Exploring the usefulness of graph properties in protein protein interaction predictions

Neven Šumonja, Nevena Veljković and Vladimir Perović*

Center for Multidisciplinary Research and Engineering
Institute of Nuclear Sciences VINCA
Belgrade, Serbia

* vladaper@vinca.rs

Abstract

A proteome-scale map of the protein-protein interactions (PPI) enables a global view of cellular function and organization.

However, experimentally validated PPI information is sparse and biased in favor of well-studied proteins and therefore computational prediction methods add a value by prioritizing candidates and covering unexplored zones of the interactome. In this work we tested graph measures commonly used in network analysis as a new type of numerical features for encoding. In our model each protein is perceived as a node in PPI network and it is numerically encoded by selected graph properties. The encouraging evaluation results on human and yeast test sets suggest that this distinct class of attributes have considerable potential to improve methods that utilize machine learning algorithms and perform binary classification on interacting and non-interacting protein pairs.

Dataset

For the training process we used 40 balanced training sets of 28K HUMAN interactions, and 40 sets of 14K YEAST protein interactions, selected from protein interaction network analysis platform (PINA) database. For the testing process, we used 40 corresponding test sets of C1 class (both proteins in test pair are in training set) of 3K HUMAN testing pairs respectively. Same was done for YEAST interactions.

Features

PPI graph measures/metrics are used to encode the proteins

•**Authorityscore**	Kleinberg's authority centrality scores
•**Betweenness**	Vertex betweenness centrality
•**Centrclo**	Centrality score
•**Closeness**	Closeness centrality of vertices
•**Clusterfastgreedy**	Community structure via greedy optimization of modularity
•**Clusterwalktrap**	Community structure via short random walks
•**Constraint**	Burt's constraint
•**Coreness**	K-core decomposition of graphs
•**Degree**	Degree distribution of the vertices
•**Eccentricity**	Eccentricity of the vertices in a graph
•**Knn**	Average nearest neighbor degree
•**Localscan**	Local scan statistics
•**Maxcardinality**	Maximum cardinality search
•**Pagerank**	The Page Rank algorithm

Results

The method **GreFePPI**, which is based on graph properties of the PPI network, is tested and compared with six state of the art methods: M1 (Martin et al., 2005) M2 (Guo et al., 2008), M3 (Pitre et al., 2008), M4 (Shen et al., 2007), M5 (Park & Markotte, 2012), M6 (Hamp & Rost, 2015) on C1 test classes. The average values of AUCs for 40 HUMAN and 40 YEAST test sets, given in Table 1, show that GraFePPI method has the best prediction performance.

Additionally, comparison between different GraFePPI ML algorithms, given on Tables 2 and 3, and Figure 1, shows that Random forest has best performance on yeast sets, while Ensemble method outperformed others on human sets.

Method	AUC (HUMAN)	AUC (YEAST)
M1	0.81 ± 0.01	0.82 ± 0.01
M2	0.77 ± 0.01	0.76 ± 0.02
M3	0.77 ± 0.01	0.75 ± 0.02
M4	0.64 ± 0.01	0.61 ± 0.01
M5	0.85 ± 0.01	0.84 ± 0.01
M6	0.87 ± 0.01	0.87 ± 0.02
GraFePPI	0.88 ± 0.01	0.90 ± 0.01

Table 1. Performance statistics on 40 HUMAN and 40 YEAST C1 class test sets. AUROC - Area under the ROC curve.

	GraFePPI-RF	GraFePPI-GBM	GraFePPI-GLM	GraFePPI-DPL	GraFePPI-ENS
AUROC	0.866 ± 0.006	0.863 ± 0.007	0.869 ± 0.006	0.869 ± 0.006	0.878 ± 0.006
AUPR	0.872 ± 0.007	0.872 ± 0.007	0.868 ± 0.007	0.870 ± 0.007	0.881 ± 0.007
ACC	0.789 ± 0.007	0.791 ± 0.006	0.796 ± 0.006	0.796 ± 0.006	0.802 ± 0.007
F	0.780 ± 0.009	0.782 ± 0.009	0.792 ± 0.007	0.793 ± 0.007	0.799 ± 0.008
Precision	0.813 ± 0.018	0.818 ± 0.014	0.807 ± 0.013	0.807 ± 0.015	0.811 ± 0.012
Specificity	0.826 ± 0.026	0.832 ± 0.019	0.814 ± 0.019	0.812 ± 0.021	0.816 ± 0.018
Recall	0.751 ± 0.026	0.750 ± 0.022	0.777 ± 0.020	0.780 ± 0.020	0.788 ± 0.020
MCC	0.579 ± 0.014	0.585 ± 0.012	0.592 ± 0.012	0.593 ± 0.012	0.605 ± 0.013

Table 2. Performance statistics on 40 HUMAN C1 class test sets of different GraFePPI ML algorithms. AUROC - Area under the ROC curve. AUPRC - Area under the Precision-Recall curve, MCC - Matthews correlation coefficient.

