Information-dense features for the neural network-aided prediction of core interacting residues in protein interactions

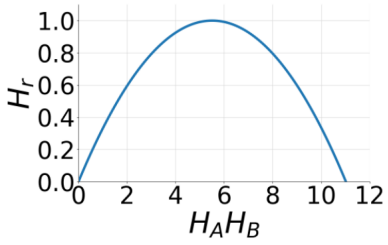
Table of contents:

- Project steps
  - Data download
  - I TASK
  - II TASK
  - OPTIONAL TASK
  - TIPS
- Bibliography

Understanding the molecular mechanisms underlying protein-protein interactions is fundamental for comprehending the basic mechanisms of cellular processes, with significant implications for therapeutic and biotechnological applications. Despite this, elucidating the binding process and the stability of the resulting complexes remains a challenging task.
A protein structure is represented by a set of atomic coordinates, where each atom has an x, y, and z position in three-dimensional space. A dimer is defined as the (non-covalent) interaction between two protein structures, which together form a protein complex. Each protein comprises a certain number of residues (amino acids that make up the protein's polypeptide chain), but only a subset of these are considered interacting residues, referred to as binding sites. They are typically determined based on their distance from the residues of the partner protein. The prediction of binding site residues between two interacting proteins remains an open challenge in computational biology. Solving this challenge would enable protein-protein docking algorithms to more efficiently identify native-like solutions.

In this context, we recently developed the Core Interacting Residues Network (CIRNet) [1], a novel predictive approach for identifying interacting residues based on neural networks. CIRNet classifies pairs of residues as interacting residues at the core of the interfaces (central residues in the two binding sites) or non-interacting residues, achieving an accuracy of approximately 0.87 on a balanced dataset.
The neural network was trained on a set of indicators providing a compact representation of protein regions, incorporating properties derived from both Coulomb potential and Lennard-Jones potential (which strongly influences the shape complementarity of interacting molecular surfaces). Specifically, the descriptors were obtained through the Zernike 2D polynomial expansion of molecular surface patches of the two interacting proteins, enabling rapid evaluation of shape [2] and electrostatic complementarity [3].

More importantly for this project, the method integrates a rapid evaluation of hydrophobicity or hydrophilicity (here, we use the term "hydropathy") complementarity. Over the decades, many hydropathy scales have been developed to numerically characterize the hydropathy profile of an amino acid. For each of these scales—some based on experimental data and others on statistical knowledge (theoretical scales)—each of the 20 natural amino acids is assigned a value quantifying its degree of hydropathy. The protocol implemented in CIRNet can accept, among its various descriptors, a value indicating the hydropathy complementarity between pairs of residues, computed starting from their hydrophobicity values. In the original version of CIRNet, the two hydropathy values are taken from the scale proposed by Di Rienzo, et al [4] (L_hydrophobicity_scale file). Hydropathy complementarity is defined as in the Figure:

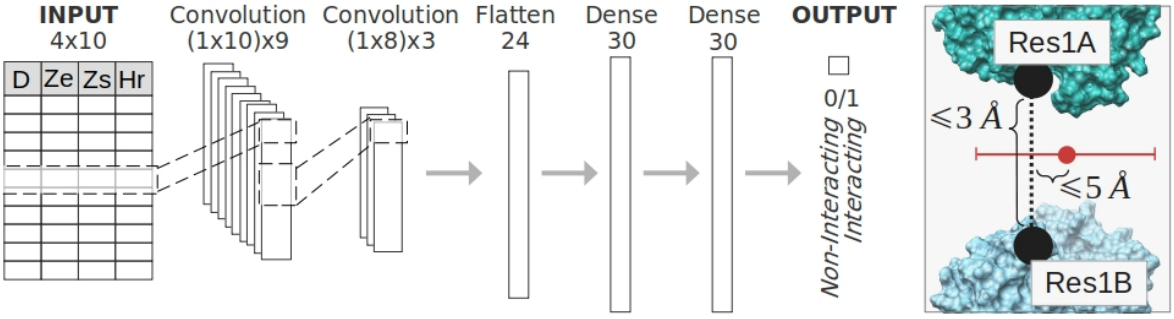


$$H_r = -a(H_A H_B)^2 + b(H_A H_B)$$

where H_A and H_B are the hydrophobicity values of residues *A* and *B* respectively, while H_r is their hydropathy complementarity given as input to CIRNet. The parameters *a* and *b* are set to 0.033 and 0.363 respectively. This expression of H_r reflects that residues with similar values of hydrophobicity (both high or low) have stronger interactions (and thus higher complementarities, corresponding to low H_r values) than those between residues with opposing characteristics.

