Extracting and Ranking Metabolic Pathways from Large Metabolic Networks using Graph Theory and Metabolic Flux Analysis

Konstantinos Mexisa, Stefanos Xeniosa, Nikos Trokanasa, Antonis Kokossisa\*

*aDepartment of Process Engineering, NTUA, Iroon Politechniou 6 Zografou, Athens,* 157 80, Greece

\*Corresponding author: akokossis@mail.ntua.gr

Abstract

Efficiently deriving biosynthetic pathways from complex biochemical networks is crucial for advancing metabolic engineering. However, efficient analysis and navigation of big biochemical networks remain a challenge. Moreover, ranking the constructed pathways introduces further complexities. We propose a novel graph-based method for extracting biologically relevant metabolic pathways within large metabolic networks. The graph links metabolites as nodes via edges, representing reactant-product connections. Edges are assigned weights defining relations between two connected nodes. By incorporating further constraints based on compounds’ molecular similarity, network connectivity and the availability of metabolites in the selected microorganism, we facilitate the rapid and reliable extraction of feasible metabolic pathways from large biochemical networks. Our method not only identifies pathways as a series of metabolites from source to target, but also delineates the specific enzymatic reactions for each step of the heterologous pathway. To illustrate the effectiveness of our approach, we provide a case study involving the bioproduction of butanol from E. coli.

**Keywords**: Metabolic Pathways, Pathway Discovery, Bioproduction, Retrobiosynthesis

* 1. Introduction

Metabolic pathways play a fundamental role in various applications, guiding the intricate conversion of source molecules into target molecules through a series of consecutive reactions. Extracting biologically relevant metabolic pathways from large biochemical networks which encompass thousands of metabolites interconnected by tens of thousands of enzymatic reactions remains a substantial challenge. Additionally, ranking the constructed pathways introduces further complexities. Historically, these pathways were crafted by hand, derived directly from experimental evidence. However, the omics era and the proliferation of computational resources have revolutionized the landscape of biochemistry research. The continuous expansion of biochemical databases not only supports fundamental research and metabolic engineering but also enables the design of non-canonical pathways not found in nature. While intuition and manual design once dominated non-natural pathway creation, the wealth of available biochemical data suggests that more efficient alternatives may be overlooked. Consequently, the development of systematic and automated pathway extraction tools has become essential to meet this challenge. These computational tools are designed to extract metabolic pathways from biochemical databases, with the primary goal of ensuring the pathways extracted are biologically relevant. They differ in how they build and search the available biochemical space, weight individual steps and complete pathways (Haffner et al., 2021). In this context, we tackle the challenge of efficiently exploring and analyzing extensive biochemical networks. We introduce a graph-based approach designed to prioritize the search for pathways featuring similar molecules in terms of molecular weight and molecular similarity, leveraging the SMILES representation. Notably, our method excels not only in extracting biologically meaningful pathways from these networks, but also in identifying the specific enzymatic reactions required for integration into a chosen host microorganism, enabling the production of a desired target molecule.

* 1. Pilot study: Butanol production form *Escherichia Coli*

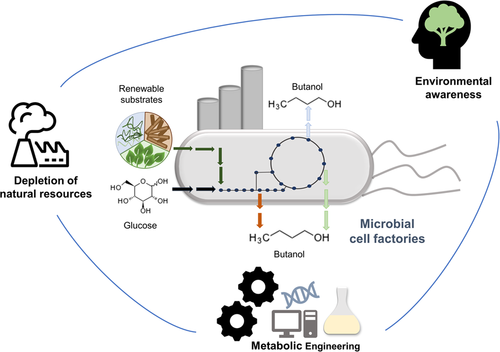
The global butanol is on the rise, owing to its expanding applications and role as a chemical precursor of several other compounds. Recently, *n-butanol*, with its superior biofuel characteristics compared to *ethanol*, have stimulated even more interest. Clostridium species naturally produce (bio)butanol. To enhance butanol production, various pathways have been transplanted into the more accessible host, *Escherichia coli* (Atsumi, et al.)*.* In our pilot study, we utilized a Genome-Scale model for the core metabolism of *E. coli.* Our method was tested on a vast biochemical network, comprising 30,081 enzymatic reactions sourced from the KEGG database. Our aim was to highlight our method’s proficiency in extracting biologically relevant pathways from complex networks and its support for efficient navigation using our proposed network structure. Through the application of our method, we successfully rediscovered a previously experimentally validated novel pathway for biobutanol production in E. coli. Furthermore, our approach uncovered additional pathways by initiating the search with metabolites from E. coli’s metabolism, showcasing the robustness and effectiveness of our method in identifying pathways of interest withing the metabolic network.

