

A ROBUST PATHFINDING ALGORITHM USING CHEMICAL  
COMPOSITION

BY

PRATIK LAHIRI

THESIS

Submitted in partial fulfillment of the requirements  
for the degree of Master of Science in Bioengineering  
in the Graduate College of the  
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Associate Professor Kaustubh Bhalerao

# ABSTRACT

Metabolic pathfinding is the task of finding preferred metabolic pathways from metabolic large reaction databases. Representing metabolism via networks enables quick enumeration of paths between two compounds. Automated pathfinding helps in working with ever increasing databases of reactions and in finding novel pathways for metabolic engineering. However, the number of pathways between two compounds can be as large as 500,000 in some metabolic models and even more as the size of the input database grows, which makes it imperative that the most relevant ones are ranked highly. While graph theoretic representations of metabolic networks bring speed and ease in enumeration of pathways, they also create the challenge of biochemically insensible shortcuts through pool or currency metabolites.

In the past, strategies to circumvent such irrelevant pathways have included weighing networks using the degree of nodes or the manual curation of edges in the metabolic network. The former method wrongfully penalizes some primary metabolites central to metabolism, while the latter requires someone to complete manual curation. KEGG RPAIR database is an annotation to describe reactions in terms of reactant pairs and has been used for metabolic pathfinding. Here, I first study a few different centrality measures to identify currency metabolites and identify one better than the degree centrality. I then describe a method to augment the KEGG RPAIR based pathfinding method using a chemical composition score and evaluate its ability to augment and replace the role of RPAIRs in pathfinding. The new algorithm is validated against a set of 30 biochemical pathways in *E.coli*. Since this method uses chemical composition as a fallback measure, it can be used in the absence of explicit RPAIR information, thus allowing the identification of putative paths not possible via methods using the RPAIR database alone.

*To my parents and my brother, for their love and support and my grandfather  
for his unwavering faith and blessings.*

# ACKNOWLEDGMENTS

I would like to thank everyone who has helped me during this research project and while writing this thesis. First, I would like to thank my advisor for his guidance and support. This project would not have been possible without his regular comments, discussions and insights.

Next I would like to thank all the members of the Bhalerao lab for providing an environment where working on this project was fun and for the discussions that helped me along the way. Last, I would like to thank my friends and family for their support.

# TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION	1
1.1	Data Sources and Nature of Data	3
1.2	Representation of Metabolism	6
1.3	Structure of Metabolic Networks	6
1.4	Currency Metabolites	8
1.5	Motivation	9
CHAPTER 2	METABOLIC PATHFINDING	10
2.1	Early Work	10
2.2	Graph Based Methods	12
2.3	Atom Mapping Based Methods	14
CHAPTER 3	MATERIALS AND METHODS	19
3.1	Input Datasets	19
3.2	Reference Pathways	20
3.3	Graph Structure	24
3.4	Centrality Measures	25
3.5	Rpair Prediction	26
3.6	$k$ Shortest Paths	26
CHAPTER 4	RESULTS	29
4.1	Centrality Measures	29
4.2	Pathfinding	30
4.3	Robustness	34
CHAPTER 5	SUMMARY AND FUTURE WORK	37
REFERENCES		38

# CHAPTER 1

## INTRODUCTION

One important aspect of how organisms work is to understand metabolism. Common questions investigated include what energy sources can a cell survive on? What compounds does it produce and release to the environment? To answer these questions a key task is to identify metabolic pathways. Metabolism encompasses nutrient uptake, energy production, synthesis of proteins, DNA and other molecules that are required for a cell to survive and proliferate. These processes are tightly regulated. Enzyme expression is one of the primary ways regulating the rates of reactions and the uptake and release of compounds from the surrounding environment. Metabolic pathways are a coherent series of chemical reactions that convert nutrients from the environment to cell products and by-products. Metabolic pathways abstract a specific set of biochemical functions. We then have a description of how metabolism operates in an organism. This knowledge can be applied to metabolic engineering which is defined as “the improvement of cellular activities by manipulations of enzymatic, transport, and regulatory functions in the cell with the use of recombinant DNA technology” [1]. *E.coli* is a popular host for metabolic engineering. Such applications in *E.coli*, include for example, the production of glucaric acid using genes from *Saccharomyces cerevisiae* and mice [2], production of terpenoids- amorphadiene from a synthetic gene and a gene from *Saccharomyces cerevisiae* [3], 1,3-propanediol [4] and 1,2,4 butanetriol [5].

A metabolic network is composed of reactions that connect metabolites. It is a representation of all metabolic pathways and hence overall metabolism in an organism. Of particular interest for metabolic engineering is small molecule metabolism which is the set of chemical reactions that act upon small and medium sized molecules. These molecules are essential macromolecules like proteins, nucleic acids, lipids, sugars, co-factors, modulators of enzyme activity [6]. Metabolic pathways have been inferred from mutation experiments on model organisms such as *E.coli* and a few other bacteria, the yeast *S.cerevisiae* and focused studies in mouse and human. In such methods a gene is mutated and hence the corresponding enzyme is not produced. All products dependent on the reactions catalyzed by this enzyme are not produced and need to be provided in the environment, which is an auxotrophy phenotype. By observing all mutated phenotypes genes can be clustered into groups. Each group then corresponds to a metabolic pathway. With the large amount of genomic data being generated these days there has been an effort at *in – silico* metabolic reconstruction.

One set of tools are metabolic pathfinding tools. These methods model the compounds and reactions as a graph and use graph theory to find metabolic pathways which are difficult to infer using biochemical experiments. Apart from the tedium involved, biochemical methods are limited in their inability to identify pathways with lethal phenotypes or multiple alternatives (branched pathways because there is a combinatorial explosion of possible pathways).

Some of the challenges in *in – silico* pathfinding are, poor quality of data, inadequate quality of model, false positives (due to biochemically irrelevant pathways), and false negatives due to incomplete networks [7].

## 1.1 Data Sources and Nature of Data

MetaCyc [8, 9] and KEGG (Kyoto Encyclopedia of Genes and Genomes) [10, 11] are large metabolic pathways databases that have been maintained and expanded for some years now and are the most popular. The number of reactions in the two databases are similar. Metacyc (version 16.0 release Feb 17th, 2012) has more reactions (8,692 vs. 10,262) while KEGG (as of Feb 17th, 2012) has a lot more compounds (16,586 vs 11,991) and total reactions. Metacyc contains many more pathways from plants, fungi, actinobacteria that are not found in KEGG, while KEGG contains many pathways for xenobiotic degradation, glycan metabolism, metabolism of terpenoids and polyketides not present in Metacyc [12].

In these databases, pathways and reactions have varying levels of associated data. For example Metacyc, KEGG pathways and reactions have EC (Enzyme Classification) numbers [13] (Figure 1.1). Metacyc also has cross references to KEGG compound and reaction identifiers. KEGG in addition has RPAIRS (Reaction Pairs). RPAIRS are substrate-product pairs assigned to each reaction based on chemical transformation patterns called RDM (Reaction Difference Match) patterns and EC numbers [14, 15].

To generate RDM patterns and EC numbers, KEGG has classified atoms and their microenvironments into 68 atom types. These atom types are used in a graph based method to identify chemical similarities. This method uses an algorithm to find common isomorphic subgraphs of two compound graphs. The compound graphs are a 2D representation of the chemical structure with well detailed vertex labels taking into account the physiochemical environmental properties of atoms.



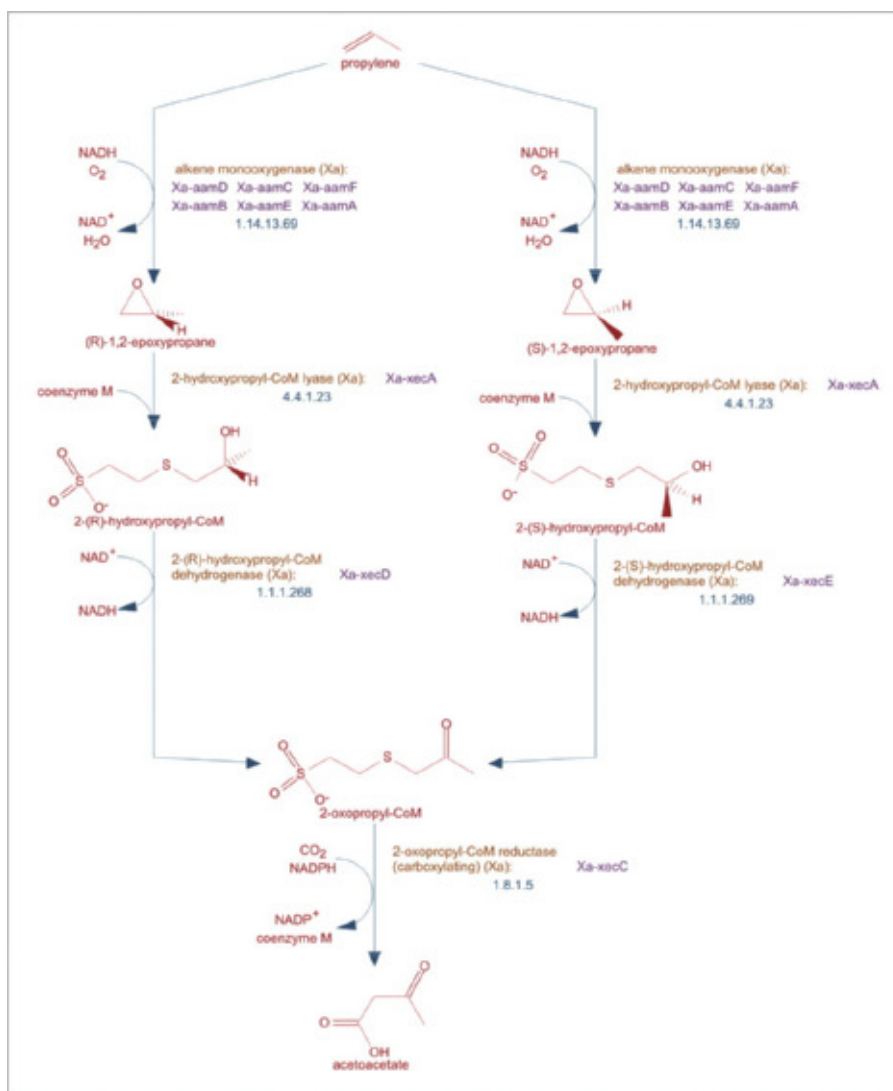


Figure 1.1: An example pathway in Metacyc reused with permission from [8]