Depending on the input hydropathy scale, CIRNet's performance in identifying residue-residue interacting pairs will vary. Thus, the main objective of this work is to investigate the impact of different types of hydropathy scales on CIRNet's ability to predict interacting residue pairs.

The following Figure shows CIRNet architecture. For each pair of residues *A* and *B*, it receives a 4x10 matrix with the shape, electrostatic, and hydropathy complementarity values between *A* and *B* (first row) and the complementarity values between *A* and the first nine neighbors of *B*.



In this project, you will use CIRNet to identify core interacting residues from a dataset of protein dimers and analyse how different hydropathy scales influence its performance.

Project steps

Data download

Download the directories:

CODES: all the starting python codes.

codes
└──cnn.py -> CIRNet architecture

DATASET: training and testing data.

dataset
├── train
    ├── classification.txt -> True classification of the residue pair as core interacting (1) or not (0)

```
    ├── dataset.txt -> True classification, name of the complex, residue on the first protein (A_n), residue on the second protein (B_n)
    ├── shape.txt -> Shape complementarity between A_n and B_n and between A_n and the nine neighbors of B_n
(B_n,B_n_1,...,B_n_10)
    ├── el.txt -> Electrostatic complementarity between A_n and B_n and between A_n and the nine neighbors of B_n
    ├── hr.txt ->  Hydropathy complementarity (defined according to the L_hydrophobicity_scale)  between A_n and B_n and between
A_n and the nine neighbors of B_n
    └── dist.txt -> Distance of B_n,B_n_1,...,B_n_10 from B_n
└── test
    ├── classification.txt
    ├── dataset.txt
    ├── shape.txt
    ├── el.txt
    ├── hr.txt
    └── dist.txt
```

HYDROPATHY: Original and new hydropathy scales

hydropathy -> L_hydrophobicity_scale.csv [4] and other 27 scales

The 27 scales were extracted from the AAindex database in Python, using the aaindex Python package.
Each file is named as the AAindex accession number of that scale.

I TASK

Train and test CIRNet with the original dataset **(based on the L_hydrophobicity_scale)**. Find the optimal threshold for the NN classification in interacting and non-interacting pairs and study how the accuracy varies for different residue pair type (you can find the residue names in the file dataset.txt.).

Study the effect of each new hydropathy scale downloaded from Github by training and testing CIRNet: for each scale, define an appropriate hydropathy complementarity formula and write a new hr.txt file. Perform the same analysis as in the first point.

II TASK

Compare the results obtained for all the proposed hydropathy scales.

Define a new scale combining the proposed ones.

OPTIONAL TASK

Try to improve CIRNet accuracy. You can modify the NN structure or add more specific thresholds to the NN prediction.

TIPS

I TASK: To find the optimal classification thresholds you can start by plotting the distributions of the NN prediction stratified in true interacting and non-interacting residues. To evaluate the accuracy you can compute the F1-score.

I TASK: Residues can be classified according to their chemical nature in Polar (P), Hydrophobic (H), and Charged (C).

H = ['GLY', 'ALA', 'VAL', 'LEU', 'ILE','MET','PHE', 'TYR', 'TRP']

P = ['SER', 'PRO','THR','CYS','ASN', 'GLN']

C = ['HIS', 'LYS','ARG','ASP','GLU']

The chemical composition of the interfaces is reflected in the hydropathy complementarity values between core interacting residue pairs (PP, HH, CC, PH, PC, HC). You can find the residue names in the file dataset.txt.

II TASK: You can compute a Principal Component Analysis (PCA) on all the 28 scales. Each PC will be defined by a combination of the 28 scales, and can be seen as a “new scale”. The contribution of each scale to the PC can be examined through the loadings.

Bibliography

[1] G. Grassmann, et al. ‘Compact assessment of molecular surface complementarities enhances neural network-aided prediction of key binding residues’. arXiv preprint:2407.20992. 2024.

[2] E. Milanetti, et al. ‘2D Zernike polynomial expansion: Finding the protein-protein binding regions.’ Computational and structural biotechnology journal, 19, 29-36. 2021.

[3] G. Grassmann, et al. ‘Electrostatic complementarity at the interface drives transient protein-protein interactions’. Scientific reports, 13(1), 10207. 2023.

[4] L. Di Rienzo, et al. ‘Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network.’ Frontiers in molecular biosciences 8: 626837. 2021.