Figure 1: Microbial cell factories can convert renewable substrates into target metabolites replacing chemical-based processes

* 1. Materials and Methods
     1. Biochemistry: Finding a retrosynthetic route through metabolism

In metabolic engineering, retrosynthesis aims to identify enzymatic reaction paths linking a target molecule to a cellular precursor. This process differs from organic chemistry retrosynthesis as it operates within a concurrent, system-level biochemical context, often involving heterologous pathways with shared reaction conditions and enzyme promiscuity. These characteristics create interconnected networks with unknown reactions and multiple potential carbon flow paths, affecting product yield, a critical economic metric (Geng-Min Lin, et al.).

* + 1. Metabolic Network construction

We constructed an undirected graph using reactions sourced from the KEGG database, where each enzymatic reaction was decomposed into pairs of reactants and products. This graph links metabolites as nodes via edges, representing reactant-product connections. Edges are assigned a weight that defines the relation between two connected nodes. Out of the 30,081 enzymatic reactions sourced from the KEGG database, we extracted 55,979 substrate-product pairs. Among these pairs, 10,747 had associated information from KEGG RPAIR. Within this subset, 5,148 pairs were categorized as *'main'* KEGG RPAIRs. A *'main'* KEGG RPAIR signifies the primary biotransformation occurring in a specific enzymatic reaction.

* + - 1. RPAIR Prediction using Machine Learning

To complete reactant-product pairs with missing RPAIR data, we implemented a machine learning (ML) pipeline incorporating the SMILES representation of the reactants and products as inputs. SMILES strings offer a linear representation of molecules and can be further transformed into molecular fingerprints, which serve as mathematical representations of molecules. In our study, we utilized the Morgan Fingerprint, a fixed-length binary vector representing molecular structure. Both reactants and products in each pair were converted into binary vectors of equal length. These vectors were then combined into a unified vector for each pair, serving as input for the ML model. The model's output determined whether the pair was a *main* pair or not. We divided pairs with known RPAIR data into an 80:20 training-validation split and used these sets to train a *Random Forest Classifier*, which was subsequently employed to predict missing RPAIR values for the entire dataset.

A blue arrows pointing to a black background

Description automatically generated

Figure 2: ML pipeline to predict missing RPAIR data

* + 1. Calculation of weighted reactant-product pairs

In our approach, edges in the graph are assigned weights that define the relationship between the connected nodes. To enable the utilization of state-of-the-art shortest-path graph search algorithms, reactants of similar molecular weight should be close to each other, while pairs with significantly different weights are distanced further apart. This arrangement is based on the atom-conserving nature of reactants with similar molecular weights, enhancing the likelihood of reactions, a key aspect in pathway analysis optimization. Shorter paths typically represent more biologically feasible and efficient pathways. As a retrosynthetic path gets longer, each added enzyme is likely to decrease the amount of product made from the precursor. To achieve this, we transformed the difference in molecular weights into a distance metric:

where and are the molecular weight of the substrate and product of a reaction pair respectively.

* + 1. Pruning the Metabolic Network
       1. Pruning for core pathway focus

The general picture of a metabolic network is that of a highly connected complex network of metabolites and reactions, sharing many features with other large-scale biological networks. To distill the network into biologically meaningful pathways, we employed a pruning strategy by retaining solely the *main* reactant-product pairs. The removal of ­*non-main* pairs led to a significantly reduces, more focused network, concentrated on the core biotransformations. The pruned network comprises 7,997 nodes and 11,783 edges.

* + - 1. Currency metabolites

Metabolic networks feature metabolites like ATP and NAD, commonly known as “currency” or “hub” metabolites, as they participate in numerous reactions. These hub metabolites, while essential for the metabolism, present a challenge in metabolic network analysis. Their prevalence in the network often obscures the underlying pathway-like structure. To clarify the network’s actual properties, it is common to exclude these metabolites during analysis. However, no consensus exists on what constitutes a currency metabolite. In our study, we opted to remove 209 well-known hub metabolites (e.g., H2O, ATP, NAD) from the network.

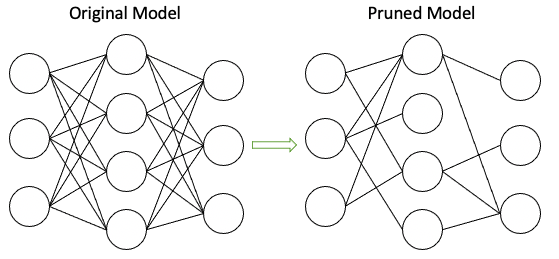


Figure 3: Metabolic network pruning

* + 1. Pathway refinement via molecular similarity

In addition to employing a weighted-shortest path approach with weights determined by the molecular weights of reactant-product pairs, we incorporated an additional constraint to refine the resulting metabolic pathways. This constraint involved evaluating the molecular similarity of every reactant-product intermediate pair within the metabolic pathway. We calculated molecular similarity using Fingerprint Similarity, extracted from the SMILES representation of the molecules. In our analysis pipeline, we took a step further by sorting the top-10 weighted shortest paths based on the average similarity between these intermediate metabolites. This step allowed us to prioritize pathways that not only exhibited efficient weight-based connections but also demonstrated a higher degree of molecular similarity among the metabolites involved, thus enhancing the biological relevance of the identified metabolic pathways.

A blue arrows pointing to a black background

Description automatically generated

Figure 4: Molecular similarity calculation

* + 1. Selecting specific enzymatic reactions

To successfully integrate a heterologous pathway into a chosen microorganism, it is imperative to specify not only the precursor metabolites starting from a given molecule but also the precise enzymatic reaction to introduce to the host cell. Complicating matters, a single reactant-product pair extracted from a large metabolic network, can belong to multiple enzymatic reactions, necessitating the careful selection of the most suitable reaction for incorporation. To address this challenge, we examine the metabolites available within the selected microorganism. Subsequently, we focus on selecting reactions that exclusively utilize the already present metabolites to generate a new metabolite. This new metabolite then serves as the starting point for subsequent reactions, ultimately culminating in the synthesis of the desired target molecule. This strategic approach ensures the compatibility of the introduced pathway with the host microorganism and paves the way for successful heterologous pathway integration.

* + 1. Finding metabolic pathways with graph search

Our method employs Yen's k-shortest loop-less path algorithm, utilizing the Python package *NetworkX*, to extract the shortest pathways from the weighted network of reactant-product pairs (Haffner et al., 2021). This algorithm takes as input a weighted graph representing the network, a source compound, a target compound, and a specified maximum number of shortest paths (k) to be identified. After identifying the specified number of k-shortest paths, the algorithm ceases its search and makes these pathways available for analysis. The algorithm then ranks these k-shortest paths according to the average molecular similarity at each reaction step. This ranking prioritizes pathways with the highest overall similarity, ensuring the selection of metabolic routes that are both biologically relevant and chemically consistent for our analysis.