The vertex labelling is done computationally on the basis of connection patterns of atoms and the functional groups they belong from the initial MDL/MOL file which are manually curated. RDM patterns are created by aligning the chemical structure of a substrate-product pair of compounds and identifying the reaction center, matched region and difference region. The RDM pattern is the KEGG atom type changes at these loci. The RDM pattern is then used to assign a category to a reactant pair.

The categories are:

**Main** describing main changes on substrates

**Trans** focused on transferred groups for transferases

**Cofac** describing changes on cofactors for oxidoreductases

**Ligase** describing the consumption of nucleoside triphosphates for ligases

**Leave** describing the separation or addition of inorganic compounds for such enzymes as lyases and hydrolases

The EC number is also inferred from the RDM pattern [15]. However, the reactant-pairs in the RPAIR database are further manually curated to weed out any errors in computational assignment. The computational assignment discussed above can be erroneous in cases where the overlapping atoms are few compared to the atoms in the the reactant-pair.

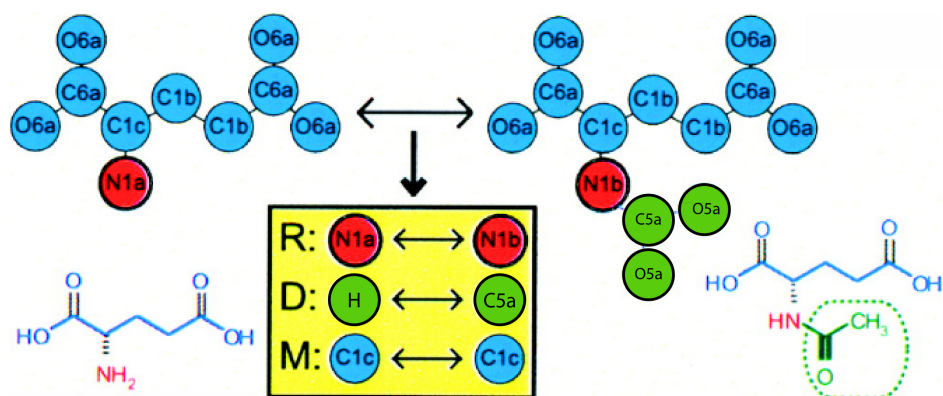


Figure 1.2: An example RDM pattern for a pair of compounds. Reaction center in red, Match in blue, Difference in Green. Reused with permission from [15]

## 1.2 Representation of Metabolism

Metabolism can be modelled using graph theory, flux balance analysis and petri nets. Here we use a graph theory based approach to solve the problem of pathfinding. A Metabolic network is a representation of metabolites (compounds) and their conversions (reactions). Reactions may be spontaneous or may be catalyzed by enzymes. While organisms regulate enzymes in response to environmental conditions we make the approximation in pathfinding that all annotated enzymes are expressed at all times and reactions are catalyzed at a significant velocity. We therefore ignore reaction rates in such problems. A metabolic pathway is a coherent series of successive reactions for a specific function (i.e it takes input compound(s) and converts them to output compound(s)), *e.g* gluconeogenesis which accepts pyruvate as input and produces glucose. When using graphs to model metabolism we can use compounds as nodes and reactions as links (compound graphs) or the reverse, that is reactions as nodes and compounds as links (reaction graph). In addition, a bipartite graph with both reactions and compounds as nodes can also be used. The compound graph and reaction graph are useful for structural analysis because some graph algorithms do not work with bipartite graph. The drawback is that biochemically irrelevant shortcuts may occur in path finding [16].

## 1.3 Structure of Metabolic Networks

Some early work using metabolic graphs was used to gain key insights into the structure of metabolic graphs such as hub metabolites, small world nature of metabolic graphs etc. [17, 18] These methods using a compound graph structure found that like other real world networks metabolic networks too

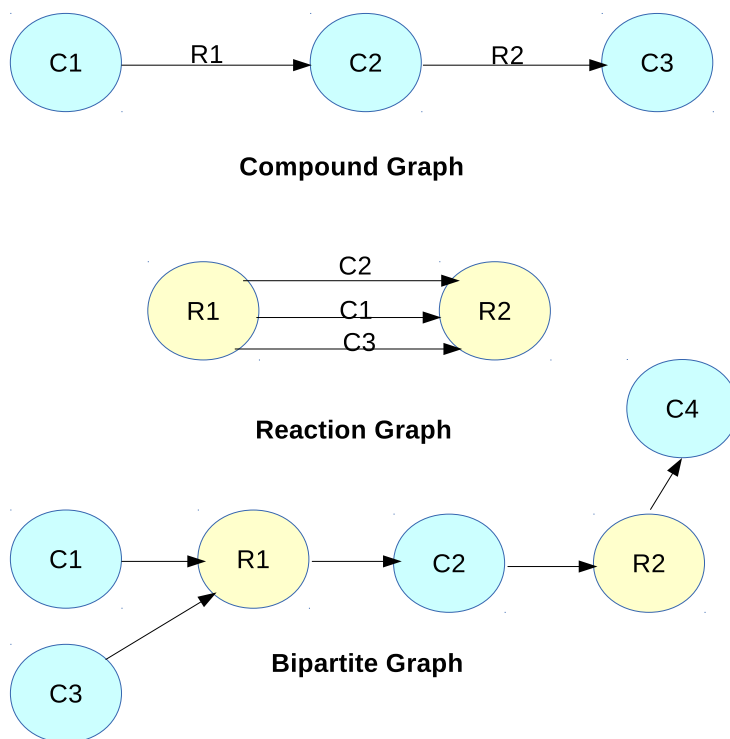


Figure 1.3: Different Graph Models.

follow the small world structure [19] and the node degrees follow a power law distribution. A small number of compounds are highly connected and most compounds have few connections. Additionally, most compounds are within 3 steps of each other and about 20 compounds are the most well connected in metabolic models of all organisms studied. These compounds on inspection were found to be currency metabolites or pool metabolites, *e.g.* ATP, ADP, NAD, H<sub>2</sub>O, H<sup>+</sup> etc., which are typically cofactors involved in energy and redox levels. A later study [20] on 80 organisms, with the removal of currency

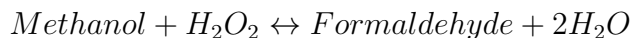
metabolites, found that the average length of all paths between two nodes was 7-8 steps and that the structure of metabolic networks could be grouped into 3 domains of organisms- eukaryotes, archaea and bacteria. They also found a different, more meaningful set of compounds that are hub nodes.

These hub nodes are: Glycerate-3-phosphate, D-Ribose-5-phosphate, Acetyl-CoA D-Ribose-5-phosphate, Acetyl-CoA Pyruvate, D-Xylulose 5-phosphate, D-Fructose 6-phosphate, 5-Phospho-D-ribose 1-diphosphate, L-Glutamate, D-Glyceraldehyde 3-phosphate, L-Aspartate, Propanoyl-CoA, Malonyl-ACP Mal, Succinate Acetate, D-Ribose-5-phosphate, D-Fructose 6-phosphate, 5-Phospho-D-ribose 1-diphosphate, D-Glucose 6-phosphate.

These compounds are intermediates in the pentosephosphate pathway, the citric acid cycle, glycolysis, TCA cycle, amino acid synthesis and are primary metabolites central to most metabolism.

## 1.4 Currency Metabolites

One major problem in metabolic path finding is the distortion of pathfinding solutions due to currency or pool metabolites pointed out by many papers [21] [20]. Currency metabolites have a high degree centrality because they are cofactors or side-products in many reactions. In a metabolic graph the presence of these compounds leads to biochemically irrelevant shortcuts. The following example illustrates this problem:



If we consider that all substrates are connected to all products we get a path  $\text{Methanol} \rightarrow \text{H}_2\text{O} \rightarrow \text{L} - \text{Lysine}$  which is meaningless. Lysine cannot be produced from methanol via water.

