2nd ML4NGP MEETING ON MACHINE LEARNING AND NON-GLOBULAR PROTEINS
May 14 - 17, 2024 – Thessaloniki, Greece

Collective Classification Algorithms in Identifying Intrinsically Disordered Proteins within Protein-Protein Interaction Networks

Milana Grbić[1], Nenad Vilendečić [1], Milan Predojević [1], & Dragan Matić[1]
[1]Faculty of Natural Science and Mathematics, University of Banja Luka

Introduction

Intrinsically Disordered Proteins (IDPs) are vital for cellular functions like transcriptional regulation, translation transcriptional regulation, translation, and cell cycle control [1]. Unlike regular proteins, IDPs lack stable structures, which allows them to act as central hubs in protein-protein interaction (PPI) networks, crucial for signaling pathways. Traditional methods used for IDP classification primarily rely on information obtained from secondary structures or amino acid sequences. Integrating data from PPI networks can improve IDP classification as PPI networks leverages the rich information contained in protein interactions which can enhance biological relevance of classification results. In this research, we combine data from different sources, including PPI networks and protein sequences, and test accuracy of classification models.

Data, Tools, & Resources

The data used in this study pertain to the model organism *S. cerevisiae* - yeast.

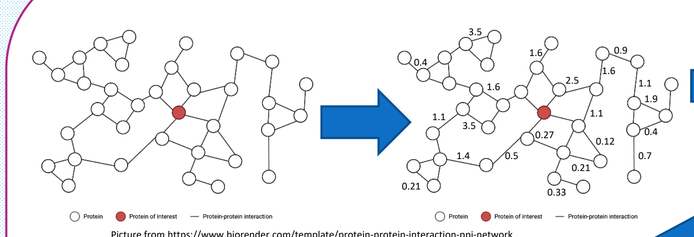
- The Protein-Protein Interaction (PPI) network – BioGRID [2] ,
- IDPs from the DisProt database [3-5],
- Gene expression information obtained by SPELL engine [6],
- Protein sequences [7],

For data extraction and classification the following tools and resources were used:

- Node2vec+ tool [8] – for extraction of features from weighted network, based on random walks,
- SMOTEEN [9] – for sampling training set,
- MinMaxScaler [10] – for normalization of dataset,
- GridSearchCV [10] – for tuning hyperparametars of KNN.

Methodology

Data preparation



Picture from https://www.biorender.com/template/protein-protein-interaction-ppi-network

128 features obtained by node2vec+

<YDR143C, $v_{1,1}$,$v_{1,2}$,…,$v_{1,128}$>
<YER068W, $v_{2,1}$,$v_{2,2}$,…,$v_{2,128}$>
…
<YMR207C, $v_{n,1}$,$v_{n,2}$,…,$v_{n,128}$>

60 features extracted from protein sequences

<YDR143C, $a_{1,1}$,…$a_{1,10}$, $ba_{1,1}$,…, $ba_{1,10}$,…,$be_{1,1}$,…,$be_{1,10}$>
<YER068W, $a_{2,1}$,…$a_{2,10}$, $ba_{2,1}$,…, $ba_{2,10}$,…,$be_{2,1}$,…,$be_{2,10}$ >
…
<YMR207C, $a_{n,1}$,…$a_{n,10}$, $ba_{n,1}$,…, $ba_{n,10}$,…,$be_{n,1}$,…,$be_{n,10}$ >

For each pair (P,Q) of proteins in the PPI network, the Adjusted Correlation Score (ACS) is calculated.
ACS is a measure of weighted correlation for the genes corresponding to the considered proteins P and Q by using the SPELL engine [6].