* 1. Results
     1. Extracting biologically relevant pathways

To demonstrate the effectiveness of our approach, we present two illustrative pathway searches. In the first instance, our goal was to establish a biochemical connection between *tyrosine* and *caffeate*. The resulting pathway, with a length of just two steps, indicates its high quality, as it involves only a minimal number of intermediate reactions. Furthermore, the pathway we extracted aligns with KEGG's phenylpropanoid biosynthesis map, underscoring its biological relevance and validity.

A chemical structure of a molecule

Description automatically generated

Figure 5: Extracted pathway connecting tyrosine and caffeate

In the second example, we searched for a pathway connecting the compounds *tyrosine* and *syringin*. The extracted pathway is part of the KEGG pathway map for phenylpropanoid biosynthesis, and it can therefore be called a confirmed, biologically meaningful pathway. These two examples of pathway search problems illustrate the capacity of our method to efficiently extract biologically relevant pathways from large biochemical networks. The algorithm robustly handled searches for long pathways of eight and more biotransformation steps, as they are usually present in secondary metabolism.

A diagram of chemical formulas

Description automatically generated

Figure 6: Extracted pathway connecting tyrosine and syringin

* + 1. Pathway discovery and host integration

A diagram of a chemical reaction

Description automatically generatedTo demonstrate our approach's efficacy, we conducted a case study on butanol bioproduction in E. coli. E. coli, a well-studied microorganism with versatile genetic tools, lacks the natural capacity for 1-butanol production. Our method rediscovered a six-step metabolic pathway, starting from acetyl-CoA, to enable 1-butanol production in this host organism. To construct our host organism representation, we utilized the *E. coli* core metabolism Genome-scale metabolic model, from which we extracted the metabolites assessable for participation in the enzymatic reactions required for the heterologous metabolic pathway. Employing the methodology outlined in *Section 3*, our algorithm successfully determined the precise enzymatic reactions to incorporate into the host organism, enabling the production of 1-butanol. To assess the biological significance of the extracted pathway, we integrated it into the host organism’s genome-scale model and conducted Flux Balance Analysis (FBA). Within this analysis, we imposed anaerobic conditions and we set the biomass production as the optimization function. The FBA algorithm subsequently calculated the flux for the biomass reaction, indicating the pathway’s biological relevance and feasibility within the host organism.

Figure 7: Schematic representation of 1-butanol production in engineered E. coli

* 1. Conclusions

In this study, we introduced a novel graph-based method for the efficient extraction of biologically relevant metabolic pathways from large biochemical networks. Our approach combines network analysis, molecular similarity constraints, machine learning and state-of-the-art pathway search algorithms to facilitate the discovery of pathways with practical applications in metabolic engineering. We demonstrated the effectiveness of our method through a case study focusing on 1-butanol bioproduction in E. coli, highlighting its potential to develop valuable bioproduction processes in microorganisms. Our approach streamlines pathway discovery, enhances the likelihood of success in practical bioproduction processes, and offers a valuable asset for metabolic engineers, bridging the gap between theory and application in the field of metabolic engineering.

References

Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., Smith, K. M., Brynildsen, M. P., Chou, K., Hanai, T., & Liao, J. C. (2008). Metabolic engineering of Escherichia coli for 1-butanol production. Metabolic Engineering, 10(6), 305–311

Jasmin Hafner, Vassily Hatzimanikatis, NICEpath: Finding metabolic pathways in large networks through atom-conserving substrate–product pairs, Bioinformatics, Volume 37, Issue 20, October 2021, Pages 3560–3568

Geng-Min Lin, Robert Warden-Rothman, Christopher A. Voigt, Retrosynthetic design of metabolic pathways to chemicals not found in nature, Current Opinion in Systems Biology, Volume 14, 2019, Pages 82-107

**Acknowledgments**

The research project was supported by the Hellenic Foundation for Research and  
Innovation (H.F.R.I.) under the 2nd Call for H.F.R.I. Research Projects to support  
Faculty Members & Researchers (Project Number: 3817).