AI607: Graph Mining and Social Network Analysis (Fall 2024)

Term Project: Metabolic Reaction Prediction

Release: September 26, 2024
Progress Report: November 8, 2024, 11:59 pm
Final Report: December 6, 2024, 11:59 pm
Final Presentation: December 9 & 11, 2024

The ultimate goal of this project is to practice data mining research by inferring the outcomes of metabolic reactions. In this project, you will design, implement, and evaluate your approach for predicting whether metabolic reactions proceed in a given direction, identifying a missing metabolite required to complete a reaction, as well as predicting the resulting metabolite sets produced from specific source metabolites. Also, you will (a) write a progress report, (b) write a final report, and (c) present your approach. While details of the following steps will be announced later, tentative schedules are as follows:

- Progress Report - November 8, 2024, 11:59 pm
- Final Report - December 6, 2024, 11:59 pm
- Presentation - December 9 & 11, 2024

This is a team project, and each team should consist of two or three members. You can find your teammates by all means (e.g., Classum), and one progress report should be submitted per team.

Your submission will be evaluated based on

- Presentation (final report & oral presentation) - 40%,
- Novelty of your proposed approach - 20%,
- Validity of your proposed approach - 20%,
- **Accuracy - 20%**.

Note that accuracy is not our only concern. Instead of spending all your time optimizing the accuracy, we recommend spending more time on developing novel and valid approaches and making your presentation clear and complete.

1 Problem: Metabolic Reaction Prediction

1.1 Introduction

Predicting metabolic reactions is important for optimizing drug development, personalizing medicine, engineering microorganisms for biotechnology, and understanding disease mechanisms. This project aims to predict the outcomes of metabolic reactions by leveraging data mining and machine learning techniques applied to existing reaction data.

A metabolic reaction can be represented by source metabolites (reactants) and destination metabolites (products). The source metabolites represent the starting set of metabolites, while the destination metabolites represent the resulting set after the reaction. The sizes of both source and destination metabolite sets may vary. For instance, in Figure 1, we present several types of metabolic reactions where metabolites interact to form new metabolite sets. This demonstrates how two source metabolites can combine to form completely new destination metabolites.

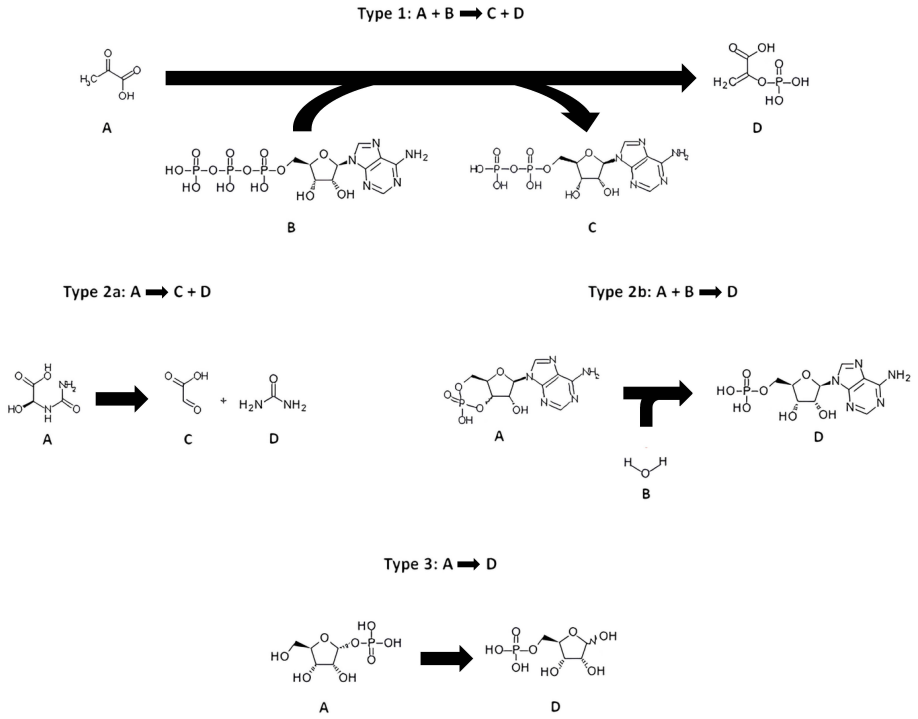

Figure 1: Examples of different types of metabolic reactions. In Type 1, metabolites A and B react to form C and D. Type 2a shows that A is broken down into C and D, while Type 2b demonstrates that A and B are combined to form D. In Type 3, A is directly converted to D without intermediates.

