,11-diene) [29], which was not identified by the other algorithms. All the sub-networks reported in Table 1 can be found in Supplementary Results §5.4.

We also tested the performance of MetQuest by finding sub-networks of varying size cut-offs (10–25) between glucose and other key metabolites involved in the central carbon metabolism on several genome-scale metabolic networks of different sizes (Table 2). We used a uniform set of seed metabolites consisting of essential co-enzymes and co-factors for all the simulations (Supplementary Table S5) and determined the sub-networks of sizes 10, 15, 20 and 25. During each simulation, we measured the time taken and the number of sub-networks generated for the target metabolites pertaining to the given size cut-off β. We carried out all the simulations on an Intel Core i7-2600 Desktop with 24GB RAM, running Ubuntu 14.04. Although we see an increase in the simulation time as the model size and the cut-off β increases (Supplementary Figure S2), MetQuest never fails to generate pathways, even at higher values of β.

3.5 MetQuest scales well to large metabolic networks

Beyond individual networks, it is also interesting to identify sub-networks in communities of organisms and understand their metabolic interactions, which are known to play a critical role in determining the stability of microbial communities [14]. However, community metabolic networks are much larger in size, presenting many more challenges for path-finding. Analysing such large graphs of microbial communities demands a scalable and efficient algorithm, which requires only minimum information such as network topology. Further, the algorithm should be capable of identifying longer
pathways to capture many metabolic exchanges happening within a community. The existing algorithms, such as OptCom [44] and cFBA [23], are based on constraint-based techniques, and require well-curated metabolic models. Also, they do not directly lend themselves to identify metabolic interactions in draft network reconstructions. To this end, we use MetQuest on joint graphs (metabolic networks) of multiple organisms and identify sub-networks between different metabolites. Specifically, we construct the community bipartite graph of a microbial community by considering a common extracellular space through which metabolite exchanges can happen. Further, we analyse the sub-networks generated by MetQuest and show that it can correctly recover previously reported metabolic interactions.

### 3.5.1 MetQuest correctly predicts metabolic exchanges in a synthetic E. coli community

To study the metabolic interactions in E. coli, [43] performed computational analyses on genetically modified E. coli strains. Specifically, two different E. coli strains with knockouts of b2276 and b3708 genes were modelled together and the metabolic interactions were predicted. We simulated these gene knockouts by removing the reactions catalyzed by these genes, after considering the Gene-Protein-Reaction relationships. We then constructed the community bipartite graph and applied MetQuest to predict the metabolic exchanges between these two strains. We assumed seed metabolites, with glycolate as source (as given in [43]) and applied MetQuest to find all sets of sub-networks to every metabolite within the scope of glycolate, pyruvate and both these sources independently, with a size cut-off $\beta = 20$.

From the sub-networks generated, we determined the metabolites that can be potentially exchanged between the strains. We were able to recover the reported acetate and formate exchanges happening between the two strains. In addition, we observe that acetate (from $\Delta b2276$) participates in the production of several important target metabolites in $\Delta b3708$ such as L-glutamine, inosine and several other co-enzymes of central carbon metabolism. In addition, we also found other metabolites such as alpha-ketoglutarate, ethanol, acetaldehyde exchanged between the organisms, which can be potential candidates for experimental verification (Supplementary Results § 5.5).