Calculate PPI weights

Determining the features based on information about amino acids in the context of IDPs involves consideration of the following properties [11]:
A) Order/disorder promoting amino acids and,
B) five physicochemical properties:
BA) Aromatic/Aliphatic
BB) Polar/Non-Polar
BC) Non-Zero/Zero
BD) Hydrophobic/Hydrophilic
BE) Positive/Negative

Extracting features from protein sequences

For each property *prop* in {A, BA-BE}:
For a protein sequence, $S = \{p_1, p_2, p_3,\ldots,p_N\}$, where $p_1, p_2, \ldots, p_N$ are the successive residues, calculate the binary sequence, *F(S)* through an indicator function *f*, as $F(S)=\{f(p_1),f(p_2),\ldots, f(p_N)\}$ where, $f(p_j) = 1$ if p_j has the property *prop*.
Calculate two inter-arrival distances (IADs) array for successive residues i.e. determine the distance between two successive residues: (*array1*) distances between ones, and (*array2*) distances between zeros.
For each IAD array:
Build the frequency histogram based on values of the IAD;
Construct the histogram with five intervals chosen in advanced;
Convert the frequency histogram to a probability distribution using standard statistical procedures;
Derive 5 probability values from the frequency histogram;
Constitute the final array of 10 features based on the property *prop* (total 60 features).

Data classification

Form training and test sets 70:30 stratified → Resample the training set by using SMOTEEN → Scale data MinMaxScaler → Adjust hyperparameters of KNN by using GridSearchCV → Train the model on the resampled training set → Evaluate the model

Results

Method		Node2vec+		Node2vec+ With A features		Node2vec+ With B features		Node2vec+ With A and B features	
F1 for non IDP		0.93		0.91		0.92		0.90	
F1 for IDP		0.20		0.18		0.19		0.19	
Confusion matrix		non IDP	IDP	non IDP	IDP	non IDP	IDP	non IDP	IDP
	non IDP	1414	173	1356	231	1369	218	1333	254
	IDP	32	26	30	28	29	29	26	32

Conclusion

Based on the provided table, it is evident that combining different groups of attributes yields similar results. The highest F1 score is attained when exclusively employing network-derived attributes (Node2vec+), whereas utilizing all the considered attributes predicts the most intrinsically disordered proteins (IDPs).
Preliminary findings suggest that integrating attributes from both network and sequence has potential, opening avenues for further methodological improvements. In order to further investigate the capability of this approach, it should be applied on other networks of different organisms, including human networks. Additionally, combining existing attributes with those derived from other protein characteristics could be a promising direction for future research.

References

[1] P. Tompa “Intrinsically disordered proteins: a 10-year recap”. Trends in Biochemical Sciences, 2012, 37(12), pp. 509–516.
[2] T. Nepusz, H. Yu and A. Paccanaro “Detecting overlapping protein complexes in protein-protein interaction networks.”, Nature Methods, 9(5), 471-472, 2012.
[3] F. Quaglia, et al. “DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation”. Nucleic Acids Research, 2022, 50.D1: D480-D487.
[4] A. Hatos, et al. “DisProt: intrinsic protein disorder annotation in 2020”. Nucleic acids research, 2020, 48.D1: D269-D276.
[5] D. Piovesan, et al. “DisProt 7.0: a major update of the database of disordered proteins”. Nucleic acids research, 2017, 45.D1: D219-D227.
[6] M.A. Hibbs, et al. “Exploring the functional landscape of gene expression: directed search of large microarray compendia”. Bioinformatics, 2007, 23 (20), 2692–2699.
[7] https://www.yeastgenome.org/ (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.
[8] R. Liu, A. Krishnan, “PecanPy: a fast, efficient and parallelized Python implementation of node2vec”. Bioinformatics, 2021, 37(19), pp. 3377–3379,
[9] G. E. Batista, R. C. Prati, and M. C. Monard “A study of the behavior of several methods for balancing machine learning training data”. ACM SIGKDD explorations newsletter, 2004, 6(1), 20-29.
[10] Pedregosa et al. “Scikit-learn: Machine Learning in Python”, JMLR 12, 2011, pp. 2825-2830.
[11] D. Chaurasiya, R. Mondal, T. Lahiri, A. Tripathi, T. Ghinmine, “IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy”. Journal of Biomolecular Structure and Dynamics, 2023, 1-9.

This poster is based upon work from COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology).

ML4NGP

Funded by the European Union

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY