Bioinformatics, 32(8), 2016, 1259–1261
doi: 10.1093/bioinformatics/btv732
Advance Access Publication Date: 14 December 2015
Applications Note

Systems biology

CytoGEDEVO—global alignment of biological networks with Cytoscape

Maximilian Malek,[1,2,*] Rashid Ibragimov,[1,3] Mario Albrecht[2,4] and Jan Baumbach[5]

[1]Saarland University, 66123 Saarbrücken, Germany, [2]Institute for Knowledge Discovery, Graz University of Technology, 8010 Graz, Austria, [3]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany, [4]BioTechMed-Graz, Graz, Austria and [5]University of Southern Denmark, Odense, Denmark

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on 25 June 2015; revised on 3 December 2015; accepted on 9 December 2015

Abstract

Motivation: In the systems biology era, high-throughput omics technologies have enabled the unraveling of the interplay of some biological entities on a large scale (e.g. genes, proteins, metabolites or RNAs). Huge biological networks have emerged, where nodes correspond to these entities and edges between them model their relations. Protein–protein interaction networks, for instance, show the physical interactions of proteins in an organism. The comparison of such networks promises additional insights into protein and cell function as well as knowledge-transfer across species. Several computational approaches have been developed previously to solve the network alignment (NA) problem, but only a few concentrate on the usability of the implemented tools for the evaluation of protein–protein interactions by the end users (biologists and medical researchers).

Results: We have created CytoGEDEVO, a Cytoscape app for visual and user-assisted NA. It extends the previous GEDEVO methodology for global pairwise NAs with new graphical and functional features. Our main focus was on the usability, even by non-programmers and the interpretability of the NA results with Cytoscape.

Availability and implementation: CytoGEDEVO is publicly available from the Cytoscape app store at http://apps.cytoscape.org/apps/cytogedevo. In addition, we provide stand-alone command line executables, source code, documentation and step-by-step user instructions at http://cytogedevo.compbio.sdu.dk.

Contact: malek@tugraz.at

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

1 Introduction

An increasing amount of data generated by omics technologies has become available in recent years. Similar to the ubiquitous availability of sequence data since the mid-2000s, individual protein–protein interactions (PPIs) have been organized as networks and collected in online resources. Understanding networks is crucial to complement our knowledge about protein function and evolution, which was initially gained through the DNA sequencing projects of the last decades. We believe that understanding differences and commonalities of the topology of networks from different organisms holds great potential to gain deeper insight into how a cell works as a whole (Kuchaiev and Pržulj, 2011). Network alignment (NA) is a method to find a mapping between nodes in two or more networks, according to some quality criterion. This allows transferring knowledge between the aligned nodes. NA is generally categorized into *global* and *local* methods: global NA results in a one-to-one mapping between nodes, and local NA results in either many-to-many or one-to-many mappings, depending on the method. Further distinction is

made between *pairwise* and *multiple* NA, which denotes the number of networks that a method can align at the same time. An exhaustive survey on the field is given in (Faisal *et al.*, 2015).

Most tools are command line utilities with varying degrees of stability (some are prone to crash, even under normal operation) (Clark and Kalita, 2014) and almost none offer a user interface usable for subsequent network analysis. GraphCrunch2 (Kuchaiev *et al.*, 2011) has the dated GRAAL method built-in but lacks visualization. Some web servers exist, which offer NA with limited visualization, such as NatalieQ (El-Kebir *et al.*, 2014) or NetAligner (Pache and Aloy, 2012). GASOLINE (Micale *et al.*, 2014) is a Cytoscape app for multiple local NA with a user interface and visualization capabilities. With the exception of these, none of the existing tools are particularly user-friendly or perform any other function besides aligning a set of networks given as input files and return the alignment result in an output file.