## 1.5 Motivation

In the next chapter we describe and do a thorough literature review of the most successful methods for metabolic pathfinding. All methods require some level of curated data. The requirement can range from complete description of acceptable compound transformation to some computational prediction followed by manual curation. Even for the methods that have mostly computation prediction of acceptable compound transformation, a complete description of compound structure information is needed. While, annotating reaction databases with the relevant information is a continuous process, there are some gaps. We have described here a method that uses annotation data where available and compensates in its absence.

# CHAPTER 2

## METABOLIC PATHFINDING

The earliest pathfinding approaches involved enumeration of paths from source to target in a directed graph. This was proved to be naïve when it was found that there were approximately 500,000 paths from glucose to pyruvate [22] and close to 350,000 from chorismate to tyrosine [23] due to the cyclic nature of the graph.

### 2.1 Early Work

An algorithm using artificial intelligence was created to find metabolic paths from a source to a target compound [24] [25]. This method used a small database of 70 reactions and 100 compounds. It enforced the exchange of carbon in biochemically valid pathways and required information of enzyme mechanism for each reaction. Another algorithm [26] [27] [28] [29] also found metabolic pathways but it required information about which compounds are present in the pathway. This work used a database of 250 reactions and 400 compounds. The work of Kuffner *et al.* [22] used Petri-Nets. Petri-Nets are bipartite graphs with two types of nodes- compounds (places) and reactions (transitions). Pathways are generated based on a “firing rule”. A firing rule typically is formed from the stoichiometry of a reaction. This method models metabolism as a concurrent process. Such a method fails to capture common situations involving, external metabolites or reversible reactions. These issues

were addressed by using colored nodes and specific types of transitions [30] [31] and used to model sucrose breakdown in potato tubers [32], pathway from chorismate to tryptophan [33]. However, these methods do not account for currency metabolites and enumerate a large number of pathways. To avoid open enumeration the current focus in the field has shifted to finding  $k$ -shortest pathways.

The key intuition in most path finding methods is that given a source and a target compound a directed path will provide insights into the intermediate reactions. Once the intermediate reactions are known the stoichiometry can be easily calculated. McShan *et al.* [34] viewed metabolism as a biochemical state space. Reactions are partitioned into two components- the chemical component and the biocatalytic component which represent the transformation and the catalysis of a reaction respectively. Each compound is represented as a vector of 145 features derived from the atoms and bonds, making a compound a point in the feature hyperspace. A chemical transformation or reaction is simply the difference in the feature vectors of the two compounds involved. Metabolic pathfinding is then reduced to the problem of finding state transitions from the source to the target compound. The state space is searched using the  $A^*$  algorithm and costs for transformations are calculated based on the Manhattan distance or the Euclidean distance of the transformation (difference of compound vectors).



## 2.2 Graph Based Methods

Authors have tried to circumvent the currency metabolite problem in metabolic graphs using different strategies. The earliest strategy was to remove a set of currency metabolites from the graph [35] [36] [37] [38] which were the most highly connected compounds, creating an adjusted graph. Removing compounds such as H<sub>2</sub>O, NAD, ATP, AMP leaves the resulting graph intact but removing a compound such as  $\beta$ -Alanine will break the graph into two resulting in many compounds being not reachable to each other.

Another strategy was to use connectivity as a measure to weight compounds and penalize them. This method has been used in a number of methods either to bias paths against currency metabolites or to rank pathways [39] [40] [41] [42].

Setting aside or penalizing currency metabolites globally is misleading. Currency metabolites must be determined locally, in the context of the reaction. For example in the following set of reactions:



If we remove H<sub>2</sub>O, ATP, ADP and Orthophosphate the two reactions (D-Glucose-1-phosphate  $\rightarrow$  ADP-glucose, Pyridoxal  $\rightarrow$  Pyridoxalphosphate ) can

still be retrieved. However the last two reactions cannot be found. This can happen when a currency metabolite is not only a co-factor in all reactions. For example, in the production of 1-Methyl-4-pyridone-3-carboximide from NAD<sup>+</sup> if NAD<sup>+</sup> is removed from the graph then the pathway from NAD<sup>+</sup> Methyl-4-pyridone-3-carboximide cannot be recovered.

So, it is necessary to integrate chemical knowledge into the process of pathfinding. Pathway Hunter [40] uses a metabolite mapping scoring function to determine relevant links between compounds. It uses the fingerprint algorithm in the CDK (Chemistry Development Kit) [43] to calculate the number of ‘on’ bits in all compounds of a reaction. Then for each substrate-product pair it calculates a similarity score using the Tanimoto coefficient [44].

$$S = \frac{|A \cap B|}{|A \cup B|}$$

where the numerator is the number of bits ‘on’ in both compounds and the denominator is the number of bits ‘on’ in either of the two compounds. This similarity score is multiplied by the percentage atomic mass contribution of the pair in the reaction to decide the most biochemically relevant substrate-product pair for a reaction.

A source for chemical knowledge is the group of KEGG databases. KEGG RPAIR discussed earlier contains mapping between substrate-product pairs that describe their biochemical relationship in a reaction. Faust *et al.* [41] have leveraged this database to create a novel heuristic. They created two new graph models: Rpair Graph and Reaction Specific Rpair Graph. The Rpair Graph has one node for each RPAIR which is connected to its constituent

compounds. The Reaction Specific Rpair Graph instantiates a separate node for each reaction that an RPAIR is involved in, which amounts to a node for each reaction with its associated RPAIRS connected to the compounds of the RPAIRS. The best results were obtained using Rpair Graphs with compound weighting and main-trans RPAIR filtering. However, these results are not the most precise as they do not provide any information about the reaction. The Reaction Specific Rpair Graph fills this void and is approximately as accurate as the Rpair Graph. A peculiarity of this method is that it requires the first and last reaction to be set. An analysis of this method [45] found that accuracy declines considerably when only the input and output compounds are specified.

## 2.3 Atom Mapping Based Methods

While the method described in [41] uses atom mapping rules, it does so to construct the metabolic graph. Further generation of pathways is done using graph theory. On the other hand there are methods that use atom mapping rules to guarantee that the product metabolite has atleast one atom from the source metabolite to target metabolite [46] [47] [48] or to ensure that there is a sequence of transformations where substrate and target substructures are isomorphic [49].

PathPred [49] uses the KEGG RDM pattern match and chemical structure match to search for paths. It accepts a query compound and its MOL file and does a global similarity search using SIMCOMP [14] [50] to generate matched compounds with their RDM patterns. These RDM patterns are then matched to the query compound to generate matched patterns. Next, the

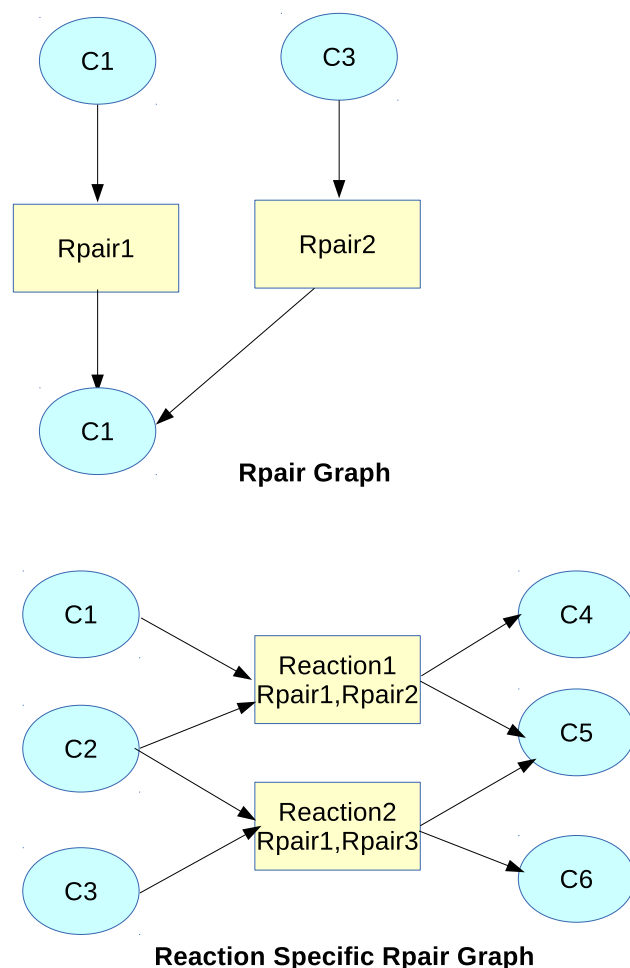


Figure 2.1

matched patterns are used to create transformed compounds from the query compound. These transformed compounds and their RDM patterns are added to the a pool of compounds and previous two steps are repeated until no more new matched patterns can be generated. In which case the transformed compounds are used as query compounds in the first step and the cycle continues a specified number of times. A Jaccard coefficient based scoring function between the query compound and the matched patterns is used to score and rank pathways. The method is restricted to two RDM pattern