In this project, we will explore several predictive tasks related to metabolic reactions through three key objectives:

- **Reaction Direction Prediction**[1]**:** Predict whether the direction of the reaction is correct or incorrect given a set of source and destination metabolites.
- **Missing metabolite Prediction:** Predict the missing source metabolite in cases where the source metabolite set is incomplete to complete the reaction.
- **Destination Prediction:** Predict the resulting destination metabolite set given a set of source metabolites.

1.2 Task 1: Reaction Direction Prediction

1.2.1 Goal

Determine whether a given metabolic reaction proceeds in the correct direction, which is defined as the transformation of a set of source metabolites into a set of destination metabolites. For each reaction, you will predict whether the reaction proceeds as True (correct direction) or False (incorrect direction).

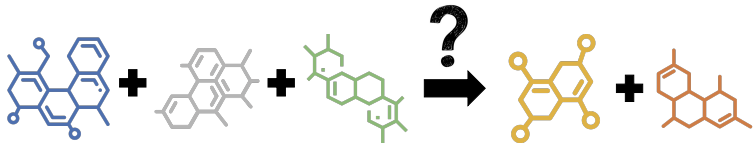

Figure 2: An illustration of Task 1.

1.2.2 Data Description

Each metabolic reaction is represented by two metabolite sets: source metabolites (reactants) and destination metabolites (products). The task is to determine whether the source metabolite set transforms into the destination metabolite set in the correct direction.

You are provided with four data files to complete this task. Each file has a header line containing the column names, and data lines starting from the second line. The structure of the columns is described below, followed by dataset-specific information.

Common Columns All files share the following common columns:

- **reaction_id**: A unique identifier for the reaction.
- **source**: A set of metabolite IDs representing the source metabolites (reactants).
- **destination**: A set of metabolite IDs representing the destination metabolites (products).

[1]The correct direction indicates source $\rightarrow$ destination, while the incorrect direction indicates source $\leftarrow$ destination.

1. Training Data (`Classification_training.csv`) This file contains the full information for training your model, including the reaction identifier, source metabolites, destination metabolites, and the correct direction label. Each row represents one reaction.

- **Additional column:**
  - **direction**: The true or false label indicating whether the reaction proceeds in the correct direction.

Example row:

```
7305,{20,33},{0,486,982},True
```

This indicates that for reaction #7305, the source metabolite set is $\{20, 33\}$, the destination metabolite set is $\{0, 486, 982\}$, and the reaction proceeds in the correct direction (True).

2. Validation Query Data (`Classification_valid_query.csv`) This file is used to validate your model. It contains the reaction identifier, source metabolites, and destination metabolites, but does not include the direction label. You will use this data to predict whether the reaction proceeds in the correct direction and compare your predictions against the correct answers provided in the validation answer file.
Example row:

```
8452,{17,23},{102,305,890}
```

This row represents reaction #8452, where the source metabolite set is $\{17, 23\}$ and the destination metabolite set is $\{102, 305, 890\}$. Your model should predict whether the reaction proceeds in the correct direction for this reaction.

3. Validation Answer Data (`Classification_valid_answer.csv`) This file contains the correct direction labels for the validation query data. After making predictions using the query data, you can compare them with the actual direction labels from this file to assess your model's accuracy.

- **Additional column:**
  - **direction**: The true or false label indicating whether the reaction proceeds in the correct direction.

Example row:

```
8452,True
```

For reaction #8452, the reaction proceeds in the correct direction (True).

4. Test Query Data (`Classification_test_query.csv`) This file contains reactions without the direction labels for testing your model on unseen data. The structure is the same as the validation query data. You will submit your predictions for these reactions.
Example row:

```
9021,{45},{400,789}
```

This row represents reaction #9021, where the source metabolite set is $\{45\}$ and the destination metabolite set is $\{400, 789\}$. Your task is to predict whether the reaction proceeds in the correct direction.

1.2.3 Submission

To complete this task, you need to:

- **Make Predictions:** For each reaction in the test query file, predict whether the direction is correct.
- **Submit Results:**
  - Your submission file should be named `Classification_test_answer.csv`.
  - The file should follow the same format as `Classification_valid_answer.csv`, with the same structure of columns.
  - The columns of the file should be:
    - reaction_id: The unique identifier for the reaction.
    - direction: Your predicted value for the reaction direction (True or False).
  - Ensure that the predicted directions are aligned correctly with the respective reactions in the test query file.

1.2.4 Evaluation

The predictions are evaluated using **Accuracy**, which is defined as the percentage of correct predictions out of the total number of test queries. The formula for accuracy is:

$$\text{Accuracy} = \sum_i \text{score}_i \times \frac{100}{\text{N}} \in [0, 100],$$

where N is the total number of test queries. score_i is 1 if your prediction for the i-th query is correct, and 0 otherwise. You should report accuracy on the validation set in your report.

1.2.5 Baseline Models

To provide a reference point for your model's performance, consider these simple baselines:

- **Random guessing:** outputs "True" or "False" each with probability $1/2$. Its accuracy is 50.00.
- **Counting:** computes the possibility of each reaction as follows: $P(S, D) = \frac{\sum_{(s_i, d_j) \in S \times D} \mathbb{1}(s_i, d_j)}{|S \times D|}$ where $S = \{s_1, s_2, ..., s_{|S|}\}$ represents the source of a reaction and $D = \{d_1, d_2, ..., d_{|D|}\}$ represents the destination of a reaction. Here, $\mathbb{I}(s_i, d_j)$ is 1 if the pair (s_i, d_j) exists in the training dataset, and 0 otherwise. If the calculated possibility is greater than 0.5, the output is "True"; otherwise, it is "False". Its accuracy is 67.29.

1.3 Task 2: Missing metabolite Prediction

1.3.1 Goal

Predict the missing metabolite in the source metabolite set for a given metabolic reaction. Instead of predicting just one metabolite, you are required to submit the top 10 candidate metabolite IDs, ranked in order of estimated likelihood.

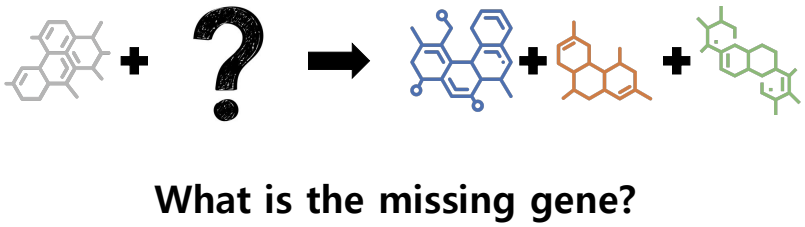

Figure 3: An illustration of Task 2.

1.3.2 Data Description

Each metabolic reaction is represented by two metabolite sets: source metabolites (reactants) and destination metabolites (products). In this task, some reactions have a missing metabolite in the source set that needs to be predicted. Both the source and destination metabolite sets can vary in size and consist of metabolite IDs, where each metabolite has a unique identifier from 0 to $n-1$, with n being the total number of metabolites.

You are provided with four data files to complete this task. Each file has a header line containing the column names, and data lines starting from the second line. The structure of the columns is described below, followed by dataset-specific information.

Common Columns All files share the following common columns:

- **reaction_id**: A unique identifier for the reaction.
- **source**: A set of metabolite IDs representing the **incomplete** source metabolites (reactants), with one metabolite missing.
- **destination**: A set of metabolite IDs representing the destination metabolites (products).

1. Training Data (`Completion_training.csv`) This file contains the full information for training your model, including the reaction identifier, source metabolites, destination metabolites, and the missing metabolite (to be predicted). Each row represents one reaction.

- **Additional column:**
  - **missing_metabolite**: The missing metabolite ID that needs to be predicted.

Example row:

`7305,{33},{0,486,982},20`

This indicates that for reaction #7305, the incomplete source metabolite set is $\{33\}$, the destination metabolite set is $\{0, 486, 982\}$, and the missing metabolite is 20.

2. Validation Query Data (`Completion_valid_query.csv`) This file is used to validate your model. It contains the reaction identifier, incomplete source metabolites, and destination metabolites but does not include the missing metabolite. You will use this data to predict the missing metabolite and compare it against the correct answers provided in the validation answer file.
Example row:

`8452,{17,23},{102,305,890}`

This row represents reaction #8452, where the incomplete source metabolite set is $\{17, 23\}$ and the destination metabolite set is $\{102, 305, 890\}$. Your model should predict the missing metabolite for this reaction.

3. Validation Answer Data (`Completion_valid_answer.csv`) This file contains the correct missing metabolites for the validation query data. After making predictions using the query data, you can compare them with the actual missing metabolites from this file to assess your model's accuracy.

- **Additional column:**
  - **missing_metabolite**: The correct missing metabolite ID.

Example row:

`8452,79`

For reaction #8452, the missing metabolite is 79.

4. Test Query Data (`Completion_test_query.csv`) This file contains reactions without the missing metabolite for testing your model on unseen data. The structure is the same as the validation query data. You will submit your predictions for these reactions.
Example row:

`9021,{45},{400,789}`

This row represents reaction #9021, where the incomplete source metabolite set is $\{45\}$ and the destination metabolite set is $\{400, 789\}$. Your task is to predict the missing metabolite for this reaction.

1.3.3 Submission

To complete this task, you need to:

- **Make Predictions:** For each reaction in the test query file, submit your top 10 predictions for the missing metabolite.
- **Submit Results:**
  - Your submission file should be named `Completion_test_answer.csv`.

- The columns of the file should be:
  - reaction_id: The unique identifier for the reaction.
  - top 10 candidates: A comma-separated list of your top 10 predictions for the missing metabolite, ranked in descending order by probability (from most to least likely).
- Ensure that the predicted metabolite lists are aligned correctly with the respective reactions in the test query file.

Example row of `Completion_test_answer.csv`:

```
8452,{17,23,19,34,28,12,4,7,15,5}
```

This example indicates that for reaction #8452, the top 10 predicted missing metabolites, ranked in descending order of probability, are $\{17, 23, 19, 34, 28, 12, 4, 7, 15, 5\}$.

1.3.4 Evaluation

The predictions are evaluated using two metrics:

- **Hits@10:** This measures the percentage of queries for which the correct missing metabolite is ranked within the top 10 predictions. The formula is:

$$\text{Hits@10} = \frac{\sum_{i=1}^{N} \mathbb{I}(\text{correct answer} \in \text{top 10 predictions for } i)}{N} \times 100,$$

  where N is the total number of test queries, and $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the correct answer is found within the top 10 predictions and 0 otherwise.

- **MRR@10 (Mean Reciprocal Rank):** This measures how well the model ranks the correct answer in the top 10 predictions. The formula is:

$$\text{MRR@10} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank of the correct answer for query } i},$$

  where the rank is the position of the correct answer in the top 10 predictions. For example, if the correct answer is ranked 1st, the rank is 1. If the correct answer is not in the top 10, the rank is treated as infinity, and the reciprocal rank is 0.

You should report both the Hits@10 and MRR@10 on the validation set in your report.

1.3.5 Baseline Models

To provide a reference point for your model's performance, consider these simple baselines:

- **Random guessing:** Randomly predicts the missing metabolite ID for each reaction by choosing 10 integers between 0 and $n - 1$ (all possible metabolite IDs). Its Hits@10 is 0.28, and MRR@10 is 0.0007.
- **Random guessing based on interactions:** Chooses 10 integers from the metabolite IDs that have interacted at least once with the destination elements in the training set. Its Hits@10 is 11.29, and MRR@10 is 0.04.

1.4 Task 3: Destination Prediction

1.4.1 Goal

Predict the destination metabolite set for each metabolic reaction, given the source metabolite set. That is, for each reaction, you should predict a set of metabolite IDs representing the products of the reaction.

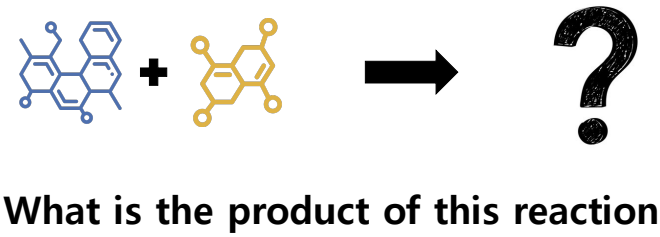

Figure 4: An illustration of Task 3.

1.4.2 Data Description

Each metabolic reaction is represented by two metabolite sets: source metabolites (reactants) and destination metabolites (products). Both the source and destination metabolite sets can vary in size and consist of metabolite IDs, where each metabolite has a unique identifier from 0 to $n-1$, with n being the total number of metabolites.

You are provided with four data files to complete this task. Each file has a header line containing the column names, and data lines starting from the second line. The structure of the columns is described below, followed by dataset-specific information.

Common Columns All files share the following common columns:

- **reaction_id**: A unique identifier for the reaction.
- **source**: A set of metabolite IDs representing the source metabolites (reactants).

1. Training Data (`Prediction_training.csv`) This file contains the full information for training your model, including the reaction identifier, source metabolites, and the destination metabolites (to be predicted). Each row represents one reaction.

- **Additional column:**
  - **destination**: The set of metabolite IDs representing the destination metabolites (products).

Example row:

```
7305,{20,33},{0,486,982}
```

This indicates that for reaction #7305, the source metabolite set is $\{20, 33\}$, and the destination metabolite set is $\{0, 486, 982\}$.

2. Validation Query Data (`Prediction_valid_query.csv`) This file is used to validate your model. It contains the reaction identifier and source metabolites, but does not include the destination metabolites. You will use this data to predict the destination metabolites and compare it against the correct answers provided in the validation answer file.
Example row:

```
8452,{17,23}
```

This row represents reaction #8452, where the source metabolite set is $\{17, 23\}$. Your model should predict the destination metabolites for this reaction.

3. Validation Answer Data (`Prediction_valid_answer.csv`) This file contains the correct destination metabolites for the validation query data. After making predictions using the query data, you can compare them with the actual destination metabolites from this file to assess your model's accuracy.

- **Additional column:**
  - **destination**: The correct set of metabolite IDs representing the destination metabolites.

Example row:

```
8452,{102,305,890}
```

For reaction #8452, the correct destination metabolite set is $\{102, 305, 890\}$.

4. Test Query Data (`Prediction_test_query.csv`) This file contains reactions without the destination metabolites for testing your model on unseen data. The structure is the same as the validation query data. You will submit your predictions for these reactions.
Example row:

```
9021,{45}
```

This row represents reaction #9021, where the source metabolite set is $\{45\}$. Your task is to predict the destination metabolite set for this reaction.

1.4.3 Submission

To complete this task, you need to:

- **Make Predictions:** For each reaction in the test query file, predict the destination metabolite set.
- **Submit Results:**
  - Your submission file should be named `Prediction_test_answer.csv`.
  - The file should follow the same format as `Prediction_valid_answer.csv`, with the same structure of columns.
  - The columns of the file should be:
    - reaction_id: The unique identifier for the reaction.
    - destination: A comma-separated list of your predicted destination metabolite set.
  - Ensure that the predicted metabolite sets are aligned correctly with the respective reactions in the test query file.