In addition, loading a previous intermediate result or an initially given mapping for further refinement is not often supported by existing software; notable exceptions are MAGNA (Saraph and Milenković, 2014) and PiSwap (Chindelevitch *et al.*, 2013), respectively, which can be used to refine existing alignments. Most tools rely on complex models (i.e. optimization functions) that make it difficult for the user to understand *why* the algorithm aligns certain nodes. This also hinders an intuitive understanding of the output.

Some methods (notably MI-GRAAL and MAGNA) offer optimizing for multiple scores, e.g. scores derived from network topology and from protein sequence alignments, which are often contradictory. Most methods integrate two or more scores using a weighted sum of the form $(\sum \alpha_i \times \text{score}_i)$ with $\sum \alpha_i = 1$. GHOST (Patro and Kingsford, 2012) attempts to solve the disparity by calculating a 'suitable' weight automatically. More recently, this tradeoff has been thoroughly evaluated for OptNetAlign (Clark and Kalita, 2015) and is still an unsolved problem.

Therefore, we have developed CytoGEDEVO as a Cytoscape app, to make the alignment process more transparent, controllable and flexible. Cytoscape is a popular tool for network analysis, visualization and exploration, which can be extended by third-party apps. Our extension is a user-friendly software for pairwise global NA, offering a user front-end that graphically aids with data analysis and follow-up studies. See the Advanced Tutorial (CytoGEDEVO web site) for a case study. A summary of and citations for related NA tools are provided in Supplementary File S3.

2 Methods

2.1 Extended GEDEVO method

For CytoGEDEVO, we have extended the previous implementation of GEDEVO (Ibragimov *et al.*, 2013), which uses an evolutionary algorithm to iteratively optimize a number of random initial alignments. Candidate alignments are modified using mutation and crossover operations to yield better scoring variants after each iteration. The algorithm terminates after convergence is detected. The original GEDEVO is a method for global topological graph alignment that minimizes the so-called graph edit distance (GED), which is the number of inserted, removed or otherwise modified edges. Since the GED is originally defined for undirected networks only, we use an extended GED model that uses per-edge modification penalties to also account for direction changes. If a network is found to be undirected, an optimized and slightly faster GED calculation method is used. A short description of the basic GEDEVO algorithm and a runtime discussion are given in Supplementary File S1.

Our extended approach is not limited to using only the GED as possible optimization criterion. Any node-versus-node distance or similarity measure can be integrated at the same time, for example BLAST bit-scores or *E*-values, scores derived from Gene Ontology (GO) similarities (Ashburner *et al.*, 2000) or graphlet signatures (Przulj *et al.*, 2006). Multiple such functions may be combined. It is also possible to solely use externally supplied data and entirely disable the built-in topological measures, thus bypassing the shortcomings of topological NA strategies.

In addition to a weighted sum to combine scores, we have integrated cascaded scoring, i.e. out of multiple 'stacked' input scores, the first score is taken if it exceeds a user-defined threshold, otherwise the next score is taken and so on. This way, topology can be incorporated as a 'backup' that is only considered if a more relevant score, e.g. the BLAST-based sequence similarity of two nodes, is too low. CytoGEDEVO can use weighted sum and cascaded scoring at the same time. This addresses the tradeoff between topological and biological features—one can typically optimize for either one but not both at the same time without diminishing results in either category.

If graphlets are used as similarity function, they are directly computed on the fly using Orca (Hocevar and Demsar, 2014), which can compute graphlets of up to five nodes very quickly. Additional similarity and distance functions may be integrated as precomputed all-versus-all score files. Most aligners explicitly require BLAST bit scores or *E*-values, with a fixed requirement for these scores to represent either distance or similarity. In CytoGEDEVO, this is handled in a more flexible way, as every input file can be customized such that the algorithm can properly interpret the score type provided.

