Application Note

Systems Biology

IMAN: Interlog protein network reconstruction, Matching and ANalysis

Mohieddin Jafari[1,3*], Payman Nickchi[1], Abdollah Safari[2], Soheil Jahangiri Tazehkand [3] and Mehdi Mirzaie[4]

[1]Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69, Pasteur St, 13164, Tehran, Iran.

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

[3]School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Shahid Lavasani St, PO Box 19395-5746, Tehran, Iran.

[4]Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Jalal Ale Ahmad Highway, PO Box 14115-134, Tehran, Iran.

*Corresponding author: Tel: +98 21 6411 2466; Email: m_jafari@pasteur.ac.ir and mjafari@ipm.ir

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

IMAN is an open-source R package that offers users to reconstruct Interlog Protein Network (IPN) integrated from several Protein-protein Interaction Networks (PPIN). Users can overlay different PPINs to mine conserved common network between diverse species. Currently STRING database is applied to extract PPINs with all experimental and computational interaction prediction methods. IMAN helps to retrieve IPN with different degrees of conservation to employ for better protein function prediction and PPIN analysis.

Availability: IMAN package does not require any registration and is freely available at http://bs.ipm.ac.ir/softwares/IMAN or http://jafarilab-pasteur.com/content/software/IMAN.html.
Contact: mjafari@ipm.ir
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Nowadays, technologies have provided access to tremendous amount of interactions at the molecular level. The study of these interactions, interactome, endeavor to model cellular and molecular events (1, 2). Among these interactions, protein-protein interactions (PPI) are significant due to functional and structural description of executive molecules i.e. proteins. However, PPI detection and prediction methods are still entangling with reducing false-positive and –negative interactions (3, 4). In addition to improvement of experimental and computational methods, data integration is the best solution overall. STRING (5), BioNetBuilder Cytoscape app (6), IMP 2.0 (7), PINALOG (8), HIPPIE (9) are using this solution to reconstruct and refine PPI networks (PPIN). Recently, an evolutionarily conserved network with communal nodes and less false-positive links, Interlog Protein Network (IPN), was introduced as a benchmark for the evaluation of clustering algorithms (10). IPN clears up the arisen and remained interactions during evolution and help to excavate the remnants of ancestor PPIN (10-14). In this study, we provide a freely available R package to integrate several PPINs and retrieve IPNs.

2 Methods

The Interlog protein network reconstruction, Matching and ANalysis (IMAN) package will enable users to define any arbitrarily list of proteins with their UniProt accession number, and seek for evolutionarily conserved interactions in the integrated PPIN. This package allows us to define any list of proteins for up to a maximum of four species and takes arbitrary arguments for different alignment purposes. Briefly speaking, the method takes the following steps to accomplish this goal. First, the intra species similarity of the given lists of proteins is searched to find paralogs and prioritize the lists based on Needleman-Wunsch algorithm. Second, the proteins in each species are aligned with the proteins of the other species using the well-known Needleman-Wunsch algorithm to find orthologs. The results of this alignment are stored in a score matrix and the proteins having the identity score above a threshold are selected for further processing. These are orthologous proteins sets (OPSs) of four species and are assumed to conserve common sequence during the evolution. Third, we map the UniProt identifier of OPSs to STRING for each species individually. Then the network of these proteins is retrieved from the STRING based on the version and selected criteria which is provided by the user. The user can choose to run the algorithm for different versions of STRING and also change the cut-off value for selecting PPI in STRING database.

At the fourth stage, which is critical in conducting the final interlog network, we arrive at 2, 3, or 4 different STRING networks depending on the number of species and protein lists which user had provided at the first stage. We have also gathered OPSs which will be considered in the final calculation. If the orthologous proteins in OPSs for each of the species is detected as connected nodes in their corresponding STRING network for that specific species then there would be an edge between these two entries of OPSs in the IPN. In our algorithm, it could be possible to define number of connected proteins pairs in each OPS pair (termed as "coverage") to reconstruct an edge of them in the IPN. For more details about the algorithm please refer to (10, 15).

3 Results

In this section, we will discuss the functions, required parameters to run the algorithm, the outputs of functions, and the estimated time of running for some lists of proteins.

3.1 Package functions and the arguments

IMAN package consists of several functions responsible for matching analysis and integrating networks. On top of the R generic and user developed functions, the main functions are IMAN2, IMAN3, and IMAN4. The required parameters in each function are nearly identical and the encouraged reader is referenced to the prepared manual for further reading. Here we will briefly explain the different required parameters in each function. At the first stage, it is needed to provide 2, 3, or 4 lists of proteins when running each of the IMAN2, IMAN3, or IMAN4 functions, respectively. These lists identify the UniProt identifiers of each protein and are considered to be of character type in R statistical software. Four integer values will determine the taxonomy id of each species and should be provided as List1_Species_ID up to List4_Species_ID, respectively. An identity value (identityU) for the purpose of selecting homologous proteins should also be provided. This argument is a real value between 0 and 100. The score_threshold argument will determine the score used in the STRING database for selecting protein-protein interactions. STRINGversion will define the version of the STRING database which will be downloaded through the internet automatically. The coverage value will determine how conservative the IPN algorithm should be (Fig 1). In the other word, with more coverage we obtain more conservative and consensus IPN. InputDirectory is an arbitrary argument which will help the users to run the functions in offline mode if the algorithm has already been run for the same parameters before and the required information is downloaded.

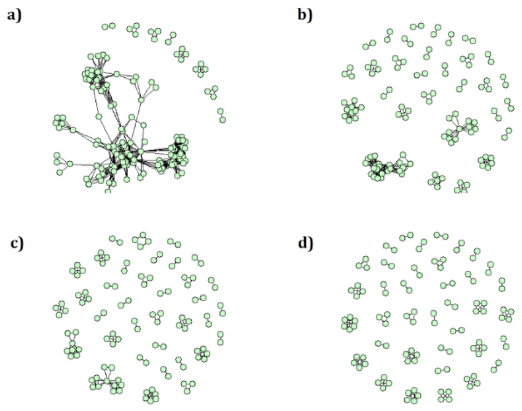

Fig1. The IPN and the effect of coverage parameter. The algorithm found the IPN for the four given lists of proteins. These lists were provided in the supplementary file. Their graph is depicted for different coverage values: a) coverage = 1, b) coverage = 2, c) coverage = 3, d) coverage = 4.

3.2 IMAN's output

The IMAN's output is the same for the 3 functions. After running each function, the following outputs are generated: an IPN edge list determining the interaction between each element in OPSs and its corresponding graph, OPS which contains the labels and proteins of each species, and based on the number of protein lists given by the user the corresponding PPIN, retrieved from STRING database, is provided as well. Supplementary File 1 provides the output for an example list.

3.3 Performance test

We ran our algorithm on lists having different number of proteins. Table1 demonstrates the required time in the minute scale in each scenario for the algorithm to generate the results on a PC equipped with Intel Core-i7 at 3.70 GHz, 32 GB of RAM. The running time will highly depend on the time that takes for the STRINGdb package to retrieve the STRING network.

Table 1. Performance result of IMAN. IMAN's functions were performed on a PC equipped with Intel Core-i7 at 3.70 GHz, 32 Gb of RAM installed, running on 64bit architecture.

IMAN	Number of Proteins in each list	Running time (min)
2	100 - 120	~ 10
3	100 - 120 - 125	~ 22
4	100 - 120 - 125 - 150	~ 35

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.

Conflict of Interest: The authors have no conflict of interest.

References

1. Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, 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. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks ? Genome Biology. 2006.
4. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, et al. Literature-curated protein interaction datasets. Nature Methods. 2009;6(1):39-46.
5. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerte-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2014:1-6.
6. Avila-Campillo I, Drew K, Lin J, Reiss DJ, Bonneau R. BioNetBuilder: automatic integration of biological networks. Bioinformatics (Oxford, England). 2007;23(3):392-3.
7. Wong AK, Park CY, Greene CS, Bongo La, Guan Y, Troyanskaya OG. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic acids research. 2012;40(Web Server issue):W484-90.
8. Phan HTT, Sternberg MJE. PINALOG: a novel approach to align protein interaction networks--implications for complex detection and function prediction. Bioinformatics (Oxford, England). 2012;28(9):1239-45.
9. Schaefer MH FJ, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. Plos One. 2012;7(2).
10. 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-.
11. 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.
12. Matthews LR. Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". Genome Research. 2001;11(12):2120-6.
13. Nguyen PV, Srihari S, Leong HW. Identifying conserved protein complexes between species by constructing interolog networks. BMC Bioinformatics. 2013;14(Suppl 16):S8-S.
14. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. Science (New York, NY). 2000;287(5450):116-22.
15. Mohieddin Jafari MS, Mehdi Mirzaie, Sayed-Amir Marashi, Mostafa Rezaei-Tavirani. Evolutionarily conserved motifs and modules in mitochondrial protein–protein interaction networks. Mitochondrion. 2013;13:7.