### 3.5.2 Predicting novel interactions in community metabolic networks

To illustrate the performance of MetQuest on a larger network, we computed multiple sub-networks in an experimentally demonstrated three-member microbial consortium [31], and determined the metabolic interactions. Towards this, we used the genome-scale metabolic models of Clostridium cellulolyticum (cc), Desulfovibrio vulgaris (dv), and Geobacter sulfurreducens (gs) from the Path2Models database [6]. We constructed the three-member community bipartite graph by connecting the genome-scale metabolic models through their common exchange reactions. We defined the seed metabolite set $S$, consisting of essential salts, co-factors, co-enzymes and a set of tRNA molecules. We specified source as cellobiose (as reported in [31]) and chose to determine sub-networks with a size cut-off $\beta = 20$, to all metabolites within the scope $M_s$ of seed metabolites. To determine the metabolic exchanges, we analysed every sub-network of 1,526 metabolites ($M_s$) with a special focus on sub-networks involving exchange metabolites.
From this analysis, we found many sub-networks having at least one exchange reaction between the organisms. We observed that the acetate and ethanol exchanges proposed in [31] predominantly led to the production of amino acids such as L-serine, L-leucine, L-aspartate and L-valine (in gs) and L-threonine and L-glycine (in dv) respectively, which contribute to the biomass, thereby enabling a stable microbial consortium.

4 Discussion

With the increasing number of draft genome-scale metabolic reconstructions, there is a pressing need for efficient algorithms to analyse these metabolic networks and generate useful predictions. It is particularly important that these algorithms perform efficiently with minimum information such as the reaction topology, and also identify pathways in large (multi-)genome-scale metabolic networks.

To this end, we propose MetQuest, a scalable and efficient graph-based algorithm for identifying all possible metabolic pathways in genome-scale metabolic models. Specifically, we focus on exhaustively determining all alternate pathways (of a particular size) between a set of source and target metabolites. We also demonstrate the application of our algorithm by identifying multiple pathways involved in several important metabolisms such as the central-carbon metabolism and the amino acid metabolism. Further, we show its usefulness in determining several diverse pathways, which take part in catechol degradation. Besides its applications on the metabolic networks of individual organisms, MetQuest, due to its scalability, can be applied to study larger and more complex networks.

A number of graph-based methods to find pathways in metabolic networks have been developed so far. These methods convert the metabolic networks either into substrate graphs, bipartite graphs or hypergraphs. However, these methods are well-suited only for smaller networks and lower cut-offs. For instance, MetaPath [17] tries to back trace the pathway from the target to the source based on the information obtained from the scope calculation. Since the metabolic networks contain many branched pathways, identifying the routes of conversion by back-tracing becomes computationally expensive, and can break down rapidly at higher cut-offs. Another method, Rahnuma [32], tries to solve the problem of path prediction on genome-scale metabolic networks by abstracting these networks into hypergraphs. Rahnuma, based on depth-first search on hypergraphs, seeks to obtain pathways between two metabolites. However, the performance of Rahnuma on metabolic networks was demonstrated only for shorter path-lengths of 6, and it may not be applicable for identifying longer pathways, such as those involved in the biosynthesis of aromatic amino acids from simple sugars. Moreover, the pathways are computed based on a condition that the metabolite can be used only once in any pathway, which inherently fails to capture cyclic pathways. MetQuest, on the other hand, can identify branched and cyclic pathways, and can perform well on larger metabolic networks for much longer cut-offs, thereby extending its application to analyse microbial communities.

We have also demonstrated the major strengths of MetQuest: in multiple examples we considered, of individual metabolic networks and real-world microbial communities, MetQuest was able to identify multiple pathways and correctly predict the metabolic exchanges/interactions taking place.
MetQuest was able to correctly predict the acetate exchange happening between the two genetically modified E. coli strains. Further, MetQuest also predicted novel interactions involving amino acids such as L-cysteine in the microbial community reported in [31].

However, as with any modelling technique, MetQuest also has its limitations. Firstly, the predictions from MetQuest rely heavily on the quality of the underlying metabolic network. Specifically, few of the metabolic interactions predicted by MetQuest could be due to model artefacts. There may be other metabolic interactions happening between microbes, which could be absent in the metabolic network (gaps in the metabolic network). Due to this reason, such interactions may not be predicted by MetQuest. Further, our algorithm does not assign any weights to metabolites/paths. Although there are other methods like cFBA and OptCom, which make more quantitative predictions, they also demand better curated metabolic networks. We primarily see the use of MetQuest as a first-line investigatory tool, to cull down from a very large set of possible communities to a more tractable set, which may be studied further using constraint-based techniques or through wet lab experiments.

In sum, we report a novel and efficient algorithm MetQuest, which (i) rapidly enumerates all possible biosynthetic pathways from a given metabolic network, (ii) efficiently handles branched and cyclic pathways, and most importantly, (iii) scales well to very large networks such as those representing microbial communities. Further, our algorithm requires only the genome-scale metabolic networks, a set of seed and target metabolites and a size cut-off, as against thermodynamic or atom-mapping information, thereby rendering it a very potent tool to perform large-scale analyses. Overall, we believe that MetQuest is a very useful tool for exploring and understanding metabolic pathways in large-scale networks, to generate testable hypotheses for further experiments.

Funding

This work was supported by the Indian Institute of Technology Madras grant ERP/1314/004/RESF/KARH to KR and the INSPIRE fellowship, Department of Science and Technology, Government of India to AR.

References

1. V. Acuna, P. V. Milreu, L. Cottret, A. Marchetti-Spaccamela, L. Stougie, and M.-F. Sagot. Algorithms and complexity of enumerating minimal precursor sets in genome-wide metabolic networks. *Bioinformatics*, 28(19):2474–2483, 2012.

2. M. Ataman and V. Hatzimanikatis. lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLoS Comput. Biol.*, 13(7):e1005513, 2017.

3. J. Berg, J. Tymoczko, and L. Stryer. *Biochemistry*. W.H. Freeman, 5th edition, 2002.

4. T. Blum and O. Kohlbacher. MetaRoute: Fast search for relevant
metabolic routes for interactive network navigation and visualization. Bioinformatics, 24(18):2108–2109, 2008.

5. E. Borenstein and M. W. Feldman. Topological signatures of species interactions in metabolic networks. J. Comput. Biol., 16(2):191–200, 2009.

6. F. Büchel, N. Rodriguez, N. Swainston, C. Wrzodek, T. Czauderna, R. Keller, F. Mittag, M. Schubert, M. Glont, M. Golebiewski, M. van Iersel, S. Keating, M. Rall, M. Wybrow, H. Hermjakob, M. Hucka, D. B. Kell, W. Müller, P. Mendes, A. Zell, C. Chaouiya, J. Saez-Rodriguez, F. Schreiber, C. Laibe, A. Dräger, and N. Le Novère. Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst. Biol., 7(1):116, 2013.

7. C. H. Chou, W. C. Chang, C. M. Chiu, C. C. Huang, and H. D. Huang. FMM: A web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res., 37:129–134, 2009.

8. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to Algorithms. The MIT Press, USA, 3rd edition, 2009.

9. S. J. Court, B. Waclaw, and R. J. Allen. Lower glycolysis carries a higher flux than any biochemically possible alternative. Nat. Commun., 6:1–8, 2015.

10. D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res., 33(Web Server):W326–W330, 2005.

11. D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Inferring meaningful pathways in weighted metabolic networks. J. Mol. Biol., 356(1):222–236, 2006.

12. Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak. An overview of data models for the analysis of biochemical pathways. Brief. Bioinform., 4(3):246–259, 2003.

13. A. Dräger and B. Ø. Palsson. Improving Collaboration by Standardization Efforts in Systems Biology. Front. Bioeng. Biotechnol., 2:1–20, 2014.

14. K. Faust and J. Raes. Microbial interactions: from networks to models. Nat. Rev. Microbiol., 10(8):538–550, 2012.

15. C. F. Feist and G. D. Hegeman. Phenol and benzoate metabolism by Pseudomonas putida: regulation of tangential pathways. J. Bacteriol., 100(2):869–877, 1969.

16. N. Hadadi, J. Hafner, A. Shahkoç, A. Zisaki, and V. Hatzimanikatis. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. ACS Synth. Biol., 5(10):1155–1166, 2016.