2.2 Cytoscape integration

By default, the result of an alignment is a single Cytoscape network that shows both input networks side-by-side, with aligned nodes in the same relative positions. Alignments performed with CytoGEDEVO are fully resumable, i.e. an alignment process can be suspended at any time, visualized, corrected, parameters may be optimized and resumed. Intermediate results may also be studied, visualized and stored for further optimization at later time points. Internal scores are imported as Cytoscape data tables, which can be used to apply visual styles to nodes and edges. Parameters are easily configurable using the app and come with explanations and error checking. For datasets where prior knowledge is available, corresponding node pairs can be assigned manually, which is expected to be particularly helpful for incomplete networks or networks from evolutionarily distant organisms. Furthermore, CytoGEDEVO can be used to import/export alignments, such that results obtained by the command-line version or other aligners can be analyzed and refined as well. This is useful if long-running computations are performed remotely and only the analysis is done on a client machine. Aligned network pairs can be visualized using pair layout variants that are derived from existing layouts, including new layouts provided by other apps. CytoGEDEVO can be used to highlight common connected subgraphs and visualize scores by applying various coloring presets to the network. See the Supplementary File S2 and documentation for examples.

3 Discussion and conclusion

CytoGEDEVO is an extension of the GEDEVO software, which provides significant graphical and functional improvements. We expect CytoGEDEVO to make NAs much more user-friendly and

applicable to end users and easier to interpret for human experts. The app features a number of built-in topological and biological alignment optimization criteria and can incorporate any number of externally supplied node similarities or distances. The basic GEDEVO global topological alignment algorithm was evaluated exhaustively in Ibragimov *et al.* (2013). See Supplementary File S1 for a summary. The underlying model is intuitively understandable, and the user can easily incorporate any custom node similarity measure. CytoGEDEVO improves the use of NA as it can be used to visualize the result of any NA tool in Cytoscape.

Acknowledgements

This work was supported by the Cluster of Excellence for Multimodal Computing and Interaction (MMCI) at Saarland University (to J.B. and M.M.) and by BioTechMed-Graz (to M.M. and M.A.).

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Chindelevitch,L. *et al.* (2013) Optimizing a global alignment of protein interaction networks. *Bioinformatics*, **29**, 2765–2773.

Clark,C. and Kalita,J. (2014) A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, **30**, 2351–2359.

Clark,C. and Kalita,J. (2015) A multiobjective memetic algorithm for PPI network alignment. *Bioinformatics*, **30**, 2351–2359.

El-Kebir,M. *et al.* (2014) NatalieQ: a web server for protein-protein interaction network querying. *BMC Syst. Biol.*, **8**, 40.

Faisal,F. *et al.* (2015) The post-genomic era of biological network alignment. *EURASIP J. Bioinform. Syst. Biol.*, **2015**, 3.

Hocevar,T. and Demsar,J. (2014) A combinatorial approach to graphlet counting. *Bioinformatics*, **30**, 559–565.

Ibragimov,R. *et al.* (2013) GEDEVO: an evolutionary graph edit distance algorithm for biological network alignment. In: *German Conference on Bioinformatics 2013, GCB 2013*. September 10–13, 2013. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Göttingen, Germany, pp. 68–79.

Kuchaiev,O. and Pržulj,N. (2011) Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, **27**, 1390–1396.

Kuchaiev,O. *et al.* (2011) GraphCrunch 2: software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, **12**, 24.

Micale,G. *et al.* (2014) GASOLINE: a Greedy And Stochastic algorithm for Optimal Local multiple alignment of Interaction NEtworks. *PLoS One*, **9**, e98750.

Pache,R.A. and Aloy,P. (2012) A novel framework for the comparative analysis of biological networks. *PLoS One*, **7**, e31220.

Patro,R. and Kingsford,C. (2012) Global network alignment using multiscale spectral signatures. *Bioinformatics*, **28**, 3105–3114.

Przulj,N. *et al.* (2006) Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics*, **22**, 974–980.

Saraph,V. and Milenković,T. (2014) MAGNA: maximizing accuracy in global network alignment. *Bioinformatics*, **30**, 2931–2940.