FINAL PROJECT BSc COURSE MACHINE LEARNING
MSc Bioinformatics & System Biology students version

Protein-Protein Ineraction (PPI) dataset

Background

Protein interactions are crucial in many biological and cellular processes (Jones & Thornton, 1996), such as transcription, signal transduction or enzymatic activity. Proteins, through their binding interfaces, interact with each other and a variety of other molecules, giving rise to all manner of cell functions. Knowledge about these interfaces provides essential clues about the mechanisms underlying associated activities. Experimental identification of interface residues is a time-consuming, costly and challenging task, while protein sequence data are ubiquitous. Consequently, many computational and machine learning approaches have been developed over the years to predict such interface residues from protein sequence.

Data

We provide a dataset (`ppi.csv`) that contains information of the protein bindings of PPI at a residue level, using only information related to sequence. It consists of 65150 residues that are part of 228 different proteins.

The dataset was obtained from the PDB database, firstly downloading 138729 protein structures of 2.5 A resolution or lower, excluding fragments, from the PDB (Berman et al., 2000). Following the annotation criterion in BioLip (Yang et al., 2012), a residue was annotated as interacting if the distance between one of its atoms and any atom of the ligand was less than the sum of their Van der Waals radii plus 0.5 A. Residues were mapped between PDB and Uniprot sequences by alignment with harsh penalties. BLASTClust was used to cluster the obtained Uniprot sequences at 25% sequence similarity. Proteins having sequences longer than 700 and shorter than 26 amino acids were removed.

In total, 6 residue feature types were used: Position Specific Scoring Matrix (PSSM) using PSIBLAST (Altschul et al., 1997), accessible surface area (RSA and ASA) predicted from NetSurfP, prediction of secondary structure, domain using Pfam database (Mistry et al., 2021), and the length of query sequence. Moreover, four windowed aggregates for each of the features are included (see further below). These additional features allow us to consider the correlation and dependency from the neighboring residues for each amino acid in a protein. This leads to a total of 128 residue features, specified below (a more detailed description of the dataset can be found **here**):

- **Rlength**: length of the protein sequence it belongs to.
- **sequence**: amino acid type
- **pssm_{amino acid type}**: Position Specific Scoring Matrix (PSSM) score for the amino acid type (to measure its evolutionary conservation).
- **normalized_abs_surf_acc**: predicted normalized absolute solvent accessibility (ASA)
- **rel_surf_acc** : predicted relative solvent accessibility (RSA). It is the predicted solvent accessible surface of a given amino acid in the input sequence divided by the maximal possible solvent accessible surface area of that amino acid.
- **prob_helix** : probability score for α-helix.

- **prob_sheet** : probability score for β-sheet.
- **prob_coil** : probability score for coil.
- **{n}_wm_{feature}** : n^{th} mean windowing approach for the local features (PSSM, RSA, ASA, and secondary structure). The window size (n) is the amount of neighbouring residues (3, 5, 7 or 9), corresponding to anterior and posterior positions, that are taken into account to calculate the unweighted mean value .
- **p_interface** : target binary variable that defines whether the residue is part of a known interface with other proteins.
- **uniprot_id** (not to be treated as input feature): protein ID to which the residue belongs to.
- **aa_ProtPosition** (not to be treated as input feature): residue's position in the protein it belongs to.
- **domain** (not to be treated as input feature): binary variable that defines whether a residue belongs to a protein domain ({1}) or not ({0}).

Task

The task consists on predicting the protein bindings of PPI at residue level, using only information related to protein sequence. Therefore, classifying whether a residue with certain characteristics (sequence-derived input features) is an interface amino acid.

These can be done by applying preprocessing techniques as you see fit and building model/s that suit best the task chosen. You can use the website to get insights about the dataset and/or make your research question more challenging. For the research question, originality will be rewarded. As master students, we expect an adequate sequence-based prediction of protein properties using machine learning.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235–242.

Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, *93*(1), 13–20.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., ... others (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412–D419.

Yang, J., Roy, A., & Zhang, Y. (2012). Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, *41*(D1), D1096–D1103.