

SOFTWARE

Open Access



IMMAN: an R/Bioconductor package for Interolog protein network reconstruction, mapping and mining analysis

Minoo Ashtiani^{1†}, Payman Nickchi^{1†}, Soheil Jahangiri-Tazehkand^{1,3†}, Abdollah Safari^{2*}, Mehdi Mirzaie^{4*} and Mohieddin Jafari^{1,5*} 

Abstract

Background: Reconstruction of protein-protein interaction networks (PPIN) has been riddled with controversy for decades. Particularly, false-negative and -positive interactions make this progress even more complicated. Also, lack of a standard PPIN limits us in the comparison studies and results in the incompatible outcomes. Using an evolution-based concept, i.e. interolog which refers to interacting orthologous protein sets, pave the way toward an optimal benchmark.

Results: Here, we provide an R package, IMMAN, as a tool for reconstructing Interolog Protein Network (IPN) by integrating several Protein-protein Interaction Networks (PPINs). Users can unify different PPINs to mine conserved common networks among species. IMMAN is designed to retrieve IPNs with different degrees of conservation to engage prediction analysis of protein functions according to their networks.

Conclusions: IPN consists of evolutionarily conserved nodes and their related edges regarding low false positive rates, which can be considered as a gold standard network in the contexts of biological network analysis regarding to those PPINs which is derived from.

Keywords: Protein-protein interaction networks (PPINs), Interolog protein network (IPN), Bioconductor, Network biology

Background

Nowadays, tremendous amount of interactions at the molecular level have been accessible by the development of the technology, endeavors to model cellular and molecular processes [1, 2]. Among these interactions, protein-protein interactions (PPIs) are remarkable due to providing functional and structural description of executive molecules i.e. proteins [3]. Nevertheless, PPI detection and prediction technologies are still entangling

with reducing false-positive and -negative interactions [4–6]. Accordingly, data integration is the best solution overall in spite of the improvement of experimental and computational methods. STRING [7], BioNetBuilder Cytoscape app [8], IMP 2.0 [9], PINALOG [10], HIPPIE [11] and BIPS [12] are using this solution to reconstruct and refine PPI networks (PPINs). In the other works, an evolutionarily conserved network with communal nodes and less false-positive links, Interolog Protein Network (IPN), was introduced as a benchmark for the evaluation of clustering algorithms [13]. IPN clears up the arisen and remained interactions during the evolution and helps to excavate the remnants of ancestor PPIN [13–17]. In this study, we present IMMAN, a package to integrate several PPINs and mine IPNs. IMMAN is free and is available as an R/Bioconductor package and also a Java program.

* Correspondence: a.saffari72@gmail.com; mirzaie@modares.ac.ir; mohieddin.jafari@helsinki.fi

[†]Minoo Ashtiani, Payman Nickchi and Soheil Jahangiri-Tazehkand contributed equally to this work.

²Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

⁴Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran

¹School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Full list of author information is available at the end of the article



