crosstalkr: an R package for the identification of related nodes in biological networks

15 November 2022

Summary

Crosstalkr is designed to facilitate the identification of functional subnetworks in graph-structured data. It is a free, open-source R package designed to allow users to integrate functional analysis using the protein-protein interaction network into existing bioinformatic pipelines. Given a set of user-provided seed proteins, crosstalkr will identify a group of proteins that have a high affinity for the provided seeds. This is accomplished using random walks with restarts, starting at the user-provided seed proteins. Affinity scores from a given random walk with restarts are compared to a bootstrapped null distribution to assess statistical significance. Random walks are implemented using sparse matrix multiplication to facilitate fast execution. The default behavior evaluates the human interactome. However, users can also provide a different graph, allowing for flexible evaluation of graph or network-structured data. Further, users can evaluate more than 1000 non-human protein-protein interaction networks thanks to integration with StringDB.

Statement of Need

In the last few decades, interest in graph-based analysis of biological networks has grown substantially. Protein-protein interaction networks are one of the most common biological networks, and represent the molecular relationships between every known protein and every other known protein.

Researchers have applied graph search and graph clustering algorithms to biological networks in an effort to derive disease-specific subgraphs (Nibbe, Koyutürk, and Chance 2010; Chitra, Park, and Raphael 2022; Pfeifer et al. 2022) or identify potential drug targets (Weaver et al. 2021; Martínez et al. 2015). One of the most well-studied algorithms in this context is the random walk with restarts. Random walks with restarts (RWR) have been used and adapted across disciplines and industries for applications ranging from internet search engines to drug target identification. (Tong, Faloutsos, and Pan 2006; Bianchini, Gori, and Scarselli 2005; Navarro et al. 2017)

There is a growing suite of tools available in R for analyzing graph-structured data (details 2022; Gatto and Christoforou 2014), including a few R packages that implement RWR (Fang and Gough 2014; Valentini 2022). These include the RANKS package, which provides tools for performing many graph-based node scoring algorithms.(Valentini (2022)). These tools require some understanding of graph data structures and ask the user to find, download, and manipulate the relevant biological networks into adjacency matrices or igraph objects. Crosstalkr compromises some of the flexibility of RANKS to provide an optimized, streamlined interface to allow users to integrate interactomic analyses into their workflow. While users can interact directly with crosstalkr to perform RWR on any graph, the package is optimized to facilitate one-line implementation of an algorithm designed to identify functional subgraphs of protein-protein interaction networks (PPI).

Design and Data Sources

compute_crosstalk

The main entrypoint for most users will be the compute_crosstalk function. If users plan to search a supported protein protein interaction network, they are only required to provide a vector of seed proteins. In this situation, compute_crosstalk will:

1. Download the requested PPI (or load it from the provided cache)
2. Process the requested PPI into a sparse adjacency matrix.
3. Perform a random walk with restart using the user provided seeds to generate affinity scores for every protein in the PPI.
4. Perform many random walks with restarts from n random seeds with a matching degree distribution to generate a null distribution of affinity score.
5. Compare the affinity scores to the null distribution to compute an adjusted p-value (using the method specified in p_adjust)
6. Remove proteins with an adjusted p-value < significance_level

Users can make use of caching to store processed PPIs and speed up future analyses substantially. Users can also make use of parallel computing by setting the ncores parameter > 1 A sample workflow demonstrating the ease of use is provided below. Here, we attempt to determine proteins that are functionally related to EGFR, KRAS, PI3K, and STAT3; proteins that are involved in growth signaling in cancer cells.

```
df <- compute_crosstalk(seed_proteins = c("EGFR", "KRAS","STAT3"),
                        cache = "./data", seed_name = "joss_ex", n = 1000,
                        significance_level = 0.99)
df %>%
  select(-c(Z,mean_p, var_p, nobs)) %>%
  slice_max(order_by = affinity_score, n = 5)
```

node	seed	affinity_score	p_value	adj_p_value
STAT3	yes	0.200	0	0
EGFR	yes	0.200	0	0
KRAS	yes	0.200	0	0
C2orf72	no	0.004	0	0
CCDC87	no	0.003	0	0

We also provide a convenience function to quickly plot the returned subgraph (Figure Figure 1). Users can specify `prop_keep` to improve readability by only plotting the top x% of identified proteins, ranked by affinity score.

```
g <- prep_stringdb(cache = "./data")
crosstalkr::plot_ct(df, g=g, prop_keep = 0.4, label_prop = 0.2)
```

Other Features

In pursuit of a one-line interactomic analysis pipeline, we developed several convencience functions that users will likely find useful in other analyses. For example, the human protein-protein interaction network crosstalkr is able to detect and convert between entrez ids, uniprot names, and ensemble ids. Users can make

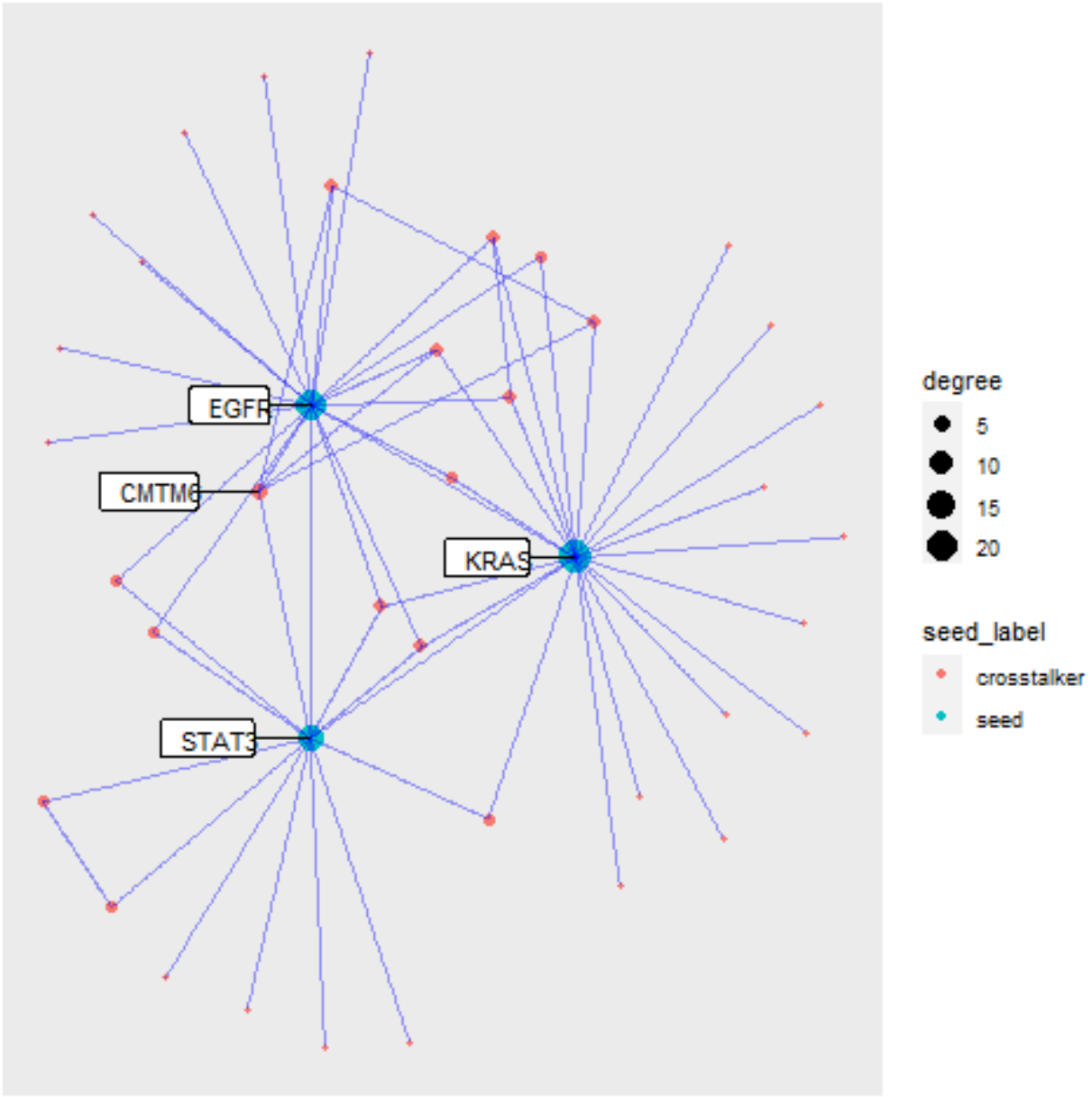

Figure 1: Protein-protein interaction subnetwork for EGFR, KRAS, and STAT3. CMTM6 was identified as a highly connected hub in the computed subnetwork.

use of the as_gene_symbol function to convert any common representation of gene identity into human-readable gene names. In addition, users can make use of single-line functions that download and clean PPIs from either StringDB or Biogrid. Crosstalkr also ships with a highly optimized implementation of the random-walk with restarts algorithm (sparseRWR), which users can apply to any graph-structured data.

Data sources

Users can leverage two high quality PPIs through crosstalkr; StringDB and Biogrid (Oughtred et al. 2021; Szklarczyk et al. 2021). Users can run their analysis using either of these resources individually or they can take the union or intersection of these networks. While Biogrid only supports the human PPI, StringDB provides high quality PPIs for more than 1500 species (Szklarczyk et al. 2021). Crosstalkr provides a user-friendly interface for all of these species.

Acknowledgements

We acknowledge contributions from Mark Chance and Mehmet Koyuturk. We would also like to acknowledge the dependencies that enable crosstalkr (Wickham, François, et al. 2022; details 2022; magrittr) et al. 2022; Hester et al. 2022; Bates et al. 2022; Wickham, Hester, et al. 2022; Wickham, Girlich, and RStudio 2022; Daniel, Ooi, et al. 2022; Daniel, Corporation, et al. 2022; Rainer, Gatto, and Weichenberger 2019) and this paper (Xie [aut et al. 2022; Wickham, Chang, et al. 2022)

Citations

Bates, Douglas, Martin Maechler, Mikael Jagan, Timothy A. Davis (SuiteSparse and 'cs' C. libraries, notably CHOLMOD AMD, collaborators listed in and dir(pattern="^+txt$", full.names=TRUE, et al. 2022. "Matrix: Sparse and Dense Matrix Classes and Methods." https://CRAN.R-project.org/package=Matrix.

Bianchini, Monica, Marco Gori, and Franco Scarselli. 2005. "Inside PageRank." *ACM Transactions on Internet Technology* 5 (1): 92–128. https://doi.org/10.1145/1052934.1052938.

Chitra, Uthsav, Tae Yoon Park, and Benjamin J. Raphael. 2022. "NetMix2: Unifying Network Propagation and Altered Subnetworks." In *Research in Computational Molecular Biology*, 193–208. https://doi.org/10.1007/978-3-031-04749-7_12.

Daniel, Folashade, Microsoft Corporation, Steve Weston, and Dan Tenenbaum. 2022. "doParallel: Foreach Parallel Adaptor for the 'Parallel' Package." https://CRAN.R-project.org/package=doParallel.

Daniel, Folashade, Hong Ooi, Rich Calaway, Microsoft, and Steve Weston. 2022. "Foreach: Provides Foreach Looping Construct." https://CRAN.R-project.org/package=foreach.

details, See AUTHORS file igraph author. 2022. "Igraph: Network Analysis and Visualization." https://CRAN.R-project.org/package=igraph.

Fang, Hai, and Julian Gough. 2014. "The 'Dnet' Approach Promotes Emerging Research on Cancer Patient Survival." *Genome Medicine* 6 (8): 64. https://doi.org/10.1186/s13073-014-0064-8.

Gatto, Laurent, and Andy Christoforou. 2014. "Using R and Bioconductor for Proteomics Data Analysis." *Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 42–51. https://doi.org/10.1016/j.bbapap.2013.04.032.

Hester, Jim, Lionel Henry, Kirill Müller, Kevin Ushey, Hadley Wickham, Winston Chang, Jennifer Bryan, Richard Cotton, and RStudio. 2022. "Withr: Run Code 'With' Temporarily Modified Global State." https://CRAN.R-project.org/package=withr.

magrittr), Stefan Milton Bache (Original author and creator of, Hadley Wickham, Lionel Henry, and RStudio. 2022. "Magrittr: A Forward-Pipe Operator for R." https://CRAN.R-project.org/package=magrittr.

Martínez, Víctor, Carmen Navarro, Carlos Cano, Waldo Fajardo, and Armando Blanco. 2015. "DrugNet: Network-Based Drug–Disease Prioritization by Integrating Heterogeneous Data." *Artificial Intelligence in Medicine* 63 (1): 41–49. https://doi.org/10.1016/j.artmed.2014.11.003.

Navarro, Carmen, Victor Martínez, Armando Blanco, and Carlos Cano. 2017. "ProphTools: General Prioritization Tools for Heterogeneous Biological Networks." *GigaScience* 6 (12): 1–8. https://doi.org/10.1093/gigascience/gix111.

Nibbe, Rod K., Mehmet Koyutürk, and Mark R. Chance. 2010. "An Integrative -Omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer." *PLOS Computational Biology* 6 (1): e1000639. https://doi.org/10.1371/journal.pcbi.1000639.

Oughtred, Rose, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, et al. 2021. "The BioGRID Database: A Comprehensive Biomedical Resource of Curated Protein, Genetic, and Chemical Interactions." *Protein Science : A Publication of the Protein Society* 30 (1): 187–200. https://doi.org/10.1002/pro.3978.

Pfeifer, Bastian, Afan Secic, Anna Saranti, and Andreas Holzinger. 2022. "GNN-SubNet: Disease Subnetwork Detection with Explainable Graph Neural Networks." *bioRxiv*, 2022.01.12.475995. https://doi.org/10.1101/2022.01.12.475995.

Rainer, Johannes, Laurent Gatto, and Christian X Weichenberger. 2019. "Ensembldb: An R Package to Create and Use Ensembl-Based Annotation Resources." *Bioinformatics* 35 (17): 3151–53. https://doi.org/10.1093/bioinformatics/btz031.

Szklarczyk, Damian, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, et al. 2021. "The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets." *Nucleic Acids Research* 49 (D1): D605–12. https://doi.org/10.1093/nar/gkaa1074.

Tong, Hanghang, Christos Faloutsos, and Jia-yu Pan. 2006. "Fast Random Walk with Restart and Its Applications." In *Sixth International Conference on Data Mining (ICDM'06)*, 613–22. https://doi.org/10.1109/ICDM.2006.70.

Valentini, Giorgio. 2022. "RANKS: Ranking of Nodes with Kernelized Score Functions." https://CRAN.R-project.org/package=RANKS.

Weaver, Davis T., Kathleen I. Pishas, Drew Williamson, Jessica Scarborough, Stephen L. Lessnick, Andrew Dhawan, and Jacob G. Scott. 2021. "Network Potential Identifies Therapeutic miRNA Cocktails in Ewing Sarcoma." *PLOS Computational Biology* 17 (10): e1008755. https://doi.org/10.1371/journal.pcbi.1008755.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and RStudio. 2022. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and RStudio. 2022. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Maximilian Girlich, and RStudio. 2022. "Tidyr: Tidy Messy Data." https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, Romain Francois, Jennifer Bryan, Shelby Bearrows, RStudio, https://github com/mandreyel/ (mio library), Jukka Jylänki (grisu3 implementation), and Mikkel Jørgensen (grisu3 implementation). 2022. "Readr: Read Rectangular Text Data." https://CRAN.R-project.org/package=readr.

Xie [aut, Yihui, cre, Abhraneel Sarma, Adam Vogt, Alastair Andrew, Alex Zvoleff, Amar Al-Zubaidi, et al. 2022. "Knitr: A General-Purpose Package for Dynamic Report Generation in R." https://CRAN.R-project.org/package=knitr.