RESEARCH

Prediction of Protein-Protein Interactions on the Human and Rice Interactomes

Nicolas A. Lopez-Rozo*, Jorge Finke and Camilo Rocha

*Correspondence:
nicolaslopez@javerianacali.edu.co
Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali, CO
Full list of author information is available at the end of the article

Abstract

Background: Abstract A recent study in network-based prediction of protein-protein interactions (PPIs) reveals that two proteins are more likely to interact, the higher the number of paths of length 3 between them (normalized by the geometric average of their interactions). This paper extends previous work on mapping binary interactions by taking into account the learning of features (embeddings) of the PPI network. In particular, we implement a gradient boosted decision tree model (XGBoost) using handcrafted features (including the normalized measure) and embeddings from an algorithm that generates a low-dimensional representation of nodes (node2vec).

Results: Our main result shows that while the measure remains an important feature for predicting interactions, better performance is achieved when in addition embedding features are considered. The proposed approach is validated for the human and rice interactomes. For both cases, the combination of both types of features yield higher AUC values.

Conclusions: As found on this study on both human and rice, when information from handcrafted features based on neighborhood is enhanced with vector representations from random walks, the prediction power of the model improves. Besides, a supervised learning model can be trained for predicting unknown interactions based on such information. Finally, the developed framework can also be applied to interactomes of other organisms for which PPI networks have recently become available.

Keywords: PPI; Prediction; Protein Interaction; Machine Learning; Python; XGBoost; node2vec

Introduction

Proteins are key actors of biological processes inside cells. Rather than carrying out tasks as single agents, they are part of dynamic networks of protein-protein interactions (PPI) [1]. Such networks underlie a variety of interdependent mechanisms, including signal transduction, homeostasis control and stress responses. Furthermore, PPI networks play an important role in physiological and developmental processes such as protein phosphorylation, transcriptional co-factor recruitment and transporter activation [2].

A common way to create PPI networks (or validate particular protein-protein interactions) is the *Yeast-Two-Hybrid* technique (also known as *two-hybrid screening* or *Y2H*). Figure 1A illustrates the biological basis of Y2H: the expression of a specific reporter gene is activated by the binding of a DNA-binding Domain (DB) and an Activation Domain (AD) of a Transcription Factor, which in turn binds

to an Upstream Activation Sequence (UAS). To evaluate an interaction between two proteins, the Y2H approach fuses one protein to the DB domain (known as *bait*) and another protein to the AD (known as *prey*). If the proteins interact, the reporter gene expression is activated by the AD (Fig. 1B). Otherwise, if proteins fail to interact, the reporter gene is not expressed (Fig. 1C).

Figure 1 The Yeast-2-Hybrid technique offers an experimental approach for constructing PPI networks.

Based on the outcome of numerous Y2H experiments, undirected networks of interactions between proteins can be constructed. However, relying only on experimental validation of PPIs is detrimental to research due to its hindrance in costs, accuracy, required manpower and time [3, 4]. Besides, the number of missing interactions for pairs of proteins on many organisms leverage the usage of computational methods for prediction of such interactions.

A common way of predicting interactions is usually based on social networks analysis, more specifically on the triadic closure principle (TCP). TCP states that the higher the number of common neighboring nodes between two nodes, the higher the probability that they interact [5]. TCP can be addressed mathematically by counting the number of shared neighbors of a pair of nodes, also known as the Common Neighbors algorithm. By raising the adjacency matrix (A) of the network to the second power ($\mathbf{A^2}$), TCP can be considered for further analyses. However, previous studies show that the mentioned approach usually fails because it does not consider the structural and chemical properties of the proteins [6, 7].

A variety of methods for predicting interactions in PPI networks have been proposed in recent years [8, 9, 10]. Kovacs et al (2019) introduce a network-based approach which predicts the interaction between two proteins based on the number of paths of length 3, normalized by the geometric average of their interactions. The simplest mathematical representation of this principle consists on the third power of the adjacency matrix ($\mathbf{A^3}$). Furthermore, Kovacs et al present a degree-normalized scaling for this metric, which reduces bias caused by intermediate hubs within the paths of length 3. This handcrafted measure, denoted L3, enables the proposed approach to outperform previous methods for predicting binary protein interactions in yeast (*S. cerevisiae*), Arabidopsis (*A. thaliana*), worm (*C. elegans*), fly (*D. melanogaster*), fission yeast (*S. pombe*), mouse (*M. musculus*) and humans [7].

The focus of this study is to evaluate different methods for predicting PPIs using the network of known interactions. To achieve this, human and rice PPI networks are compared using the proposed methods (A2, A3, L3), as well as a low-dimensional representation of nodes (`Node2Vec`)[11]. After that, combinations of the two types of methods are tested. In the case of the human network, two different versions of the human interactome [12] are used for training and another interactome version is used for validation (*HI-III*)[7]. For the rice interactome, a portion of the known interactions removed and then predicted with different techniques based on the known interactions. The main contributions of this paper are i) a general framework for link prediction in non-directed networks and ii) two applications of this framework to biological networks with a structural insight.

Structure: This paper is organized as follows. *Materials and Methods* describes the methodological steps and key milestones in the preparation of the networks, model parameters and experimental configurations. *Results* presents the main results of the models for the human and rice interactomes. *Conclussions* presents the main conclusions of this study. Finally, *Appendix* presents supplementary information and figures which complement the results described in this paper.

Sub-heading for section

Text for this sub-heading...

Sub-sub heading for section

Text for this sub-sub-heading...

Sub-sub-sub heading for section Text for this sub-sub-sub-heading...

In this section we examine the growth rate of the mean of Z_0, Z_1 and Z_2. In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E\big[Z_1(vT_x)\big] = \int_0^{v\wedge 1} Z_0(uT_x)\exp(\lambda_1)\,du.$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0}\log x$, then we can heuristically estimate the expected value as

$$\begin{aligned} &E\big[Z_1(vT_x)\big] \\ &\quad = \frac{\mu}{r}\log x \int_0^{v\wedge 1} x^{1-u}x^{(\lambda_1/r)(v-u)}\,du. \end{aligned} \tag{1}$$

Thus we observe that this expected value is finite for all $v > 0$ (also see [13, 14, 15, 16, 17, 18]).

Appendix

Text for this section...

Acknowledgements
Text for this section...

Funding
Text for this section...

Abbreviations
Text for this section...

Availability of data and materials
Text for this section...

Ethics approval and consent to participate
Text for this section...

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Text for this section. . .

Authors' contributions
Text for this section . . .

Authors' information
Text for this section. . .

Author details
Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali, CO.

References

1. Lin, J.-S., Lai, E.-M.: Protein–protein interactions: Co-immunoprecipitation. In: Journet, L., Cascales, E. (eds.) Bacterial Protein Secretion Systems: Methods and Protocols, pp. 211–219. Springer, New York, NY (2017)
2. Zhang, Y., Gao, P., Yuan, J.: Plant protein-protein interaction network and interactome. Current Genomics **11**(1), 40–46 (2010). doi:10.2174/138920210790218016
3. Laraia, L., McKenzie, G., Spring, D.R., Venkitaraman, A.R., Huggins, D.J.: Overcoming chemical, biological, and computational challenges in the development of inhibitors targeting protein-protein interactions. Chemistry and biology **22**, 689–703 (2015)
4. Macalino, S.J.Y., Basith, S., Clavio, N.A.B., Chang, H., Kang, S., Choi, S.: Evolution of In Silico Strategies for Protein-Protein Interaction Drug Discovery. Molecules (Basel, Switzerland) **23** (2018)
5. Goldberg, D.S., Roth, F.P.: Assessing experimentally derived interactions in a small world. Proceedings of the National Academy of Sciences of the United States of America **100**, 4372–6 (2003)
6. Cannistraci, C.V., Alanis-Lobato, G., Ravasi, T.: From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific reports **3**, 1613 (2013)
7. Kovács, I.A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.-K., Kishore, N., Hao, T., Calderwood, M.A., Vidal, M., Barabási, A.-L.: Network-based prediction of protein interactions. Nature Communications **10**(1) (2019). doi:10.1038/s41467-019-09177-y. https://doi.org/10.1038/s41467-019-09177-y
8. Chang, J.-W., Zhou, Y.-Q., Ul Qamar, M.T., Chen, L.-L., Ding, Y.-D.: Prediction of protein-protein interactions by evidence combining methods. International journal of molecular sciences **17** (2016)
9. Chen, K.-H., Wang, T.-F., Hu, Y.-J.: Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC bioinformatics **20**, 308 (2019)
10. Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., Broackes-Carter, F., Petschnigg, J., Mills, G.B., Jurisicova, A., Stagljar, I., Maestro, R., Jurisica, I.: In silico prediction of physical protein interactions and characterization of interactome orphans. Nature Methods **12**(1), 79–84 (2015)
11. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, ??? (2016). doi:10.1145/2939672.2939754
12. Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S.D., Yang, X., Ghamsari, L., Balcha, D., Begg, B.E., Braun, P., Brehme, M., Broly, M.P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B.J., Hardy, M.F., Jin, M., Kang, S., Kiros, R., Lin, G.N., Luck, K., MacWilliams, A., Menche, J., Murray, R.R., Palagi, A., Poulin, M.M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J.M., Scholz, A., Shah, A.A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A.O., Trigg, S.A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M.E., Xia, Y., Barabasi, A.-L., Iakoucheva, L.M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M.A., Hill, D.E., Hao, T., Roth, F.P., Vidal, M.: A proteome-scale map of the human interactome network. Cell **159**, 1212–1226 (2014)
13. Koonin, E.V., Altschul, S.F., Bork, P.: Brca1 protein products: functional motifs. Nat. Genet. **13**, 266–267 (1996)
14. Jones, X.: Zeolites and synthetic mechanisms. In: Smith, Y. (ed.) Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore, pp. 16–27 (1996)
15. Margulis, L.: Origin of Eukaryotic Cells. Yale University Press, New Haven (1970)
16. Schnepf, E.: From prey via endosymbiont to plastids: comparative studies in dinoflagellates. In: Lewin, R.A. (ed.) Origins of Plastids, 2nd edn., pp. 53–76. Chapman and Hall, New York (1993)
17. Kohavi, R.: Wrappers for performance enhancement and obvious decision graphs. PhD thesis, Stanford University, Computer Science Department (1995)
18. ISSN International Centre: The ISSN register (2006). http://www.issn.org Accessed Accessed 20 Feb 2007

Figures

Figure 2 Sample figure title

Figure 3 Sample figure title

Table 1 Sample table title. This is where the description of the table should go

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Tables

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.