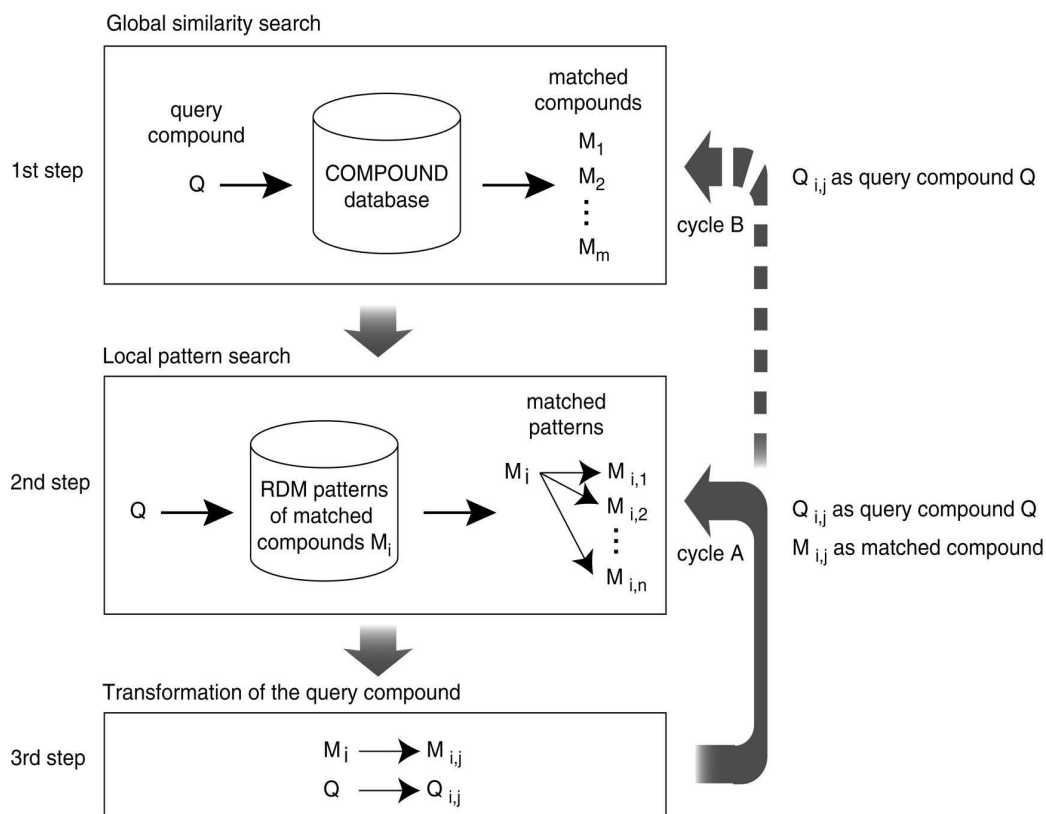


Figure 2.2: PathPred Algorithm Flowchart. Reused with permission from [49]

libraries- xenobiotic degradation in bacteria and biosynthesis pathways in plants, and one must be selected as input.

Arita *et al.* [46] first used atom mapping to guarantee transfer of at least one atom from source to target. It first enumerated all paths between source and target in the metabolite graph and then evaluated whether a carbon (or nitrogen or sulphur) atom was transferred from the source to the target. To do this evaluation it generated an atom mapping database computationally which was then thoroughly manually curated. The atom mapping generation method finds topologically maximum common subgraphs in pairs of compound graphs. However, this method fails in cases of isomerization, dimerization, cyclization, rearrangement of carbon skeleton from linear to branched, transfer of chemical

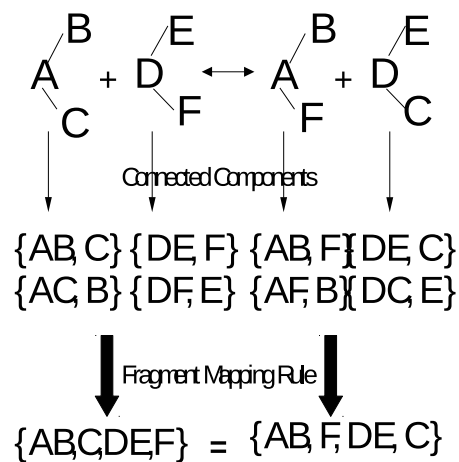


Figure 2.3: An example Fragment Mapping by graph partitioning described in [48]

groups, requiring manual curation.

A major improvement was developed by Blum *et al.* [48]. They also used the maximum common subgraph method to generate atom mapping rules but first generated fragment mapping rules, from which atom mapping rules were generated. Fragment mapping rules were generated using a chemical graph partitioning algorithm. This algorithm, given a cut size  $C$ , removes  $C$  edges from the graphs of all compounds in a reaction. This creates a sets of connected components (fragments) for each reactant. Finally, to generate fragment mapping rules, all pairs of combinations of connected components of substrates and products that are chemically equivalent are selected. Now that the fragments of substrates and products are mapped to each other most of the cases where the method by Arita *et al.* failed can be handled. A worked example for cut size 1 is shown in Figure 2.3. These become fragment mapping rules which can be used to infer atom mapping rules. To handle

multiple mapping rules per reaction, the reactions were clustered according to the first 3 digits of the EC number associated with the reaction and the fragment mapping rule most common in the cluster is used to create a reaction mechanism rule.

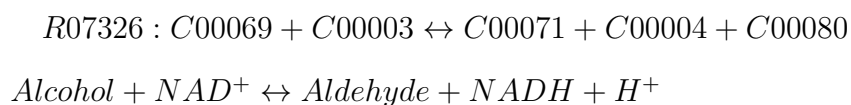
Alternatively, Heath *et al.* [47] used the manually curated alignment mapping for compounds in KEGG RPAIR to generate an atom mapping graph. In this graph, for each RPAIR, compound nodes were connected mapping nodes which stored information to map the atoms of the compound node on the incoming edge to the compound node on the outgoing edge. Then using a depth-first search all reachable nodes were found. During this depth first search at each node the mapped atoms were also stored as a transition history. Using the transition history  $k$ -shortest paths can be found.

# CHAPTER 3

## MATERIALS AND METHODS

### 3.1 Input Datasets

The graph model was built using data from the KEGG website. All reactions were downloaded from the KEGG Reaction database [10, 51, 11]. For each reaction downloaded, the reaction equation with stoichiometric coefficients, KEGG reaction identifiers, chemical formulas and all RPAIRS [15, 14, 52] were downloaded. Additionally all corresponding enzyme identifiers were also downloaded [15, 14, 52]. To filter reactions for those present in *E.coli* MG1655, all its genes were downloaded from the KEGG API using the list operation (<http://rest.kegg.jp/list/eco>). This operation lists all genes with enzyme identifiers. Only complete enzyme identifiers were selected. Enzyme identifiers representing a subclass or class of enzymes were discarded. These enzyme identifiers were then cross linked with the enzyme identifiers corresponding to downloaded reactions to filter non-*E.coli* MG1655 reactions. Further, to filter generic reactions the KEGG IUBMB (International Union Biochemistry and Molecular Biology) reaction heirarchy was utilized. An example of a generic reaction is shown below.





All generic reactions and reactions with substrates that also participate in a generic reaction were removed. All glycans and their corresponding compound identifiers were downloaded from KEGG. Any reaction with a glycan identifier was modified to contain the corresponding compound identifier.

## 3.2 Reference Pathways

In order to test the pathfinding method a list of E. Coli reference pathways was used [53]. These pathways were then looked up in EcoCyc [54] to curate the putative paths. These pathways are listed in Table 3.1. Since pathfinding does not find branched pathways unbranched linear paths were manually curated and side compounds were removed.

Table 3.1: Test data set of the curated pathways

Reference Pathway	Annotated Path
Gluconeogenesis	C00022 → C00074 → C00631 → C00197 → C00236 → C00118 → C00354 → C00085 → C01172
Glycolysis	C01172 → C00085 → C00354 → C00118 → C00236 → C00197 → C00631 → C00074 → C00022
Proline Biosynthesis	C00025 → C03287 → C01165 → C03912 → C00148

\*

————Continued On Next Page————

\*

Table 3.1: Test data set of the curated pathways

Reference Pathway	Annotated Path
Ketoglutarate Metabolism	C02780 → C01062, C06473 → C00257 → C00345
Pentose Phosphate Pathway	C01172 → C01236 → C00345 → C00199 → C00117, C00231 → C05382 → C00118
TCA cycle	C00036 → C00158 → C00417 → C00451 → C00026 → C00091 → C00042 → C00122 → C00149
NAD Biosynthesis	C00049 → C05840 → C03722 → C01185 → C00857 → C00003
Arginine Biosynthesis	C00025 → C00624 → C04133 → C01250 → C00437 → C00077 → C00327 → C03406 → C00062
Spermidine Biosynthesis	C00019 → C01137 → C00315
Threonine Degradation	C00188 → C03508 → C00037
Serine Biosynthesis	C00197 → C03232 → C01005 → C00065

