CS177 Bioinformatics: Software Development and Applications

Guidelines for Course Projects

School of Information Science and Technology, ShanghaiTech University

Spring, 2024

Lecturer: Associate Professor Jie Zheng (郑杰), Email: zhengjie@shanghaitech.edu.cn

Teaching Assistants:

- Miss Siyu Tao (陶思宇), Email: taosy2022@shanghaitech.edu.cn
- Mr. Ruoyin Zhang (张若隐), Email:zhangry12023@shanghaitech.edu.cn

1. General information and rules

The projects are mainly about implementing, comparing and improving algorithms and deep learning models in bioinformatics. Through a hands-on project, students can test their understanding of computational problems and ideas in bioinformatics, and learn how to make innovation in data science. Moreover, students will be trained to present their project ideas and results clearly through oral presentations and written reports, paving the way for becoming successful researchers or engineers in the future. Hereafter, "you" refers to every student who has been officially enrolled in the course.

Each project is **group-based** and each group comprises 2-3 students**,** but you should not rely on anyone outside of your team for the project. The 3 instructors (Dr. Jie Zheng, and his TAs, Siyu Tao and Ruoyin Zhang) are the persons who will make judgment and evaluation of your performance in the project.

2. Timeline：

- **Project start:** March 28, 2024
- **Mid-term personal interviews:** Week 11 (date TBD)
- **Oral representation:** Week 16 (date TBD)
- **Submission of final report (with other project files):** Week 16 (date TBD)

3. Description of projects

There are 2 project options as described below. Each group is required to choose only one project. Each project is to be chosen by at most 2 groups. *If more than two groups happen to choose the same project, the instructors will select two groups by interviews.* Once choosing a project, you should not switch to another one, unless absolutely necessary.

3.1 Project 1: Leveraging multi-omics data for context-specific synthetic lethality (SL) prediction

Background

Synthetic lethality (SL) is a type of genetic interaction in which the simultaneous inactivation of two genes leads to cell death, while the inactivation of a single gene does not affect the cell viability, as illustrated in Figure 1. Therefore, SL can be used to selectively kill cancer cells by targeting SL partners of cancer-specific genetic abnormalities while keeping normal cells alive. However, high-throughput wet-lab screenings of SLs are often costly and face various challenges (e.g., high cost and off-target effects). Therefore, deep learning (DL) methods for SL prediction have become popular in predicting SLs.

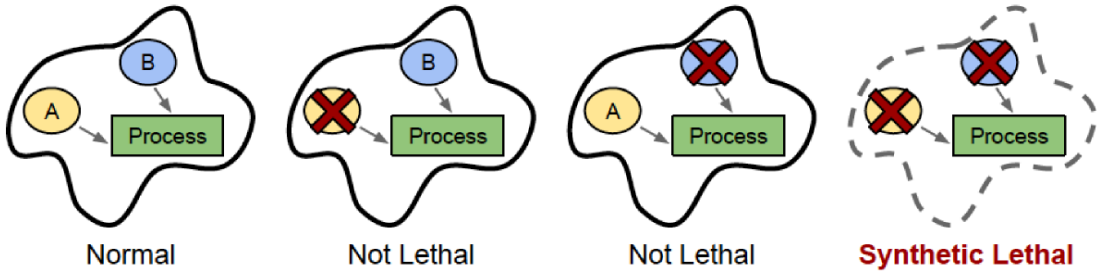


Figure 1. The concept of synthetic lethality.

Many machine learning approaches, including both cell-line-free methods [1-3] and cell-line-specific methods [4-9], have been proposed for predicting SLs. However, recent studies have shown that many SL relationships occur specifically in certain cell lines, e.g., two genes often have SL relationship in one cell line but not in other cell lines. In other words, SL pairs are significantly heterogeneous among cell lines. Therefore, the heterogeneity of SL pairs across cell lines necessitates models for more precise SLs prediction. Multi-omics data have been demonstrated to reflect the context-specific genetic information of cell lines or cancer types. However, there are still unexplored but crucial issues, e.g. how to select effective multi-omics data combination to enhance the context-specific SL prediction.

Goal: In this project, we aim to leverage powerful deep learning (DL) techniques to learn from the multi-omics data and thereby increase the performance of predicting context-specific SLs.

Sub-goals and recommended steps:

1. Gather and integrate multi-omics data for capturing context-specific information. You are recommended (but not required) to select the multi-omics data from the following list:

- Biomedical knowledge graph (BKG): PrimeKG, SynLethKG;
- Protein sequence: UniProt;
- Gene expression data, copy number variation, mutation, and methylation: CCLE or TCGA;
- Protein-Protein interaction (PPI) network: STRING.

Furthermore, the studies mentioned in the background may shed light on the selection of multi-omics data combinations.

2. Select and implement the DL methods for learning the representations of the multi-omics data. Moreover, the following DL methods are recommended as encoders of the two types of data:
   - Graph-structure data: GNN;
   - Protein sequence data: ESM2.

   To assist you in constructing your model more effectively, we provide a model framework for reference (shown in Figure 2).

Figure 2. The recommended model framework.

Given a gene pair (g_i, g_j) and a cell line c_k, you can first adopt a cell line representation learning strategy, i.e., representing it as an embedding h_{ck}, to better capture the cell line information. Then, adopt the BKG and protein sequence to learn gene embeddings. Specifically, a GNN encoder (e.g., GCN) can be used to generate gene representations from the BKG, while ESM2 is applied to produce gene representations from the protein sequence. Then, you can concatenate the above representations as the final representation of the triplet (g_i, g_j, c_k). Finally, feed the final representation into a classifier (e.g., Multi-layer perceptron) to predict the SL relationship between the two gene in the cell line.

3. Evaluate the performance of your proposed model and the baselines compared with it in the context-specific SL prediction scenarios. Note that we formulate the context-specific SL prediction as a binary classification task, i.e., 1 for positive SL pairs and 0 for non-SL pairs.
   In this project, you need to complete ***at least*** the following two experiments:

   1) **Cell-line-specific SL prediction:** For each cell line, use 5-fold cross-validation (CV) to randomly split the gene pairs in this cell line. All the compared models are to be trained and tested independently for the cell line.

2) **Cell-line-adapted SL prediction:** In this experiment, we evaluate the models in a cross-cell-line prediction scenario, namely (Jurkat, A375)⟶A549, i.e., the first two cell lines provide the training data, and the last cell line the test data.

Here, we provide the SL label data:

https://epan.shanghaitech.edu.cn/l/TF3uIz (提取码：CS177)

The baselines to be compared (you can include more):

1) KG4SL [1]
2) SLMGAE [2]
3) MVGCN-iSL [4]

Besides, it is recommended to use the following classification metrics:

- Area under the receiver operating characteristic curve (AUC),
- Area under the precision-recall curve (AUCPR),
- F1 score,
- Balanced accuracy (BACC).

4. A discussion about the multi-omics data combinations of the proposed methods is recommended. For example, which types of data play important roles in predicting SLs and why?

Useful links:

SynLethDB 2.0: http://synlethdb.sist.shanghaitech.edu.cn/v2

PrimeKG: https://zitniklab.hms.harvard.edu/projects/PrimeKG/

SLKB: https://slkb.osubmi.org/

ESM2: https://github.com/facebookresearch/esm

PyG: https://pytorch-geometric.readthedocs.io/en/latest/

SL benchmark: https://github.com/JieZheng-ShanghaiTech/SL_benchmark?tab=readme-ov-file#benchmarking-of-machine-learning-methods-for-predicting-synthetic-lethality-interactions

References

[1] Wang S, Xu F, Li Y, et al. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers[J]. Bioinformatics, 2021, 37(Supplement_1): i418-i425.

[2] Hao Z, Wu D, Fang Y, et al. Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(10): 4041-4051.

[3] Wang S, Feng Y, Liu X, et al. NSF4SL: negative-sample-free contrastive learning for ranking synthetic lethal partner genes in human cancers[J]. Bioinformatics, 2022, 38(Supplement_2): ii13-ii19.

[4] Fan K, Tang S, Gökbağ B, et al. Multi-view graph convolutional network for cancer cell-specific synthetic lethality prediction[J]. Frontiers in Genetics, 2023, 13: 1103092.

[5] Xing Y, Pu M, Tian K, et al. Cell context-specific Synthetic lethality Prediction and Mechanism

Analysis[J]. bioRxiv, 2023: 2023.09. 13.557545.
[6] Wan F, Li S, Tian T, et al. EXP2SL: a machine learning framework for cell-line-specific synthetic lethality prediction[J]. Frontiers in Pharmacology, 2020, 11: 507703.
[7] Tepeli Y I, Seale C, Gonçalves J P. ELISL: early–late integrated synthetic lethality prediction in cancer[J]. Bioinformatics, 2024, 40(1): btad764.
[8] Zhao X, Liu H, Dai Q, et al. Multi-omics Sampling-based Graph Transformer for Synthetic Lethality Prediction[C]. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023: 785-792.
[9] X. Liu, S. Tao and J. Zheng. Meta Learning for Low-Data Prediction of Cancer-Specific Synthetic Lethality as Drug Targets. 2023 IEEE International Conference on Knowledge Graph (ICKG), Shanghai, China, 2023, pp. 255-262, doi: 10.1109/ICKG59574.2023.00037.

3.2 Project 2: Evaluation of quality of predicted protein models

Background

Protein structure prediction plays important roles in biological research, particularly with the advent of deep learning techniques that have significantly accelerated its progress and transformation. Techniques such as AlphaFold2, RoseTTAFold, and ESMFold have ushered in a new era of high-accuracy protein structure prediction. Despite the remarkable internal confidence estimation provided by tools like AlphaFold2's pLDDT, relying solely on a single metric to assess the quality of the predicted protein structures may be insufficient.

Therefore, the model quality evaluation plays a key role in structural biology. It involves not just differentiating between high and low-quality protein models, but also identifying inaccurately modeled regions within high-quality structures for further refinement. Such evaluations are essential, as the reliability and applicability of these models directly impact the efficiency of processes like target identification and drug design.

Goal: In this project, we focus on the assessment of the quality of protein models by implementing a baseline and trying to improve it. By doing so, you can understand better the critical roles of model quality evaluation in structural biology.

Sub-goals and recommended steps:

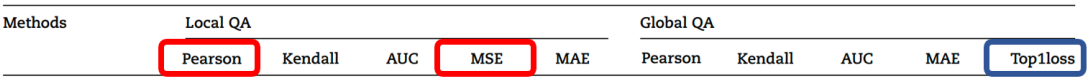
1. Select QATEN or GraphQA as the baseline model. Here, the links to the two baselines are provided below.
   - https://github.com/CQ-zhang-2016/QATEN
   - https://github.com/baldassarreFe/graphqa

2. Retrain and improve the chosen baseline method using the following training dataset: https://epan.shanghaitech.edu.cn/l/IF2WXi
   Furthermore, you need to predict lDDT values for each residue in the protein model in this project. These predictions should be:
   - **Recorded in the B-Factor Field**: Predicted lDDT scores must be inserted into the B-factor

field (columns 61-66 of the PDB format). Each score should reflect the local structural accuracy of the corresponding residue.

- **A Measure of Structural Accuracy**: lDDT is a measure of how accurately local structural motifs are predicted, with higher values indicating better accuracy.

3. Evaluate the enhanced method on a *new test dataset* for prediction accuracy. Note that we will provide the test dataset *after the mid-term personal interviews.* Moreover, you should use at least two evaluation metrics to assess the performance, including Pearson correlation coefficient and Mean Squared Error (MSE).



References

[1] Baldassarre F. et al. (2021) GraphQA: Protein model quality assessment using graph convolutional networks. Bioinformatics, 37, 360–366.

[2] Hiranuma N. et al. (2021) Improved protein structure refinement guided by deep learning based accuracy estimation. Nat. Commun., 12, 1340.

[3] Chen, C. et al. (2023). 3d-equivariant graph neural networks for protein model quality assessment. Bioinformatics, 39(1), btad030.

[4] Dong L. et al. (2024): Assessing protein model quality based on deep graph coupled networks using protein language model. Briefings in Bioinformatics, 2024, 25(1): bbad420.

[5] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.

END OF CS177 PROJECT GUIDELINES (SPRING, 2024)