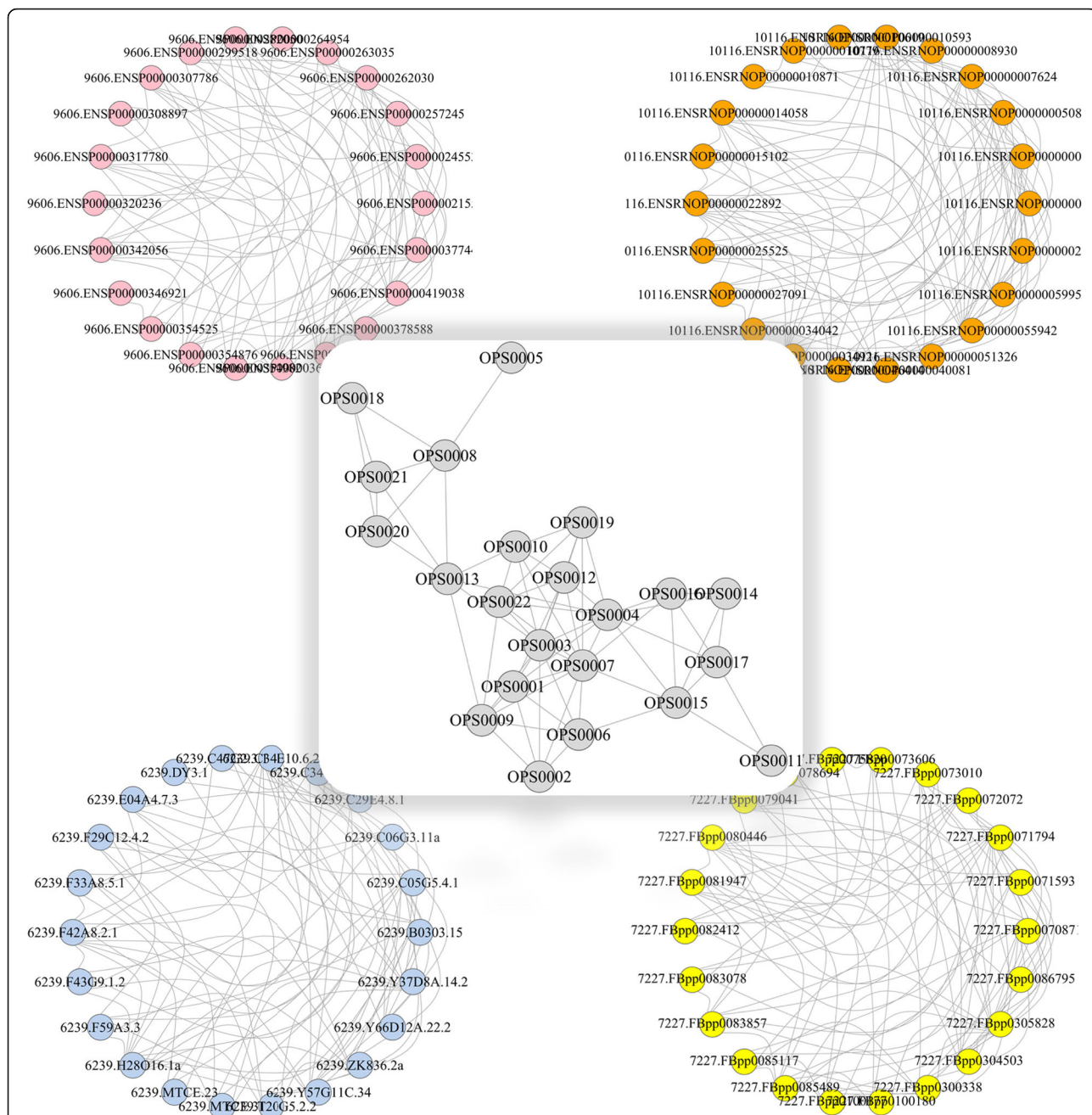


Fig. 1 The IPN derived from four PPINs of sample species named; *H. sapiens* (top-left), *M. musculus* (top-right), *D. melanogaster* (bottom-left) and *C.elegans* (bottom-right). The size of IPN is proportional to evolutionary distance of selected species. The IPN edges is less than or equal to the smallest related PPINs. The IPN nodes are orthologous set of proteins which is abbreviated as OPS and STRINGdb IDs were used to label nodes of PPINs

Implementation

IMMAN enables users to define two to four arbitrarily lists of proteins (by UniProt accession number) as inputs, and seek for evolutionarily conserved interactions in the integrated PPIN or IPN as the output. Briefly speaking, the method takes the following steps to accomplish this goal.

- Step 1. First, the amino acid sequence of each protein of input list is automatically retrieved from UniProt database.
- Step 2. In the second step, IMMAN infers the orthologous proteins. To this end, the Needleman-Wunsch algorithms is employed to compute the pairwise sequence similarities. The reciprocal best

hits are retrieved and applied in the next step to increase the chance of discovering the orthologous pairs. The user can adjust different parameters of alignment algorithm as well as the sequence similarity cutoff for orthology detection.

- Step 3. In this step, the nodes of the IPN are specified. Each node of the network is defined as a set of mutually orthologous proteins (OPS) such that each OPS belongs to a set of species involved in the analysis.
- Step 4. In the fourth step, for each species, the PPINs are singly extracted according to the proteins constitute the OPSs or IPN nodes. The PPINs are retrieved from STRING database. Next, the user can adjust the minimal confidence score of STRING networks.
- Step 5. Finally, the edges of the interolog network are extracted. To this end, for every OPS pair, the number of protein pairs (p_{ik}, p_{jk}) are considered such that p_i and p_j are connected in the PPIN of the species k . If this number exceeds a predefined cutoff (coverage cutoff), there would be an edge between the aforementioned nodes. The coverage cutoff can be also specified by the user to tune conservedness.

Results

After running IMMAN, the node list and the edge list of inferred IPN is produced. Additionally, IMMAN outputs the graphical representation of the network. The graphical output of IMMAN are produced using GraphViz [18] and igraph [19] in Java and R applications, respectively. The graphical representation of IMMAN on a sample dataset is depicted in Fig. 1. In this figure, the IPN derived from the original four different PPINs (Node No. \cong 30) related to *H. sapiens*, *M. musculus*, *D. melanogaster* and *C. elegans* is represented. The resulting network contains 23 nodes and 97 to 66 edges depends on coverage parameters. Note that the higher coverage cutoff results in more stringent and conserved network. The sample dataset is available in Additional file 1.