1.4.4 Evaluation

The predictions are evaluated using the following metrics:

- **Average Precision:** Measures the average precision across all test queries. Precision for each query is calculated as the proportion of correctly predicted destination metabolites out of all predicted metabolites:

$$\text{Average Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i \cap \hat{D}_i|}{|\hat{D}_i|},$$

  where D_i is the true destination metabolite set for the i-th reaction, $\hat{D}_i$ is the predicted destination metabolite set, and N is the total number of test queries.

- **Average Recall:** Measures the average recall across all test queries. Recall for each query is calculated as the proportion of correctly predicted destination metabolites out of the true destination metabolites:

$$\text{Average Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i \cap \hat{D}_i|}{|D_i|}.$$

- **Average F1-Score:** The harmonic mean of precision and recall, averaged across all test queries:

$$\text{Average F1-Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i},$$

  where Precision_i and Recall_i are the precision and recall for the i-th query.

You should report the average precision, recall, and F1-score on the validation set in your report. In these formulas, D_i represents the true set of destination metabolites for the i-th reaction, and $\hat{D}_i$ represents the predicted set of destination metabolites for that reaction.

1.4.5 Baseline Models

To provide a reference point for your model's performance, consider these simple baselines:

- **Random guessing:** Predicts the destination metabolite set by choosing k integers between 0 and $n-1$ (all possible metabolite IDs) uniformly at random, where k is proportional to the size distribution of destination metabolite sets in the training data. Its average precision, recall, and F1-score are 0.068, 0.07, and 0.07 respectively.

- **Random guessing based on interactions:** Predicts the destination metabolite set by choosing k integers from metabolite IDs that have interacted at least once with the source metabolites in the training set. Its average precision, recall, and F1-score are 4.69, 4.30, and 4.16 respectively.

1.5 Notes

- We may run the submitted code on another query dataset if your answer is suspiciously similar to any other group's answer.

- You may encounter some subtleties when it comes to implementation, please come up with your design and/or contact Heechan Moon (heechan9801 at kaist.ac.kr) and Kyuhan Lee (kyuhan.lee at kaist.ac.kr) for discussion. Any idea can be taken into consideration when grading if it is written in the *readme* file.
- Unlike the other assignments, you can use any programming language and any external library.

2 How to submit your project

2.1 Progress Report

Submit your progress report written on the attached template to KLMS by November 8, 2024, 11:59 pm. The file should be named report-[your student ids].pdf (e.g., report-20189000_20199000_20209000.pdf). Details will be announced soon.

2.2 Final Report Submission

1. Submit project-[your student ids].tar.gz (e.g., project-20189000_20199000_20209000.tar.gz) to KLMS. Your submission should contain the following files:
   - **final_report.pdf**: this file should be written on the attached template with LaTeX. This file should contain the accuracy of your predictions in the validation sets for direction prediction, missing gene prediction, and destination prediction tasks.
   - **test_prediction.tar.gz**: this file should contain the following files:
     - `Classification_test_answer.csv`
     - `Completion_test_answer.csv`
     - `Prediction_test_answer.csv`
   - **readme.txt**: this file should contain the names of individuals from whom you received help and the natures of help that you received. This includes help from friends, classmates, lab TAs, course staff members, etc. In this file, you are also welcome to write any comment that can help us grade your assignment better, your evaluation of this assignment, and your ideas. This file also should describe how to run your code.
   - **code.tar.gz**: your implementation.
2. Make sure that no other files are included in the tar.gz file.

2.3 Video Presentation Submission

1. Submit project-[your student ids].tar.gz (e.g., project-20189000_20199000_20209000.tar.gz) to KLMS. Your submission should contain the following files:
   - **slides.pdf**: slides used for the final presentation.
   - **video.mp4**: recorded video presentation that does not exceed 5 minutes.
2. Make sure that no other files are included in the tar.gz file.