17. T. Handorf and O. Ebenhöh. MetaPath Online: A web server implementation of the network expansion algorithm. Nucleic Acids Res., 35:613–618, 2007.
18. T. Handorf, O. Ebenhöh, and R. Heinrich. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, 61:498–512, 2005.

19. A. P. Heath, G. N. Bennett, and L. E. Kavraki. An Algorithm for Efficient Identification of Branched Metabolic Pathways. *J. Comput. Biol.*, 18(11):1575–1597, 2011.

20. K. Jahreis, E. F. Pimentel-Schmitt, R. Brückner, and F. Titgemeyer. Ins and outs of glucose transport systems in eubacteria. *FEMS Microbiol. Rev.*, 32(6):891–907, 2008.

21. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, 44:D457–D462, 2016.

22. P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, 11(1):40–79, 2010.

23. R. a. Khandelwal, B. G. Olivier, W. F. M. Röling, B. Teusink, and F. J. Bruggeman. Community flux balance analysis for microbial consortia at balanced growth. *PLoS One*, 8(5):e64567, 2013.

24. A. Kumar, S. Kumar, and S. Kumar. Biodegradation kinetics of phenol and catechol using *Pseudomonas putida* MTCC 1194. *Biochem. Engg. J.*, 22(2):151–159, 2005.

25. M. Latendresse, M. Krummenacker, and P. D. Karp. Optimal metabolic route search based on atom mappings. *Bioinformatics*, 30(14):2043–2050, 2014.

26. W. Lu, T. Tamura, J. Song, and T. Akutsu. Integer programming-based method for designing synthetic metabolic networks by Minimum Reaction Insertion in a Boolean model. *PLoS One*, 9(3):e92637, 2014.

27. W. Lu, T. Tamura, J. Song, and T. Akutsu. Computing Smallest Intervention Strategies for Multiple Metabolic Networks in a Boolean Model. *J. Comput. Biol.*, 22(2):85–110, 2015.

28. H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003.

29. V. J. J. Martin, D. J. Pitera, S. T. Withers, J. D. Newman, and J. D. Keasling. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.*, 21(7):796–802, 2003.

30. K. McClymont and O. S. Soyer. Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways. *Nucleic Acids Res.*, 41(11):e113, 2013.

31. L. D. Miller, J. J. Mosher, A. Venkateswaran, Z. K. Yang, A. V. Palumbo, T. J. Phelps, M. Podar, C. W. Schadt, and M. Keller. Establishment and metabolic analysis of a model microbial community for understanding trophic and electron accepting interactions of subsurface anaerobic environments. *BMC Microbiol.*, 10(1):149, 2010.
32. A. Mithani, G. M. Preston, and J. Hein. Rahnuma: Hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831–1832, 2009.

33. M. L. Mo, B. O. Palsson, and M. J. Herrgård. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.*, 3(1):37, 2009.

34. J. Monk, J. Nogales, and B. O. Palsson. Optimizing genome-scale network reconstructions. *Nat. Biotechnol.*, 32(5):447–452, 2014.

35. Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Koteru, S. Goto, and M. Kanehisa. PathPred: An enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, 38:138–143, 2010.

36. J. Nogales, B. Ø. Palsson, and I. Thiele. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Syst. Biol.*, 2(1):79, 2008.

37. M. A. Oberhardt, B. Ø. Palsson, and J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, 5(320):320, 2009.

38. J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol.*, 7(1):535, 2011.

39. J. Pey, J. Prada, J. E. Beasley, and F. J. Planes. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol.*, 12(5):R49, 2011.

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

41. N. D. Price, J. L. Reed, and B. Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, 2(11):886–897, 2004.

42. A. Ravikrishnan and K. Raman. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Brief. Bioinform.*, 16:1057–1068, 2015.

43. E. Tzamali, P. Poirazi, I. G. Tollis, and M. Reczko. A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities. *BMC Syst. Biol.*, 5(1):167, 2011.

44. A. R. Zomorrodi and C. D. Maranas. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.*, 8(2):e1002363, 2012.