Bioinformatics, 37(14), 2021, 2070–2072
doi: 10.1093/bioinformatics/btaa942
Advance Access Publication Date: 2 November 2020
Applications Note

Systems biology

COSIFER: a Python package for the consensus inference of molecular interaction networks

Matteo Manica [1,2,*,†], Charlotte Bunne[1,3,†], Roland Mathis[1,†], Joris Cadow[1,*,†], Mehmet Eren Ahsen[4], Gustavo A. Stolovitzky[4,5] and María Rodríguez Martínez[1,*]

[1]Cognitive Computing and Industry Solutions, IBM Research Europe, Rüschlikon, ZH 8803, Switzerland, [2]Institute of Molecular Systems Biology, ETH Zürich, Zürich, ZH 8093, Switzerland, [3]Institute for Machine Learning, ETH Zürich, Zürich, ZH 8092, Switzerland, [4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674, USA and [5]Translational Systems Biology and Nanobiotechnology, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

.Received on April 17, 2020; revised on September 25, 2020; editorial decision on October 24, 2020; accepted on October 26, 2020

Abstract

Summary: The advent of high-throughput technologies has provided researchers with measurements of thousands of molecular entities and enable the investigation of the internal regulatory apparatus of the cell. However, network inference from high-throughput data is far from being a solved problem. While a plethora of different inference methods have been proposed, they often lead to non-overlapping predictions, and many of them lack user-friendly implementations to enable their broad utilization. Here, we present *Consensus Interaction Network Inference Service* (COSIFER), a package and a companion web-based platform to infer molecular networks from expression data using state-of-the-art consensus approaches. COSIFER includes a selection of state-of-the-art methodologies for network inference and different consensus strategies to integrate the predictions of individual methods and generate robust networks.

Availability and implementation: COSIFER Python source code is available at https://github.com/PhosphorylatedRabbits/cosifer. The web service is accessible at https://ibm.biz/cosifer-aas.

Contact: tte@zurich.ibm.com or dow@zurich.ibm.com or mrm@zurich.ibm.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of high-throughput technologies has enabled the molecular determination of many cellular components, however, the inference of the gene regulatory networks that govern the internal cellular machinery is still a long-standing challenge for which no single, best-performing method has been proposed. To understand the advantages and limitations of different network inference approaches, the *Dialogue on Reverse Engineering Assessment and Methods* (DREAM) project assessed the performance of over 30 network inference methods (Marbach *et al.*, 2012). One of the main conclusions of the study is that no single method performs optimally across all settings and datasets, but a high-confidence consensus network built by integrating predictions across distinct methods shows robustness across species and datasets and achieves the best overall performance. This finding has been called the *Wisdom of the Crowds* for network inference.

Here, we introduce *Consensus Interaction Network Inference Service* (COSIFER), a Python package with a companion web based platform, that provides a service for inferring gene interaction networks from molecular data. COSIFER integrates 10 different inference methods, chosen due to their performance, complementary theoretical approaches and scientific acceptance. COSIFER (Fig. 1) improves the user-friendly accessibility of existing inference methods and provides three different consensus approaches inspired by the Wisdom of the Crowds (Ahsen *et al.*, 2019; Marbach *et al.*, 2012; Wang *et al.*, 2014).

2 Materials and methods

2.1 Selection of methods

COSIFER currently integrates 10 unsupervised inference methods. Methods have been selected based on their inference performance, as reported by previous studies (Iyer *et al.*, 2017; Marbach *et al.*, 2012; Maetschke *et al.*, 2014), as well as their distinct theoretical approaches, which has been shown to be an important factor for robust performance of consensus networks (Dietterich, 2000). Thus

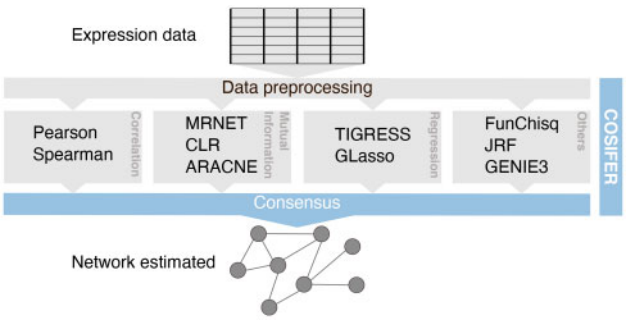

Fig. 1. COSIFER workflow. COSIFER implements 10 different unsupervised network inference methods and three integration strategies to produce a high-confidence consensus network. In addition, COSIFER implements several utilities to preprocess data including: on-the-fly decompression based on file extension, data standardization, mean imputation (imputation of missing values using the mean of the measured data) and the possibility of inferring pathway-specific networks if the user provides a .gmt file. A user can select different preprocessing approaches, inference methods and consensus strategy to process expression data, and COSIFER returns pairwise interaction networks between the measured molecular entities. The resulting consensus network and single-method inferred networks are stored as edge lists in gzipped.csv format. The edge list is composed by triplets, listing the interacting entities and the interaction intensity, which is a real value $\in [0, 1]$ associated with the strength of the predicted interaction. No threshold is applied to the output

COSIFER comprises methods based on correlation, mutual information, regression, tree ensembles and functional χ^2-test based methods. Supplementary Table S1 provides an overview of the methods included in COSIFER. The theoretical approach as well as the default parameters used by every method are discussed in the Supplementary Section S1.

2.2 Integration strategies

As previously discussed (Marbach *et al.*, 2012), assuming that the ranks of an individual edge predicted by different methods are independent, the central limit theorem of probability theory states that the average rank distribution will approach a Gaussian distribution whose variance shrinks with the number of integrated methods. Hence, a consensus approach can approximate the true edge distribution given that a sufficient number of independent methods are integrated.

COSIFER allows for single-method or consensus network inference. In practice, each method outputs a weighted adjacency matrix, i.e. a real-valued matrix, where the elements indicate the strength of an interaction. In the consensus inference mode, the matrices are first scaled between 0 and 1 using min-max scaling to obtain comparable matrices across methods, and then integrated using one of three possible consensus strategies. These strategies are: (i) the Wisdom of the Crowds (WOC) (Marbach *et al.*, 2012), where a consensus network is built by averaging the rank of the interactions predicted by each method and assigning a zero rank to those interactions not predicted by a method; (ii) WOC (hard), a WOC variant that also averages ranks, but only across the methods that have predicted an interaction, i.e. no zeros are added to the average; (iii) similarity Network Fusion (SNF) (Wang *et al.*, 2014), an iterative method based on the use of a sparse kernel approximation of the similarity matrices that represent the interactions and (iv) SUMMA (Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions) (Ahsen *et al.*, 2019), a novel unsupervised ensemble learning algorithm that estimates the AUROC (Area under the Receiver Operating Characteristic) of each individual method and assigns a weight to each method proportional to its estimated performance. In all integration schemes, the reported final score is the edge interaction score in the consensus network. All methods are reported in the Supplementary Table S2.

2.3 COSIFER Python package

COSIFER is available as a Python package, distributed under the MIT license, on GitHub at https://github.com/PhosphorylatedRabbits/cosifer. The Python module implements utilities, functions and classes to handle expression data for network inference and consensus network prediction. Its object-oriented nature permits seamless inclusion of additional methods. During setup, it will install a script (cosifer) to run the full COSIFER pipeline from the command line. The cosifer script accepts as input a character-separated file (e.g. .csv, .tsv, etc.) containing expression data, and creates an output folder where the inference results, single-method and consensus, are stored as an edge list in compressed.csv format [compatible with most popular software tools, e.g. Cytoscape (Shannon *et al.*, 2003), NetworkX (Hagberg *et al.*, 2008)]. A user can: enable/disable data standardization with a flag; specify whether samples are on column or rows; provide a selection of inference methods to run as well as the consensus method of choice; and optionally, provide a .gmt file to perform inference on all the gene sets listed in it. For more details on COSIFER module and the script usage, a user can consult the documentation available in the Supplementary Material S1 or online at this link. In addition, a tutorial describing an application of COSIFER to breast cancer, both in the case of proteomic and transcriptomic data, is available in the Supplementary Material S2 or on GitHub at the following link.

2.4 COSIFER web GUI

COSIFER is available as a login-free service on IBM Cloud at https://ibm.biz/cosifer-aas. COSIFER web application integrates basic functionalities, such as data upload, data processing, inference and choice of consensus method. For performance reasons, only WOC, WOC (hard) and SUMMA are available on the web GUI. The GUI also enables the selection of molecular entities and interactive network visualization using *Bokeh* (Bokeh Development Team, 2018), see Supplementary Figure S1. A detailed description of the application can be found in Supplementary Section S2, together with screenshots of the web interface in Supplementary Figure S2. A full example on RNASeq data from TCGA BRCA is reported in Supplementary Section S2.2.

3 Discussion

Although many methods for network inference have been proposed, no method systematically outperforms all others in terms of robustness and reconstruction accuracy across heterogeneous datasets. We have presented here COSIFER, a Python package with a companion web service GUI, that uses the collective knowledge gathered by a community of complementary inference approaches to generate more robust and stable predictions. COSIFER integrates 10 network inference methods, selected for their reported superior performance as well as their diverse theoretical foundations, and implements 4 different consensus strategies, including the Wisdom of the Crowds (Marbach *et al.*, 2012) (WOC), the WOC hard, Similarity Network Fusion (Wang *et al.*, 2014) (SNF) and the more recently published method SUMMA, which exploits a weighted rank aggregation approach (Ahsen *et al.*, 2019).

Robust network inference is crucial to improve our understanding of healthy and diseased molecular mechanisms. The networks can be independently analysed or used as prior information for downstream analyses either in the case of bulk data (Manica *et al.*, 2019b) or in more modern data types such as single cell data (Skinnider *et al.*, 2019). However, in the absence of known ground truth, the best inference method for each dataset and task cannot be selected with provable guarantees of optimal performance. COSIFER alleviates this challenge by enabling the inference of consensus networks that are guaranteed at least closely matche (and sometimes surpass) the performance of the best method. Furthermore, COSIFER consensus strategies can also be used to combine networks inferred from multi-modal sources, e.g. networks inferred from context-specific scientific publications (Manica *et al.*, 2019a) or databases, resulting in the prediction of higher confidence interactions.

Funding

This project received funding from the European Union's Horizon 2020 Research and Innovation Program [668858 and 826121].

Conflict of Interest: none declared.

References

Ahsen,M.E. *et al.* (2019) Unsupervised evaluation and weighted aggregation of ranked classification predictions. *J. Mach. Learn. Res.*, **20**, 1–40.

Bokeh Development Team. (2018) *Bokeh: Python library for interactive visualization*.

Dietterich,T.G. (2000) Ensemble methods in machine learning. In: *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, vol 1857. Springer, Berlin, Heidelberg. 10.1007/3-540-45014-9_1.

Hagberg,A. *et al.* (2008) Exploring network structure, dynamics, and function using networkx. *Technical report*, Los Alamos National Lab. (LANL), Los Alamos, NM, USA.

Iyer,A.S. *et al.* (2017) Computational methods to dissect gene regulatory networks in cancer. *Curr. Opin. Syst. Biol.*, **2**, 115–122. [doi: 10.1016/j.coisb.2017.04.004].

Maetschke,S.R. *et al.* (2014) Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief. Bioinf.*, **15**, 195–211. [PubMed : 23698722] [PubMed Central: PMC3956069] [doi: 10.1093/bib/bbt034].

Manica,M. *et al.* (2019a) Context-specific interaction networks from vector representation of words. *Nat. Mach. Intell.*, **1**, 181–190.

Manica,M. *et al.* (2019b) PIMKL: pathway-induced multiple kernel learning. *NPJ Syst. Biol. Appl.*, **5**, 1–8.

Marbach,D. *et al.*; The DREAM5 Consortium. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Skinnider,M.A. *et al.* (2019) Evaluating measures of association for single-cell transcriptomics. *Nat. Methods*, **16**, 381–386.

Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.