Unpacking the Metabolic Sandwich: A Comprehensive Exploration of Biochemical Pathways, Enzyme Structures, and Evolutionary Relationships

Biswajit Banerjee[1], D. Eric Smith [1,2,3] , Loren Dean Williams [1,2] , Anton S. Petrov [1,2]
[1] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332
[2] School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332
[3] Earth-Life Science Institute, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan, 〒152-8550

Introduction

The evolution of metabolic reactions and the enzymatic folds that catalyze these reactions in extant biology is a fundamental aspect of understanding life's biochemical and molecular intricacies. This project aims to reveal the evolutionary relationship between these two spaces (namely metabolic and enzymatic space), focusing on the most ancient and conserved parts of the metabolic network and the oldest protein folds. It also seeks to identify correlated evolutionary events between the early expansion of metabolic networks and the evolution of protein folds during the advent of life. To address these questions, computational tools are being developed to access, organize, and analyze metabolic and molecular data, utilizing resources from KEGG, PDB, ECOD, and M-CSA databases. These tools will enable the construction of customizable networks, consolidation of inter-database linkages, and support user-defined categorizations. The project demonstrates the practical application of these tools through the reconstruction of stoichiometric paths from prebiotic inputs to metabolites, using network expansion models proposed by Goldford et al. By mapping reaction orderings onto enzyme sequences and catalytic-domain folds, the project explores the alignment between enzyme diversification and metabolic network expansion.

Challenges

For a given enzyme, it is challenging to unambiguously find a 3D structure due to existence of a) incomplete structures; b) multiple structures; c) structures that were obtained from different closely related species; d) structures that were obtained by introducing mutations into the wild type sequences. Some of which can be incomplete or in different resolution of X-ray crystallography.
Creating a one-to-one map can be challenging as multiple biological/chemical databases have different ids that are limited to the database only and different databases do not agree with each other in some cases. Accurately mapping enzyme commission (EC) numbers with 3D structures and catalytic domains is complex and poses significant challenges.

Results

The research focuses on mapping two critical biological spaces: metabolic reactions and enzymatic folds. A comprehensive pipeline has been developed that connects metabolic and enzymatic spaces, allowing users to query reactions or compounds to find their associated enzymes and catalytic domains. This research yielded the below results:

Data Processing Pipeline: I created a pipeline (Figure 1) that maps information provided from KEGG, EBI and Uniport to the catalytic domain (ECOD) enzyme involved in catalyzing the reaction. This pipeline allows processing vast amounts of data in real-time with capabilities of integration of multiple databases besides the ones already provided. The service is created entirely in python with the help of Airflow to schedule each user query as a workflow and provide the mapping of 3D structures as result.

Algorithm 1 Find Best Matching 3D Structures for Given Metabolic Reactions

Require: List of metabolic reactions
1: Query KEGG for finding EC numbers corresponding to the metabolic reactions
2: **for** each EC number **do**
3: Query databases such as UniProt, EBI, and others for all matching 3D structures
4: Assign matching 3D structures to a variable `structures`
5: Filter `structures` by species
6: Find the structure in `structures` with the highest number of chains
7: Further filter `structures` by X-ray resolution until only one remains
8: **Output:** The remaining 3D structure is the match for the current EC number
9: **end for**

Mapping Algorithm: At the core of this pipeline lies the mapping algorithm as provided as Algorithm 1. I have created this algorithm where Researchers can provide a pathway or set of different metabolic reactions and this algorithm tries to find the most accurate 3D structure for each of the enzymes amongst the existing one across all databases.

Algorithm 2 Metabolic Network Expansion Using Seed Components

Require: Seed components
1: Initialize `network` with seed components
2: Initialize `generation` with seed components
3: **while** `generation` is not empty **do**
4: Query KEGG reactions to find reactions possible with `generation`
5: List outcomes as `gen_next` compounds
6: Add `gen_next` to `network`
7: Update `generation` to `gen_next`
8: **end while**
9: **Output:** `network` with all reachable compounds from seed components

Network Expansion: I also developed an algorithm for metabolic network expansion that begins with seed components. This algorithm constructs a metabolic network by querying KEGG reactions to identify which reactions can occur with the initial seed compounds. The resulting products of these reactions are classified as generation 1 compounds. The process continues iteratively, using the new generation of compounds along with the seed compounds to find additional reactions and expand the network further. This expansion continues until no further reactions are possible. The output is a comprehensive network of all compounds reachable from the initial seed components, providing a detailed map of metabolic connectivity.

Conclusion

Although I was able to associate the fold to catalytic functionality mapping, the current state of the pipeline can only search the metabolic space, not the fold space. To deepen the understanding of the role of folds in ancient metabolic pathways and their similarities, it is necessary to enable the pipeline to search the fold space. With that said, the current pipeline allows navigation of the metabolic space and its association with the enzymatic space. I have collected data for multiple pathways and associated them with folds. This data will help fuel research aimed at searching from the fold level and mapping fold functionality.