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

ML4NGP

Funded by the European Union

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

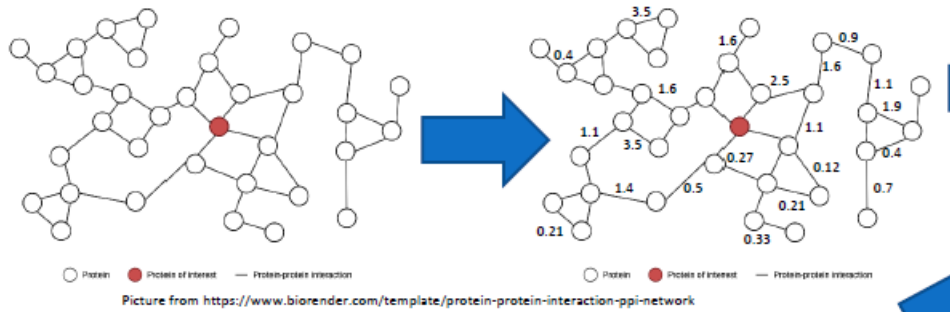
Introduction

- Traditional methods used for IDP classification?
- Integrating data from PPI networks?
- In this research?

Data, Tools, & Resources

- The data used in this study pertain to the model organism *S. cerevisiae* - yeast.
  - The Protein-Protein Interaction (PPI) network – BioGRID,
  - IDPs from the DisProt database,
  - Gene expression information obtained by SPELL engine,
  - Protein sequences,
- For data extraction and classification the following tools and resources were used:
  - Node2vec+ tool – for extraction of features from weighted network, based on random walks,
  - SMOTEEN – for sampling training set,
  - MinMaxScaler – for normalization of dataset,
  - GridSearchCV – 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

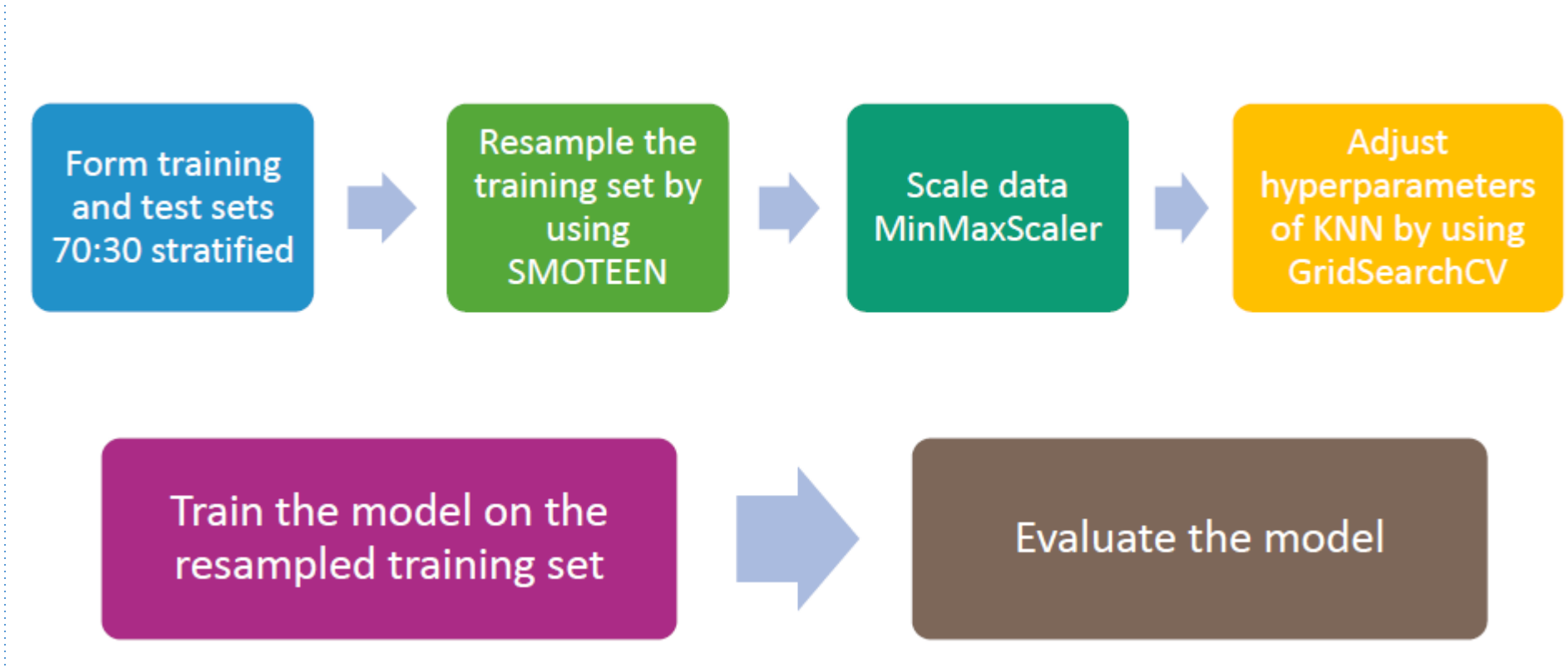
Determining the features based on information about amino acids in the context of IDPs involves consideration of the following properties [11]:
A) Order/disorder promoting amino acids and,
B) five physicochemical properties:
- BA) Aromatic/Aliphatic
- BB) Polar/Non-Polar
- BC) Non-Zero/Zero
- BD) Hydrophobic/Hydrophilic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

Methodology - Data classification



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

Thank you!

- Conclusion?
- Future work?
- Poster session I Poster F05!