\*

————Continued On Next Page————

\*

Table 3.1: Test data set of the curated pathways

Reference Pathway	Annotated Path
Histidine Biosynthesis	C00119 → C02741 → C04896 → C04916 → C04666 → C01267 → C01100 → C00860 → C01929 → C00135
Tyrosine Biosynthesis	C00251 → C00254 → C01179 → C00082
Coenzyme A Biosynthesis	C03492 → C04352 → C01134 → C00882 → C00010
Pentathenoate Biosynthesis	C00141 → C00966 → C00522 → C00864 → C03492
Tetrahydrofolate Biosynthesis	C00568 → C00921 → C00415 → C00101
Flavin Biosynthesis	C00044 → C01304 → C01268 → C04454 → C04732 → C04332 → C000255 → C00061 → C00016
Heme Biosynthesis	C01051 → C03262 → C01079 → C02191 → C00032
Pyrimidine Ribonucleotide Synthesis	C00064 → C00169 → C00438 → C00337 → C00295 → C01103 → C00105 → C00015 → C00075 → C00063

\*

————Continued On Next Page————

\*

Table 3.1: Test data set of the curated pathways

Reference Pathway	Annotated Path
Pyrimidine DeoxyRibonucleotide Synthesis	C00075 → C00460 → C00365 → C00364 → C00363 → C00459
Rhamnose Degradation	C00507 → C00861 → C01131 → C00424 → C00186 → C00022
Fucose Degradation	C01019 → C01721 → C01099 → C00424 → C00186 → C00082
Entner Duodoroff Pathway	C00345 → C04442 → C00118
Anearobic Respiration	C00022 → C00024 → C00158 → C00417 → C00451 → C00026
Arginine Biosynthesis	C00062 → C03296 → C03415 → C05932 → C05931 → C00025
Proline Degradation	C00148 → C03912 → C01165 → C00025
Glycolate Degradation	C00160 → C00048 → C01146 → C00258 → C00197
Glycerol Degradation	C00116 → C00093 → C00111 → C00118
Glutamate Biosynthesis	C00064 → C00006

\*

————Continued On Next Page————

\*

Table 3.1: Test data set of the curated pathways

Reference Pathway	Annotated Path
Phenylalanine Biosynthesis	C00251 → C00254 → C00166 → C00079
Allantoin Degradation	C02350 → C00499 → C02091 → C00603 → C00048
Cysteine Biosynthesis	C00065 → C00979 → C00097

### 3.3 Graph Structure

The reaction data was used to create a directed bipartite graph. For each compound the graph has a compound node. For each reaction the graph has a reaction node. Each reaction node has incoming directed edges from all substrate compounds and outgoing edges to all products. If the reaction is reversible the graph has directed edges from compounds to reaction in both directions. An example is in the Figure 3.1.

Each node in the graph stores some data. Compound nodes store compound formulas, and reaction nodes store a dictionary of RPAIRS, a list of substrates, a list of products. Edges between compound and reaction nodes store stoichiometric coefficients of the compounds in the reaction. The graph was implemented using Networkx [55]

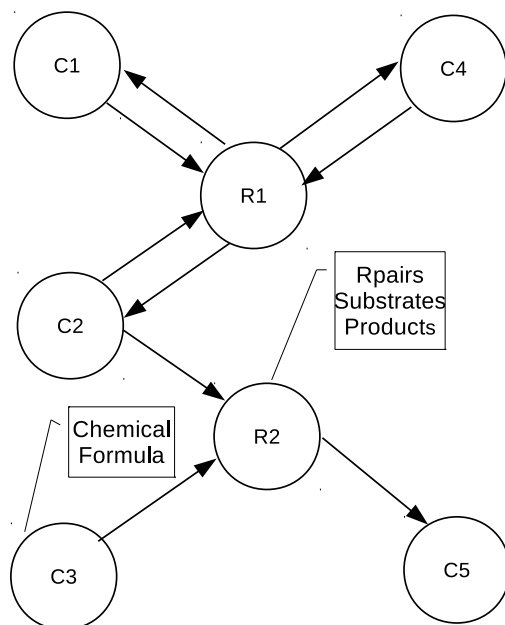


Figure 3.1: Graph Structure. R1,R2 are reaction nodes and C1-C5 are compound nodes. R1 is reversible, R2 is irreversible. Reaction nodes store RPAIR dictionary, list of substrates, list of products. Compound nodes store chemical formula

### 3.4 Centrality Measures

Various centrality measures were studied to differentiate between currency metabolites and non-currency metabolites. Closeness centrality [56] [57] is a measure of how close a vertex in a graph is to all other nodes.

$$c(v_i) = 1/(\sum_j d(v_i, v_j))$$

Betweenness centrality [58] is the number of shortest paths from all vertices to all other vertices that pass through v.

$$c(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} n_{jk}(v_i) / n_{jk}$$

where  $n_{jk}$  is the number of shortest paths between vertices  $v_j$  and  $v_k$  and  $n_{jk}(v_i)$  is the number of such paths that

contain  $v_i$

Page Rank centrality [59] measures direct and indirect importance of a vertex. These centrality measures were calculated using the in-built functions in Networkx [55]

### 3.5 Rpair Prediction

Since the RPAIR database is not complete, a method to fill in the gaps was created. Any pair of compounds that is a main or trans or ligase RPAIR will have significant structural and chemical similarity. Since there is a significant proportion of reactions with RPAIR data, checking for chemical similarity should suffice in pathfinding. The reactions with RPAIR data will restrict the possible paths and for reactions without RPAIR data a chemical composition constraint will eliminate the most irrelevant paths.

The chemical composition similarity score is a simple atomic composition similarity. For a pair of compounds the atomic profile is generated which has counts for each atomic element. For each common element between the two input compounds  $\min(c_i, c_j)$  where  $c_i$  and  $c_j$  are counts of the element in the two compounds is added to the similarity score. This similarity score is divided by the larger atomic count.

### 3.6 $k$ Shortest Paths

Yen's  $k$  Shortest Path Algorithm was modified to find the  $k$  shortest paths [60]. The shortest path function used was a modified Dijkstra's algorithm.

The algorithm for shortest paths ensures that a reaction vertex cannot be traversed twice. It also uses Rpair and chemical similarity scores to select

---

**Algorithm 1** Modified Dijkstra's Algorithm

---

```
1: procedure SHORTEST PATH(source,sink,graph,previousVertex)
2:   Q  $\leftarrow$  makePriorityQueue()
3:   insert(Q,(source,0))
4:   for each vertex  $u \neq$  source do
5:     insert(Q,(u, $\infty$ ))
6:   S $\leftarrow$   $\emptyset$ 
7:   if previousVertex then
8:     previous[source]=previousVertex
9:   for  $i=1$  to  $|V|$  do
10:    ( $v, \text{dist}(s,v)$ )=minPriorityQueue(Q)
11:    S=S  $\cup$  { $v$ }
12:    if  $v$  is a Reaction vertex then
13:      for  $u$  in Adj( $v$ ) do
14:        if ( $\text{prev}[v],u$ ) is an Rpair that is not of type leave then
15:          rpair=True
16:        else
17:          if similarity( $\text{prev}[v],u$ ) $>0.3 \wedge$   $\text{prev}[v],u$ 
18:            not co-reactants then
19:              sim=True
20:          if rpair  $\vee$  sim then
21:            cost=distances[ $v$ ] + weight( $v,u$ )
22:            if cost $<$ distances[ $u$ ] then
23:              distances[ $u$ ]=cost
24:              Q[ $u$ ]=cost
25:               $\text{prev}[u]=v$ 
26:          else
27:            cost=distances[ $v$ ] + weight( $v,u$ )
28:            if cost $<$ distances[ $u$ ] then
29:              distances[ $u$ ]=cost
30:              Q[ $u$ ]=cost
31:               $\text{prev}[u]=v$ 
```

---



acceptable paths. The similarity cutoff of 0.3 was chosen for similarity scores based on similarity scores of all known Rpairs. In case the source vertex is a reaction vertex the algorithm needs its previous vertex as a parameter. Such a case occurs only when the method is called from the k shortest path method and the previous vertex is known in this case.

The modified Yen’s k shortest paths algorithm is described in Algorithm 2.

Algorithm 2 uses the modified Dijkstra’s method (Algorithm 1) repeatedly

---

**Algorithm 2** Modified Yen’s Algorithm

---

```

1: procedure K SHORTEST PATHS(source,sink,graph,K)
2:   A[0]=Shortest Path(source,sink,graph)
3:   B=[ ]
4:   for k from 1 to K do
5:     for i from 0 to size(A[k-1])-1 do
6:       spurVertex=A[k-1].vertex(i)
7:       rootPath=A[k-1].vertices(0,i)
8:       for p in A do
9:         if rootPath==p.vertices(0,i) then
10:          remove p.edge(i,i+1) from graph
11:        if spurVertex is a Reaction vertex then
12:          spurPath=ShortestPath(spurVertex,sink,graph,A[k].vertex(i-
13:          1))
14:        else
15:          spurPath=ShortestPath(spurVertex,sink,graph)
16:          totalPath=rootPath+spurPath
17:          B.append(totalPath)
18:          restore edges to graph
19:        B.sort()
20:        A[k]=B[0]
21:        B.pop()
22:   return A

```

---

to calculate a shortest path with a deviation from the root path. When it calls the shortest path method with a reaction vertex as the source it passes an extra parameter which is the previous vertex of the source in the root path.

All algorithms were implemented in python.

# CHAPTER 4

## RESULTS

### 4.1 Centrality Measures

As described in the Material and Methods chapter, centrality measures were studied for their ability to distinguish between currency and non-currency metabolites. The Figures 4.1 4.2 4.3 are a series of histogram plots for the different centrality measures tested. For the betweenness and closeness centrality a list of currency compounds (ATP, ADP, NAD, NADH, NADH, NADP, NADPH,  $H^+$ ,  $H_2O$ ,  $P_i$ ,  $PP_i$ , CMP,  $CO_2$ ,  $O_2$ ) was used. From the figure it is evident that the centrality measures separate compounds into two classes to some extent. However, the boundary between the classes is not distinct. On inspection of the compounds in the currency metabolite group it turns out that some non currency metabolites are also classified as currency metabolites. Some compounds central to metabolism are misclassified in this scenario. Another reason for errors is that centrality measures do not account for chemical context. The same metabolite could act as a currency metabolite in one reaction and act as a non-currency metabolite in another reaction. For example, ATP is a currency metabolite in most reactions but is a main metabolite in nucleotide synthesis reactions.

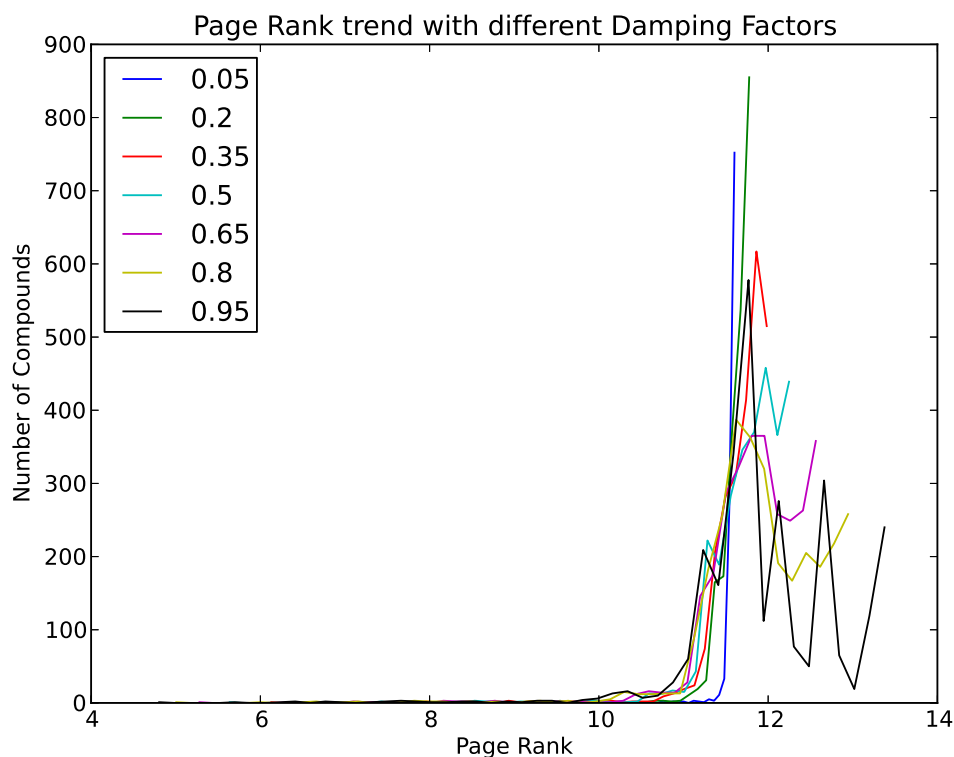


Figure 4.1: Pagerank distribution for different damping factors

## 4.2 Pathfinding

To optimize a cutoff value for the similarity score, the similarity score for all RPAIRS was calculated and the distribution of the similarity scores for different categories of RPAIRS has been plotted in Figure 4.4. A good cutoff value will maximize the number of main,trans and ligase RPAIRS while minimizing the number of leave, cofac RPAIRS. On careful inspection of the distributions, 0.3 was selected as a cutoff. All compound pairs with similarity score less than 0.3 are considered as biochemically irrelevant and compound pairs with a similarity score greater than 0.3 are considered as biochemically relevant. Another possible strategy to use similarity score is to weigh compound-reaction-compound edges in the metabolic graph using the similarity score. But, it is misguided because many transferases transfer large

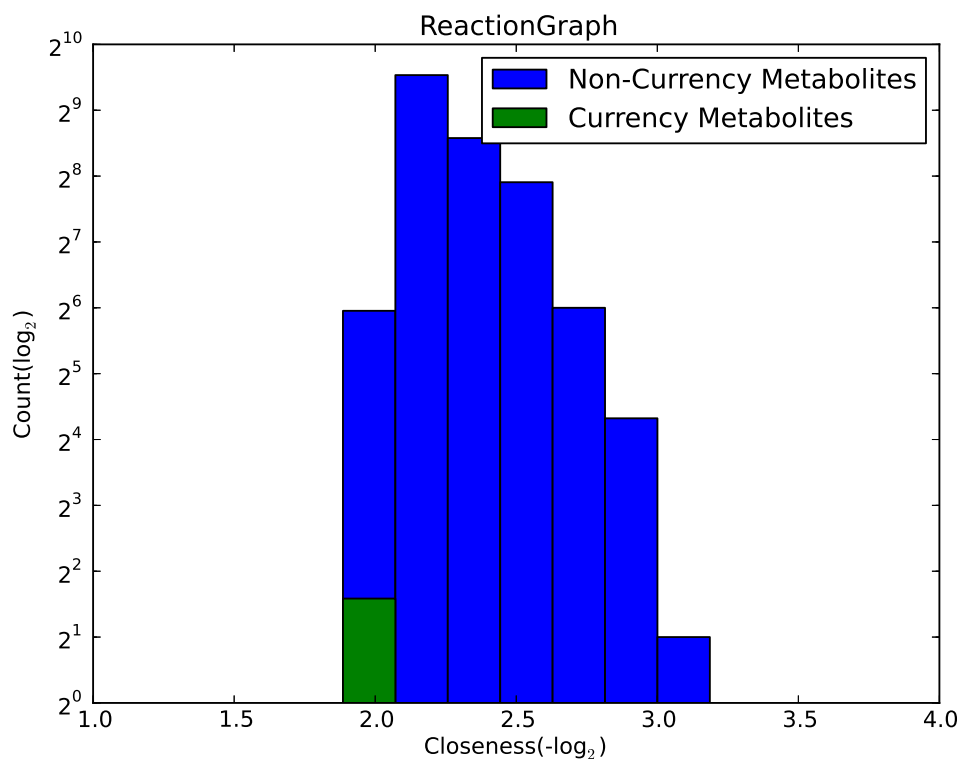


Figure 4.2: Closeness distribution.

chunks of a substrate onto the product and so the similarity scores for those compound pairs are close to 0.5 which is quite low compared to the scores of many main RPAIRS.

To test the path finding algorithm we compared our algorithm against [41] which relies solely on RPAIR annotation. This method is one of the most accurate tools available to the best of our knowledge. It also is the the only tool designed to accept any set of reactions and is not limited to the reactions of just one organism.

We implemented the Reaction Specific Rpair Graph from that paper because it is the more precise than the Rpair Graph since it provides reaction information too. We then compared the our method to it under 4 settings

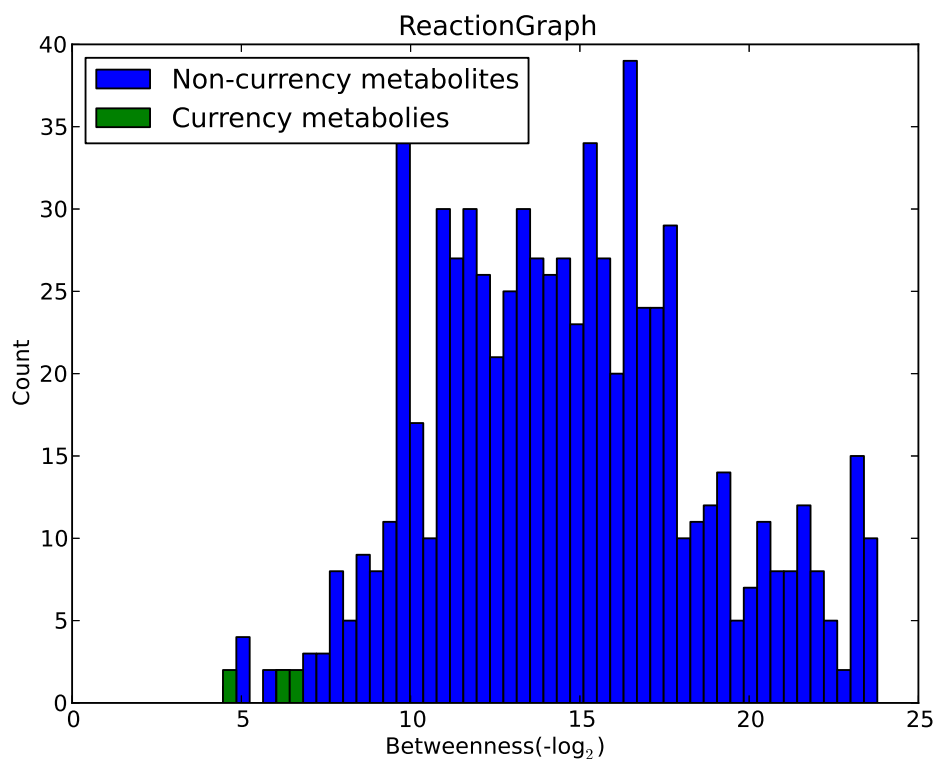


Figure 4.3: Betweenness distribution.

(Table 4.1). Our results for the list of annotated pathways in Table 3.1 exactly matched the Reaction Specific Rpair graph method run under the same settings. This is expected since the test pathways are all well studied and annotated with complete RPAIR data.

<b>RPAIR Filtering</b>	<b>Edge Weight</b>
Main	Unit
Main	Degree
Main-Trans	Unit
Main-Trans	Degree

Table 4.1

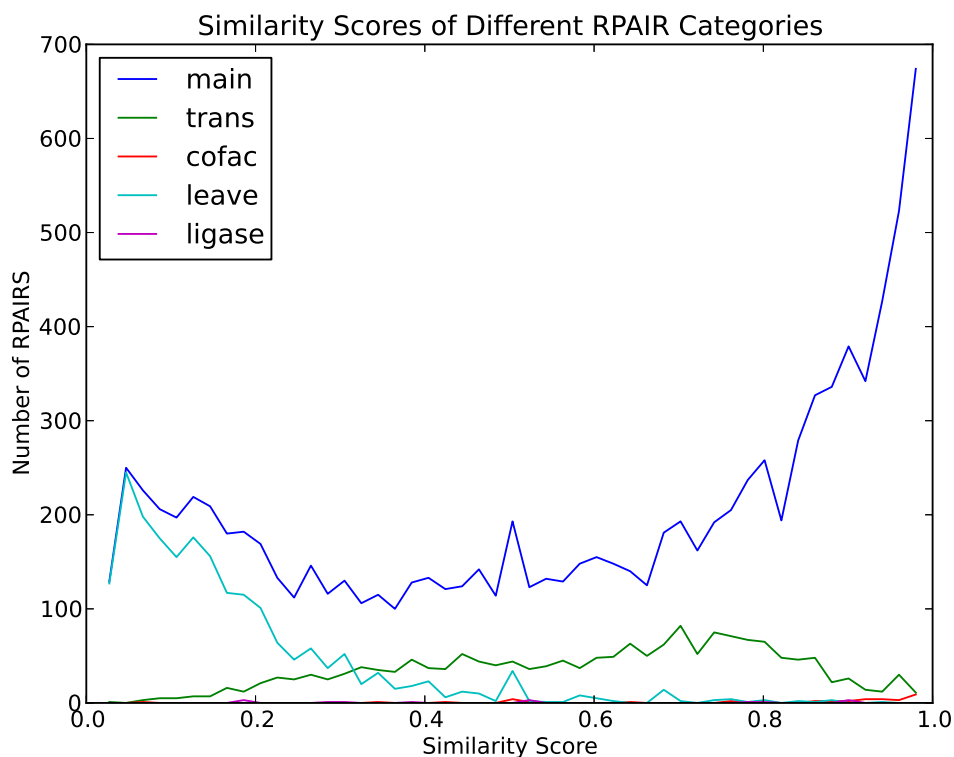



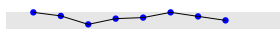
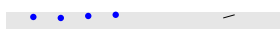
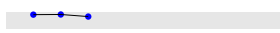




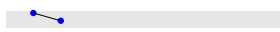

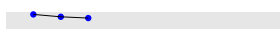
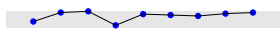
Figure 4.4: Similarity score distribution for all RPAIR categories

To, illustrate the efficacy of the similarity score we calculated the similarity score of each pair of compounds in a pathway not in our annotated pathway list and from an organism different than *E.coli*. We here look at the lysine biosynthesis pathway in *M.tuberculosis* H37Rv stored in MetaCyc. This pathway has reactions involving 5 of the 6 enzyme classes include a lyase reaction adding a pyruvate, a relatively smaller compound to L-aspartate-semialdehyde, a relatively larger compound in comparison . All compound pairs have a similarity score higher than the cutoff. So, even in case there was no RPAIR data available this pathway was successfully found.

### 4.3 Robustness

The RPAIR method depends upon the RPAIR annotation find pathways. In fact, if an RPAIR is missing, its corresponding reactions are not present in the metabolic network, potentially severely affecting its performance. Other methods discussed in chapter 2 are also similarly dependent on the annotation of reactions for accuracy. We have performed the analysis done on the lysine biosynthesis pathway on all annotated pathways in Figure 3.1 to check for robustness of our method vis-à-vis the RPAIR method. The results are in Table 4.2.

Table 4.2: Similarity Score for all reactions in pathways from Table 3.1.








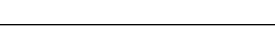

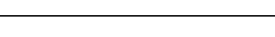
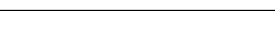







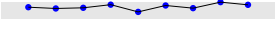

Reference Pathway	Reaction Scores
Gluconeogenesis	
Glycolysis	
Proline Biosynthesis	
Ketoglutarate Metabolism	
Pentose Phosphate Pathway	
TCA Cycle	
NAD Biosynthesis	
Arginine Biosynthesis	
Spemidine Biosynthesis	
Threonine Degradation	
Serine Biosynthesis	
Histidine Biosynthesis	

\*

—Continued On Next Page—

\*

Table 4.2: Similarity Score for all reactions in pathways from Table 3.1.

Reference Pathway	Reaction Scores
Tyrosine Biosynthesis	
CoenzymeA Biosynthesis	
Pentothenate Biosynthesis	
Tetrahydrofolate Biosynthesis	
Flavin Biosynthesis	
Heme Biosynthesis	
Pyrimidine Ribonucleotide Synthesis	
Pyrimidine Deoxy Ribonucleotide Synthesis	
Rhamnose Degradation	
Fucose Degradation	
Entner Duodoroff Pathway	
Anearobic Respiration	
Arginine Biosynthesis	
Proline Degradation	
Glycolate Degradation	
Glycerol Degradation	
Phenylalanine Biosynthesis	
Allantoin Degradation	
Cysteine Biosynthesis	
Lysine Biosynthesis	



Only two reactions in TCA cycle fall below the cutoff. These reactions transform:  $2\text{-oxoglutarate} \rightarrow \text{succinyl-CoA}$   $\text{succinyl-CoA} \rightarrow \text{succinate}$ . The similarity score for these reactions are below the cutoff because of the transfer of CoA, which is a large compound. Very few, if any reactions will be of this type because CoA is a coenzyme and hence much larger than typical secondary metabolites.

# CHAPTER 5

## SUMMARY AND FUTURE WORK

I have presented here a relaxed easy to compute criterion to predict reactant pairs. We have also implemented a graph model that utilizes this criterion as well as KEGG RPAIR data for metabolic pathfinding, for which we implemented a modified  $k$ -shortest path algorithm adapted to the graph model. On the test set of annotated pathways we were able to show that a pathfinding method using just our similarity score performed as well as the method based on the KEGG RPAIR data. One key insight obtained is that for a large number of metabolic pathways, compound transformations maintain significant amount of atomic content between the substrate and product.

The work done here can be extended to incorporate stoichiometry and develop a truly automatic pathfinding tool using both stoichiometry and pathfinding. Incorporating stoichiometric information can help elucidate an organisms preference for alternative pathways under different metabolic conditions. Currently, the only method using both pathfinding and stoichiometry requires manual specification of acceptable transformations and is restrictive in its definition of acceptable transformations to significant carbon exchange [61]. My similarity score can be computed automatically and is a more relaxed criterion.

## REFERENCES

- [1] J. E. Bailey, "Toward a science of metabolic engineering." *Science (New York, N. Y.)*, vol. 252, no. 5013, pp. 1668–75, June 1991.
- [2] T. S. Moon, S.-H. Yoon, A. M. Lanza, J. D. Roy-Mayhew, and K. L. J. Prather, "Production of glucaric acid from a synthetic pathway in recombinant *Escherichia coli*." *Applied and environmental microbiology*, vol. 75, no. 3, pp. 589–95, Feb. 2009.
- [3] 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." *Nature biotechnology*, vol. 21, no. 7, pp. 796–802, July 2003.
- [4] C. E. Nakamura and G. M. Whited, "Metabolic engineering for the microbial production of 1,3-propanediol." *Current opinion in biotechnology*, vol. 14, no. 5, pp. 454–9, Oct. 2003.
- [5] W. Niu, M. N. Molefe, and J. W. Frost, "Microbial synthesis of the energetic material precursor 1,2,4-butanetriol." *Journal of the American Chemical Society*, vol. 125, no. 43, pp. 12998–9, Oct. 2003.
- [6] L. Stryer, *Biochemistry*, 4th ed. W. H Freeman, 1995.
- [7] F. J. Planes and J. E. Beasley, "A critical examination of stoichiometric and path-finding approaches to metabolic pathways." *Briefings in bioinformatics*, vol. 9, no. 5, pp. 422–36, Sep. 2008.
- [8] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. a. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. a. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic acids research*, vol. 38, no. Database issue, pp. D473–9, Jan. 2010.
- [9] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, S. M. Paley, and A. Pellegrini-Toole, "The EcoCyc and MetaCyc databases." *Nucleic acids research*, vol. 28, no. 1, pp. 56–9, Jan. 2000.

- [10] M. Kanehisa, “A database for post-genome analysis.” *Trends in genetics : TIG*, vol. 13, no. 9, pp. 375–6, Sep. 1997.
- [11] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic acids research*, vol. 27, no. 1, pp. 29–34, Jan. 1999.
- [12] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. D. Karp, “A systematic comparison of the MetaCyc and KEGG pathway databases.” *BMC bioinformatics*, vol. 14, p. 112, Jan. 2013.
- [13] a. Bairoch, “The ENZYME database in 2000.” *Nucleic acids research*, vol. 28, no. 1, pp. 304–5, Jan. 2000.
- [14] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, “Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.” *Journal of the American Chemical Society*, vol. 125, no. 39, pp. 11 853–65, Oct. 2003.
- [15] M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa, “Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions.” *Journal of the American Chemical Society*, vol. 126, no. 50, pp. 16 487–98, Dec. 2004.
- [16] B. Junker and F. Schreiber, *Analysis of biological networks*, 2008.
- [17] B. A. Albert R. and J. H., “Power law distribution of the world wide web,” *Science*, vol. 287, p. 2115a, 2000.
- [18] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, “The large-scale organization of metabolic networks.” *Nature*, vol. 407, no. 6804, pp. 651–4, Oct. 2000.
- [19] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks.” *Nature*, vol. 393, no. 6684, pp. 440–2, June 1998.
- [20] H. Ma and A.-P. Zeng, “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms,” *Bioinformatics*, vol. 19, no. 2, pp. 270–277, Jan. 2003.
- [21] M. Huss and P. Holme, “Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks.” *IET systems biology*, vol. 1, no. 5, pp. 280–5, Sep. 2007.
- [22] R. Kuffner, R. Zimmer, and T. Lengauer, “Pathway analysis in metabolic databases via differential metabolic display (DMD),” *Bioinformatics*, vol. 16, no. 9, pp. 825–836, Sep. 2000.

- [23] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks.” *Bioinformatics (Oxford, England)*, vol. 21, no. 8, pp. 1603–9, Apr. 2005.
- [24] A. Seressiotis and J. E. Bailey, “MPS: An algorithm and data base for metabolic pathway synthesis,” *Biotechnology Letters*, vol. 8, no. 12, pp. 837–842, Dec. 1986.
- [25] A. Seressiotis and J. E. Bailey, “MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways.” *Biotechnology and bioengineering*, vol. 31, no. 6, pp. 587–602, Apr. 1988.
- [26] M. L. Mavrovouniotis and G. Stephanopoulos, “Computer-aided synthesis of biochemical pathways.” *Biotechnology and bioengineering*, vol. 36, no. 11, pp. 1119–32, Dec. 1990.
- [27] M. L. Mavrovouniotis and G. Stephanopoulos, “Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 1. A synthesis approach in the context of simple examples,” *Industrial & Engineering Chemistry Research*, vol. 31, no. 7, pp. 1625–1637, July 1992.
- [28] M. L. Mavrovouniotis, “Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 2. Formalization and analysis of the synthesis algorithm,” *Industrial & Engineering Chemistry Research*, vol. 31, no. 7, pp. 1637–1653, July 1992.
- [29] M. M.L., “Identification of qualitatively feasible metabolic pathways,” in *Artificial Intelligence and Molecular Biology*, H. L., Ed. AAAIPress/MIT Press, 1993, pp. 325–364.
- [30] I. Zevedei-Oancea and S. Schuster, “Topological analysis of metabolic networks based on Petri net theory.” *In silico biology*, vol. 3, no. 3, pp. 323–45, Jan. 2003.
- [31] K. Voss, M. Heiner, and I. Koch, “Steady state analysis of metabolic pathways using Petri nets.” *In silico biology*, vol. 3, no. 3, pp. 367–87, Jan. 2003.
- [32] I. Koch, B. H. Junker, and M. Heiner, “Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber.” *Bioinformatics (Oxford, England)*, vol. 21, no. 7, pp. 1219–26, Apr. 2005.
- [33] E. Simão, E. Remy, D. Thieffry, and C. Chaouiya, “Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E.coli.” *Bioinformatics (Oxford, England)*, vol. 21 Suppl 2, no. suppl\_2, pp. ii190–6, Sep. 2005.

- [34] D. McShan, S. Rao, and I. Shah, “PathMiner: predicting metabolic pathways by heuristic search,” *Bioinformatics*, vol. 19, no. 13, pp. 1692–1698, Sep. 2003.
- [35] J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, and S. Wodak, “Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data,” *Lecture Notes Comput. Sci.*, vol. 2066, pp. 155–172, 2001.
- [36] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, “Hierarchical organization of modularity in metabolic networks.” *Science (New York, N.Y.)*, vol. 297, no. 5586, pp. 1551–5, Aug. 2002.
- [37] D. A. Fell and A. Wagner, “The small world of metabolism.” *Nature biotechnology*, vol. 18, no. 11, pp. 1121–2, Nov. 2000.
- [38] J. van Helden, L. Wernisch, D. Gilbert, and S. J. Wodak, “Graph-based analysis of metabolic networks.” *Ernst Schering Research Foundation workshop*, no. 38, pp. 245–74, Jan. 2002.
- [39] D. Croes, F. Couche, S. J. Wodak, and J. van Helden, “Inferring meaningful pathways in weighted metabolic networks.” *Journal of molecular biology*, vol. 356, no. 1, pp. 222–36, Feb. 2006.
- [40] S. a. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg, “Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC).” *Bioinformatics (Oxford, England)*, vol. 21, no. 7, pp. 1189–93, Apr. 2005.
- [41] K. Faust, D. Croes, and J. van Helden, “Metabolic pathfinding using RPAIR annotation.” *Journal of molecular biology*, vol. 388, no. 2, pp. 390–414, May 2009.
- [42] E. Pitkänen, P. Jouhten, and J. Rousu, “Inferring branching pathways in genome-scale metabolic networks.” *BMC systems biology*, vol. 3, p. 103, Jan. 2009.
- [43] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, “The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics.” *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 493–500, 2003.
- [44] J. Willet, P. an Barnad and G. Downs, “Chemical similarity searching.” *J. Chem. Inf. Comput. Sci.*, vol. 38, pp. 938–996, 1998.
- [45] F. Planes and J. Beasley, “Path finding approaches and metabolic pathways,” *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2244–2256, May 2009.

- [46] M. Arita, “In Silico Atomic Tracing by Substrate Product Relationships in Escherichia coli Intermediary Metabolism,” *Genome Research*, vol. 13, no. 11, pp. 2455–2466, 2003.
- [47] A. P. Heath, G. N. Bennett, and L. E. Kavvaki, “Finding metabolic pathways using atom tracking.” *Bioinformatics (Oxford, England)*, vol. 26, no. 12, pp. 1548–55, June 2010.
- [48] T. Blum and O. Kohlbacher, “Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks.” *Journal of computational biology : a journal of computational molecular cell biology*, vol. 15, no. 6, pp. 565–76.
- [49] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, and M. Kanehisa, “PathPred: an enzyme-catalyzed metabolic pathway prediction server.” *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W138–43, July 2010.
- [50] M. Hattori, N. Tanaka, M. Kanehisa, and S. Goto, “SIM-COMP/SUBCOMP: chemical structure search servers for network analyses.” *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W652–6, July 2010.
- [51] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes.” *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, Jan. 2000.
- [52] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, and M. Kanehisa, “Ezyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs.” *Bioinformatics (Oxford, England)*, vol. 25, no. 12, pp. i179–86, June 2009.
- [53] J. E. B. Francisco J. Planes, “An optimization model for metabolic pathways.” *Bioinformatics*, vol. 25, pp. 2723 – 2729, 2009.
- [54] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp, “EcoCyc: a comprehensive database of Escherichia coli biology.” *Nucleic acids research*, vol. 39, no. Database issue, pp. D583–90, Jan. 2011.
- [55] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using {NetworkX},” in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, Aug. 2008, pp. 11–15.

- [56] G. Sabidussi, "The centrality of a graph." *Psychometrika*, vol. 31, no. 4, pp. 581–603, Dec. 1966.
- [57] A. Bavelas, "Communication Patterns in Task-Oriented Groups," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, p. 725, June 1950.
- [58] U. Brandes, "A Faster Algorithm for Betweenness Centrality ," *Journal of Mathematical Sociology*, vol. 25, no. 1994, pp. 163–177, 2001.
- [59] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1998.
- [60] J. Y. Yen, "FINDING THE K SHORTEST LOOPLESS PATHS IN A NETWORK. ," *Management Science*. Jul1971, vol. 17, no. 11.
- [61] J. Pey, J. Prada, J. E. Beasley, and F. J. Planes, "Path finding methods accounting for stoichiometry in metabolic networks." *Genome biology*, vol. 12, no. 5, p. R49, Jan. 2011.