Conclusions

Although, the size of IPN is tunable by several thresholds, but obviously, missing the edges in IPN is the cost of true positive discovery which is an ideal within PPI studies with inherent inconsistency [6, 20]. However, function prediction is a prominent question in molecular biology and this approach pave its way based on evolutionary mechanism [21]. All routine analysis of network biology related to PPIN become more reliable by the study of IPN. For instance, finding modules within the IPN help us to understand how evolution thinks, provides and preserves cellular mechanism of species to

characterize a given biological process [13]. Also, ranking the node's influence of IPN, based on centrality measures, can shed light on the detailed mechanism of evolutionary processes [22].

Additional file

Additional file 1: Example lists containing the UniProt accession number of four different species. (RAR 843 bytes)

Abbreviations

IPN: Interolog Protein Network; PPI: Protein-Protein Interaction; PPIN: Protein-protein Interaction Network

Acknowledgements

The authors would like to thank Dr. Mehdi Sadeghi for his valuable comments and discussions.

Funding

This work has been supported by the grant number No. BS 1395_0_01 provided by the school of biological sciences, Institute for Research in Fundamental Sciences, Tehran, Iran.

Availability of data and materials

Project name: IMMAN.
Project home page: <https://bioconductor.org/packages/IMMAN>
Archived version: 1.2.0.
Operating system(s): Platform independent.
Programming language: R.
Other requirements: None.
License: Artistic-2.0.

Authors' contributions

MJ, MM conceived and commenced the project and provided direction and feedback on the final results. PN wrote the basic R codes, gathered datasets used in the package and drafted the manuscript. AS developed and implemented the method and improved the R codes. MA built the R package IMMAN and participated in revising the code and the manuscript. SJT developed and implemented the JAVA script of the procedure. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. ²Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. ³Department of Computer Science, Shahid Beheshti University, Tehran, Iran. ⁴Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. ⁵Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland.

Received: 6 August 2018 Accepted: 28 January 2019

Published online: 12 February 2019

References

1. Rolland T, Tasan M, Charloreaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, et al. A proteome-scale map of the human Interactome network. *Cell*. 2014;159(5):1212–26.
2. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell*. 2011;144(6):986–98.
3. Srihari S, Yong CH, Wong L. Computational prediction of protein complexes from protein interaction networks: Morgan & Claypool; 2017.
4. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual J-f, Borick H, Braun P, Dreze M, et al. Literature-curated protein interaction datasets. *Nat Methods*. 2009;6(1):39–46.
5. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biology*; 2006.
6. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, et al. An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*. 2009;6(1):91–7.
7. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res*. 2016;gkw937–7.
8. Avila-Campillo I, Drew K, Lin J, Reiss DJ, Bonneau R. BioNetBuilder: automatic integration of biological networks. *Bioinformatics*. 2007;23(3):392–3.
9. Wong AK, Park CY, Greene CS, Bongo L, Guan Y, Troyanskaya OG. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res*. 2012;40(Web Server issue):W484–90.
10. Phan HTT, Sternberg MJE. PINALOG: a novel approach to align protein interaction networks—implications for complex detection and function prediction. *Bioinformatics*. 2012;28(9):1239–45.
11. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, MA A-N. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*. 2012;7(2). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031826>.
12. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Res*. 2012;40(Web Server issue):W147–51.
13. Jafari M, Mirzaie M, Sadeghi M. Interlog protein network: an evolutionary benchmark of protein interaction networks for the evaluation of clustering algorithms. *BMC bioinformatics*. 2015;16(1) 319–319.
14. Jafari M, Sadeghi M, Mirzaie M, Marashi S-A, Rezaei-Tavirani M. Evolutionarily conserved motifs and modules in mitochondrial protein–protein interaction networks. *Mitochondrion*. 2013;13:7.
15. Matthews LR. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "Interologs". *Genome Res*. 2001;11(12):2120–6.
16. Nguyen PV, Srihari S, Leong HW. Identifying conserved protein complexes between species by constructing interolog networks. *BMC Bioinformatics*. 2013;14(Suppl 16) S8–S8.
17. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*. 2000;287(5450):116–22.
18. Ellison J, Gansner E, Koutsofios L, North SC, Woodhull G: Graphviz—open source graph drawing tools. In: *International Symposium on Graph Drawing: 2001*. Springer: 483–484.
19. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006;1695(5):1–9.
20. Shin CJ, Davis MJ, Ragan MA. Towards the mammalian interactome: inference of a core mammalian interaction set in mouse. *Proteomics*. 2009;9(23):5256–66.
21. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*. 2004;14(6):1107–18.
22. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, Mirzaie M, Jafari M. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst Biol*. 2018;12(1):80.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions