Computational Biology and Data Mining

TH BINGEN
University of Applied Sciences

Kick Off Meeting: Bachelor thesis of Robin Ender

Date: *2025-05-08*

Attendees: *Robin Ender, Asis Hallab, Eric Schumbera*

The Task / Scientific Question

Are **multiple block decompositions** of protein sequences able to predict **Phase Separation Propensity**?

How to answer this question?

1. rewriting the current block decomposition algorithm to get a more meaningful output for Deep Learning Models
2. acquiring and processing curated training / test data
3. find meaningful mappings for running the block decomposition
4. run the block decomposition on the data
5. train Deep Learning Models with the output
6. compare the capabilities to other Phase Separation Predictors

Background

- phase separation is mainly driven by two forces: [1]
  - protein-protein or protein-RNA interaction domains
  - interactions between intrinsically disordered regions
- current predictors use machine learning models that are trained on properties like fraction of each amino acid, or fraction of intrinsically disordered regions
  - one also integrates structural information obtained from AlphaFold

[1] S. Hou, J. Hu, Z. Yu, D. Li, C. Liu, and Y. Zhang, "Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions," *Nature Communications*, vol. 15, no. 1, p. 2147, Mar. 2024, doi: 10.1038/s41467-024-46445-y.

The Block Decomposition Algorithm

- the block decomposition algorithm is able to find all factors of a sequence that have a balance lower or equal to the balance threshold
  - the balance threshold ensures that these blocks have a certain homogeneity
- to be less sensitive to substitutions the protein sequences are mapped before the decomposition
  - this leads to homogeneous blocks in the context of the current mapping that can be labeled to provide additional info

The Block Decomposition Algorithm

- The labeling function will give information about the main components of the block
  - if a block consists of mainly amino acids with a mapping of 1, the label would be 1
  - if a block consists mainly of two amino acids mapped to 1 and 2 it gets the label 12 ...
- leading to a decomposition that could look like:

```
[rep(0, 3), rep(1, 14), rep(0, 2), rep(12, 30), ... ]
```

The Idea

- if many different mappings are used where each represents a different property of the amino acids in it, a multidimensional block decomposition is created
- using deep learning models like convolutional neural networks (CNNs) the different decompositions can be treated similar to the color channels of a picture, like so:

0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 2 2 2 2 2 2

3 3 3 3 2 2 2 2 2 2 0 0 0 0 0 0 1 1 1 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 0 0 0

0 0 0 2 2 2 2 1 1 1 1 0 0 0 3 3 3 3 4 4 4 4 4 2 2 2 2 2 2 2 4 4 4 4 6 6 6

The Idea

- as CNNs alone are not able to work with long range interactions a hybrid with Long short-term Memory or transformers may be able to observe relations between the different channels that is able to predict attributes like phase separation

The Steps in Detail

Reimplementing the block decomposition algorithm

- currently the block decomposition algorithm outputs data in the form of a block list, where each entry is a list containing a start and end position:
  - `[[1, 20], [24, 45], ...]`
  - This output lacks a label for the block as well as compatibility with deep learning models
- The desired output would look like this, where each position is the position of a amino acid, and the number is the label:
  - `[0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2...]`

Reimplementing the block decomposition algorithm

- This will be implemented as a Python module
- if python is too slow, the algorithm will be rewritten in a low level language like Rust

Data Acquisition

- the data will be obtained from previous studies that developed phase separation predictors[1] [2]
  - these studies took their data from curated databases for phase separation
- this makes the results comparable and saves time

[1] S. Hou, J. Hu, Z. Yu, D. Li, C. Liu, and Y. Zhang, "Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions," *Nature Communications*, vol. 15, no. 1, p. 2147, Mar. 2024, doi: 10.1038/s41467-024-46445-y.

[2] Z. Chen *et al.*, "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, Jun. 2022, doi: 10.1073/pnas.2115369119.

Finding meaningful Mappings

- To find meaningful mappings the results of the previous studies will be investigated as they provide lists of features that contributed most to phase separation propensity
  - for example mappings for idr related amino acids or amino acids that are involved in pi-pi interactions should be created

Training Models

- As already described, the multidimensional block decomposition can be interpreted as a one dimensional image with many color channels, therefore 1dCNNs will be used
- They will be integrated with Long short-term memory or transformer models to account for long distance relations
- pyTorch will probably be used for this

Benchmarking

- if the trained models are capable of predicting phase separation a comparison to the other models will be made