	GraFePPI-RF	GraFePPI-GBM	GraFePPI-GLM	GraFePPI-DPL	GraFePPI-ENS
AUROC	0.900 ± 0.008	0.868 ± 0.01	0.849 ± 0.013	0.852 ± 0.011	0.885 ± 0.010
AUPR	0.910 ± 0.009	0.879 ± 0.01	0.856 ± 0.016	0.859 ± 0.013	0.896 ± 0.008
ACC	0.829 ± 0.009	0.801 ± 0.011	0.776 ± 0.012	0.777 ± 0.011	0.815 ± 0.010
F	0.821 ± 0.011	0.790 ± 0.014	0.767 ± 0.016	0.770 ± 0.017	0.809 ± 0.012
Precision	0.862 ± 0.020	0.834 ± 0.021	0.800 ± 0.028	0.796 ± 0.023	0.837 ± 0.023
Specificity	0.873 ± 0.024	0.850 ± 0.027	0.812 ± 0.041	0.806 ± 0.036	0.846 ± 0.029
Recall	0.785 ± 0.025	0.752 ± 0.031	0.740 ± 0.043	0.749 ± 0.042	0.784 ± 0.029
MCC	0.662 ± 0.018	0.605 ± 0.021	0.555 ± 0.023	0.557 ± 0.022	0.632 ± 0.021

Table 3. Performance statistics on 40 YEAST C1 class test sets of different GraFePPI ML algorithms

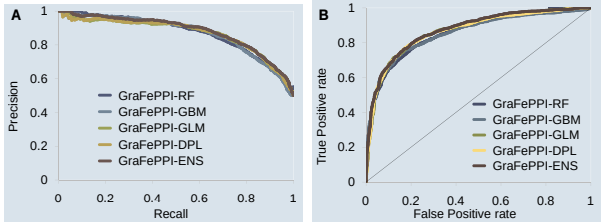


Figure 1. (A) ROC and (B) recall-precision plots for GraFePPI methods tested on human sets

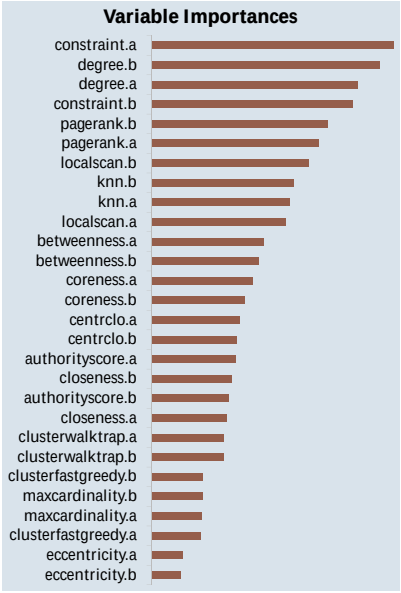
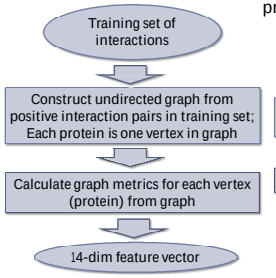


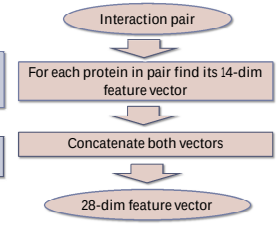
Figure 2. Variable importance measures computed by random forest algorithm

Algorithm

Calculating **GraFePPI** protein features



Generating **GraFePPI** feature vector for protein pair (if both of the proteins are in training set – C1 class)



Machine Learning

For generating of PPI classifiers, we have used different machine learning (ML) methods along with hyperparameter optimization.

According to used ML algorithms we have developed five methods

- **GraFePPI-RF** Random Forest
- **GraFePPI-GBM** Gradient Boosted Method
- **GraFePPI-GLM** Generalized Linear Model
- **GraFePPI-DPL** Deep Learning
- **GraFePPI-ENS** Ensemble of ML methods

Conclusion

GraFePPI is the novel method for protein and PPI representation which is based only on the information in PPI network. It is effective for prediction of PPIs and outperforms the state of the art methods in terms of prediction potential. It is also capable of efficiently analyzing large number of protein pairs.

Acknowledgments

This work was supported by the Ministry of Education and Science Republic of Serbia grant no. 173001.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 63(1):3-42. (2006)
Friedman HJ. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics: 1189-1232. (2001)
Deng Li & Dong Yu. Deep learning: methods and applications. Foundations and Trends in Signal Processing 7.3-4: 197-387. (2014)
Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 33:1. (2010)
Martin S, Roe D & Faulon JL. Predicting protein-protein interactions using signature products. Bioinformatics, 21(2), 218-226. (2005).
Shen J, Zhang J, Luo X, Zhu W, et al. Predicting protein-protein interactions based only on sequences information. PNAS, 104(11), 4337-41. (2007).
Guo Y, Yu L, Wen Z & Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Research 36(9): 3025-30. (2008).
Pitre S, North C, Alamgir M, Jessulat M, et al. Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. Nucleic Acids Research, 36(13), 4286-4294. (2008).
Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nature methods. 9(12):1134. (2012)
Hamp T & Rost B. Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics. 4;31(12):1945-50. (2015)
Csardi G, Nepusz T. The igraph software package for complex network research, InterJournal, Complex Systems 1695. http://igraph.org. (2006)