ANNOTATING ANTIBIOTIC RESISTANCE GENES USING GRAPH NEURAL NETWORK

goc Khoi Dang, Jun Chen

N

Introduction

- Antibiotic is the primary weapon to fight against the bacterial infection.
- Antibiotic resistance became a global threat due to the overuse and misuse of these medications.
- Antibiotic resistance protein (ARP) could deactivate or repel antibiotics.

Problem

- The protein labeling problem is a multilabel classification problem that can be processed quickly with the helping of computational protein function classifiers.
- More effectively labeling antibiotic resistance protein could help to improve the study of antibiotic resistance.

Approach

- Computational protein function classifiers.
- Building model to deal with antibiotic resistance protein function prediction
- We will use CARD dataset to train and evaluate the model.

Methods

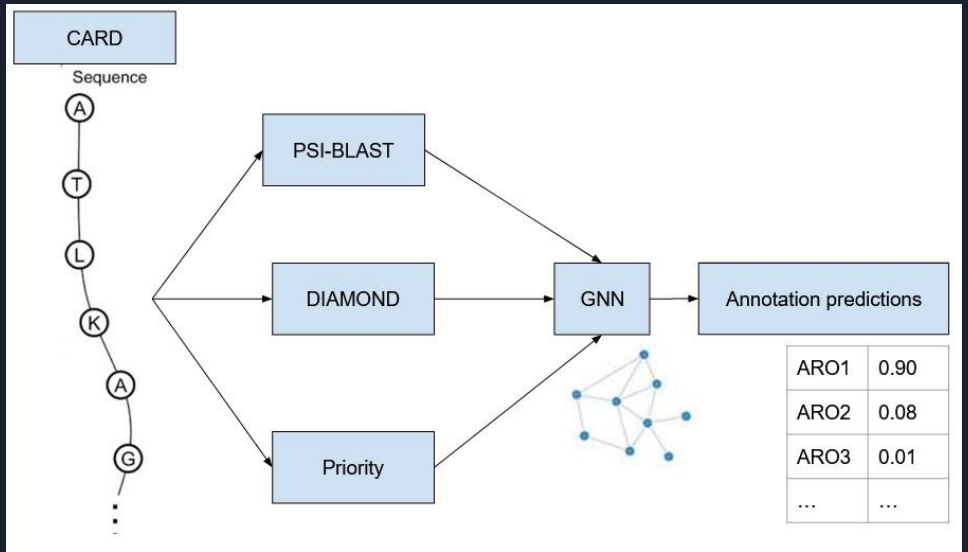
- The model is Graph Neural network.
  - Each node represent a ARO and the edge represent the relations between nodes.
  - The graph is trained with node features, edge features, and global features.
  - Each node has 12 node features which are 10 top scores from PSI-BLAST [3], DIAMOND score [4] and Priority score [1].
  - The global feature is a vector of amino acid composition with a length of 20. The edge feature is just the relation of 2 nodes. The initial values would be assigned randomly from 0 to 1.
- We plan to use PyTorch and a graph network framework library.
- We plan to use Adam optimizer with a learning rate of 0.01.

Datasets

- 4,577 annotated ARP sequences from CARD database. The dataset will be splitted into training set and testing set with the ration 8:2.

Metrics for evaluation

- F-measure score (Fmax) is the maximum F-measure over confidence score thresholds t.
- Minimum semantic distance score (Smin) is the metric that considered the unbalanced information content (IC) of GO terms.
- Area under the precision-recall curve (AUPR)



Project milestones

- 3/14 - 4/04: Code implementation
  - Khoi: Implement graph neural network
  - Jun: Implement node feature extractions.
- 4/04 - 4/25: Evaluation and tuning
  - Khoi: Evaluation metrics implementation
  - Jun: Hyper parameters tuning
- 4/18 - 4/25: Presentation (Khoi + Jun)
- 4/25 - 5/04: Final report (Khoi + Jun)

Bibliography

[1] C. Zhao, T. Liu, and Z. Wang, “Panda2: protein function prediction using graph neural networks,” NAR Genomics and Bioinformatics, vol. 4, no. 1, p. lqac004, 2022.

[2] B. P. Alcock, A. R. Raphenya, T. T. Lau, K. K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.-L. V. Nguyen, A. A. Cheng, S. Liu, et al., “Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database,” Nucleic acids research, vol. 48, no. D1, pp. D517–D525, 2020.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped blast and psi-blast: a new generation of protein database search programs,” Nucleic acids research, vol. 25, no. 17, pp. 3389–3402, 1997.

[4] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using diamond,” Nature methods, vol. 12, no. 1, pp. 59–60, 2015.