2024 OXCAM Programme – AI+ Course Group Project Proposal – BIO Track

Group 3

August 16, 2024

Programme	**Biotechnology Engineering & Healthcare Technology**		
Course Group	**BIOA-G3**	**Group Name**	**Stray Birds**
Group Members	**Dong Shuyi , Wang Xinyi , Yang Keyi , Qin Hanqing , Zheng Xi , Zhou Yangyang**		

1 Project title

Analysis of phase separation proteins based on machine learning to help assess neurodegeneration

2 Project aims and objectives

Overall aim:

By collecting data from public databank to obtain datasets , we start model training and evaluation, thus achieving a predictive tool to help assess neurodegeneration to promote global healthcare.

Specific objectives:

I . project objectives

1. Use statistical analysis of protein characteristics in database to select biomarkers for data collection.
2. Prepare reliable datasets based on selected biomarkers.
3. Choose appropriate model and find out corresponding code.
4. Encode datasets to train the model, evaluate the results.
5. Achieve an ideal predictive tool.

II . Team objectives

1. **Team realizes how to use AI in this field:** We prepare datasets based on biological analysis and statistical science, input datasets to our selected model and start training, finally the whole team analyzes the results and evaluate

the performance of the predictive model. By incorporating AI and ML approaches, our team adopt their application to project practical achievement, enabling more precise and data-driven solutions and effectively harnessing AI and machine learning tools.

2. **Team enrichment:** Through previous online learning, our team has grasped the principles of AI and developed a solid understanding of machine learning algorithms, ensuring and supporting for practical application.Through this golden opportunity our team gets to delve into the fascinating fields of AI and ML. The application of machine learning empowers the team to solve complex challenges with innovative and efficient approaches, and the team can have a further understanding of how to align AI capabilities with our project's strategic goals to drive innovation and have the ability of exploring complex interactions and non-linear effects using large data sets

3 Background and motivation

Liquid-liquid phase separation (LLPS) is a reversible process of a homogeneous fluid demixing into two distinct liquid phases, one condensed phase and one dilute phase [1]. Many biological processes have been revealed to be regulated by LLPS in biological systems, normal phase separation (PS) proteins typically return to the dilute phase once they have completed their function, while in some situations the liquid condensates can transform into solid aggregates without reverting to a liquid-like state, which are deemed to be implicated as an obvious marker related to neurodegenerative diseases [2].

Previous work has been conducted to screen out phase transition proteins with machine-learning models that integrate multimodal features [3], and develop public databanks of proteins undergoing LLPS in vitro such as LLPSDB and PhaSepDB [4] [5] [6]. However, proteins that undergo either normal PS or abnormal PS have not been studied extensively, the scientific challenge lies in understanding the conditions under which a protein changes from normal PS to abnormal PS, and what modifications (such as PTMs) occur in these conditions [7]. Based on that, we propose to analyze those PS proteins compared to those going through aberrant phase separation by using machine learning algorithms, to help predict those that have propensity to form into aberrant phase separation and the conditions in relationship with neurodegenerative diseases , providing a new insight for researchers to explore the mechanism within and better assessing neurodegenerative diseases [2].

4 Project Plan

Work Packages

(1). Group activities: 8.12—8.24

1. Group collective team building
2. Group collective learning
   - communicate and discuss with members about AI
   - share resources and create shared documents
   - get members' feedback and effectively update

(2). Database collection and pre-processing: 8.12—8.15

1. identify biomarkers
2. prepare datasets

Methods:classification of data, statistical science
Sources: LLPSDB, PhaSepDB
Personnel: All Responsible person: Wang Xinyi, Yang Keyi)

(3). Model training and evaluation: 8.15—8.18

1. encode machine learning algorithm
2. evaluate with experiments testing data

Methods: Python coding skill
Sources:prepared datasets, Github,literature
Personnel: All Responsible person:Qin Hanqing

(4). Results analysis: 8.18—8.20

1. analyze the prediction results
2. score the model to assess its performance

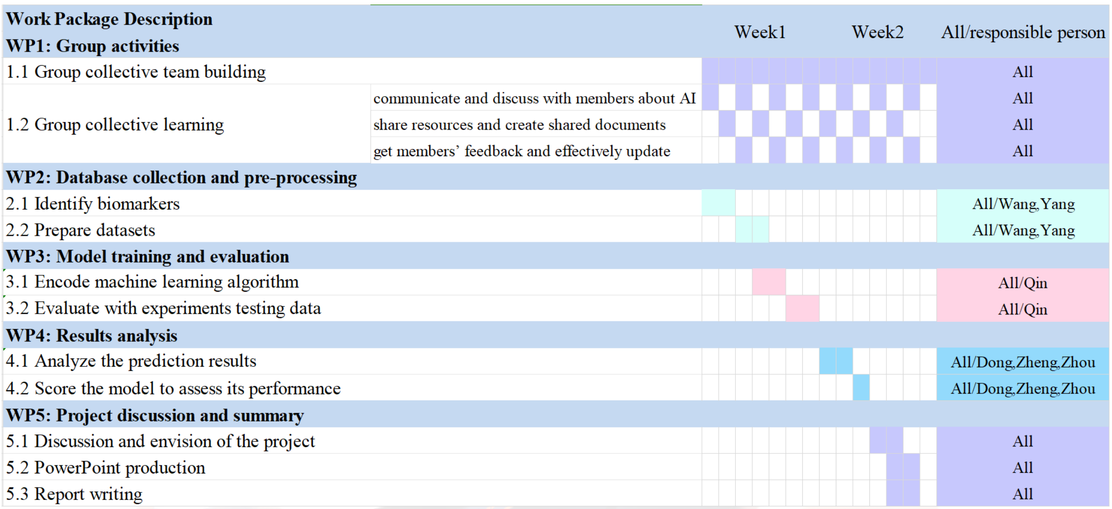
Methods:rational analysis
Sources:model prediction results
Personnel: All Responsible person:Dong Shuyi , Zheng Xi, Zhou Yangyang

(5). Project discussion and summary: 8.21—8.23

1. discussion and envision of the project
2. PowerPoint production
3. report writing

Methods:discussion and thinking
Sources:project results
Personnel: All

5 Project Timeline (Gantt Chart) & allocation of tasks



Work Package Description		Week1							Week2							All/responsible person
WP1: Group activities																
1.1 Group collective team building		■	■	■	■	■	■	■	■	■	■	■	■	■	■	All
1.2 Group collective learning	communicate and discuss with members about AI	■		■		■		■		■		■		■		All
	share resources and create shared documents		■		■		■		■		■		■			All
	get members' feedback and effectively update			■		■		■		■		■		■		All
WP2: Database collection and pre-processing																
2.1 Identify biomarkers		■	■													All/Wang,Yang
2.2 Prepare datasets				■	■											All/Wang,Yang
WP3: Model training and evaluation																
3.1 Encode machine learning algorithm					■	■										All/Qin
3.2 Evaluate with experiments testing data							■	■								All/Qin
WP4: Results analysis																
4.1 Analyze the prediction results									■	■						All/Dong,Zheng,Zhou
4.2 Score the model to assess its performance											■					All/Dong,Zheng,Zhou
WP5: Project discussion and summary																
5.1 Discussion and envision of the project												■	■			All
5.2 PowerPoint production													■	■		All
5.3 Report writing													■	■		All

Figure 1: Project Timeline (Gantt Chart) & allocation of tasks

6 References cited

References

[1] Y. Gao, X. Li, P. Li, and Y. Lin, "A brief guideline for studies of phase-separated biomolecular condensates," *Nature Chemical Biology*, vol. 18, no. 12, pp. 1307–1318, 2022.

[2] M. Ding, W. Xu, G. Pei, and P. Li, "Long way up: rethink diseases in light of phase separation and phase transition," *Protein & Cell*, p. pwad057, 2023.

[3] Z. Chen, C. Hou, L. Wang, C. Yu, T. Chen, B. Shen, Y. Hou, P. Li, and T. Li, "Screening membraneless organelle participants with machine-learning models that integrate multimodal features," *Proceedings of the National Academy of Sciences*, vol. 119, no. 24, p. e2115369119, 2022.

[4] Q. Li, X. Peng, Y. Li, W. Tang, J. Zhu, J. Huang, Y. Qi, and Z. Zhang, "Llpsdb: a database of proteins undergoing liquid–liquid phase separation in vitro," *Nucleic acids research*, vol. 48, no. D1, pp. D320–D327, 2020.

[5] K. You, Q. Huang, C. Yu, B. Shen, C. Sevilla, M. Shi, H. Hermjakob, Y. Chen, and T. Li, "Phasepdb: a database of liquid–liquid phase separation related proteins," *Nucleic acids research*, vol. 48, no. D1, pp. D354–D359, 2020.

[6] R. Pancsa, W. Vranken, and B. Mészáros, "Computational resources for identifying and describing proteins driving liquid–liquid phase separation," *Briefings in Bioinformatics*, vol. 22, no. 5, p. bbaa408, 2021.

[7] B. Shen, Z. Chen, C. Yu, T. Chen, M. Shi, and T. Li, "Computational screening of phase-separating proteins," *Genomics, Proteomics and Bioinformatics*, vol. 19, no. 1, pp. 13–24, 2021.