Bioinformatics, 33(7), 2017, 1107–1109
doi: 10.1093/bioinformatics/btw388
Advance Access Publication Date: 5 January 2017
Letter to the Editor

OXFORD

Systems biology

Rebuttal to the Letter to the Editor in response to the paper: proper evaluation of alignment-free network comparison methods

Ömer Nebil Yaveroğlu[1], Noël Malod-Dognin[2], Tijana Milenković[3] and Nataša Pržulj[2,*]

[1]Calit2, University of California, Irvine, CA, USA, [2]Department of Computer Science, University College London, London, UK and [3]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia
Contact: natasa@cs.ucl.ac.uk

Received on February 15, 2016; revised on May 27, 2016; accepted on June 14, 2016

Dear Editor,

We rebut the allegations of Ali *et al.* (2016) that we mis-read their Ali *et al.* (2014) paper and that we carried out its flawed evaluation.

Alignment-free (AF) network comparison is used to quantify the level of similarity (or equivalently, distance) between input networks, irrespective of the node mapping between the networks. The need for improving AF measures arises from the computational intractability of the underlying subgraph isomorphism problem (Cook, 1971) and from the important applications that network comparison measures have in many domains, including computational biology. *Alignment-based* (AB) network comparisons directly account for the node mapping between the networks being compared, which AF measures do not.

An AF measure called NetDis was published in September 2014 by Ali *et al.* (2014). Unfortunately, Ali *et al.* (2014) did not properly evaluate NetDis: despite the focus of their study being the introduction of a new AF network distance measure, NetDis, they only evaluated NetDis against one outdated AB method and not against *any* of the existing AF measures. For example, RGFD (Pržulj *et al.*, 2004) and GDDA (Pržulj, 2007) are AF measures that had been available for 10 and 7 years, respectively, before Ali *et al.* (2014) appeared. Another AF measure is GCD (Yaveroğlu *et al.*, 2014), which was published 4 days after the NetDis paper was submitted, but over 5 months before NetDis was published. Hence, Ali *et al.* (2014) could certainly have compared NetDis to RGFD and GDDA. In addition, Ali *et al.* could also have compared NetDis to GCD in their letter to the Editor that we are rebutting here (Ali *et al.*, 2016). It is unclear why Ali *et al.* (2014) decided not to provide such comparisons. For these reasons, Yaveroğlu *et al.* (2015) conducted an objective and comprehensive evaluation of the existing AF network distance measures, including RGFD, GDDA, GCD and NetDis, amongst others and showed that many of the claims of Ali *et al.* (2014) about NetDis were not supported by experimental evidence.

1 Null models for real-world networks

All of RGFD, GDDA, GCD and NetDis are based on *graphlets*, small induced subgraphs (Pržulj *et al.*, 2004). Once the desired graphlet size is chosen, RGFD, GDDA and GCD have no further parameters. However, NetDis depends on an additional parameter: a null model of the data network. Ali *et al.* (2014) present this as an advantage of their method. It was recognized more than a decade ago that a null model can be used to correct for background noise when counting subgraphs in a network by accounting for the background subgraph counts (Milo *et al.*, 2002). However, shortly after, it was shown that this is a double-edged sword, since the use of an inappropriate null model can lead to incorrect statistical and biological conclusions (Artzy-Randrup *et al.*, 2004). Importantly, this is likely to be the case in practice, since the null model is generally unknown for real-world data. While Ali *et al.* (2014) re-introduced a decade old idea of correcting for background subgraph counts, they failed to discuss what else this correction also involves.

To address this, we evaluated the performance of NetDis under *seven* different null models and observed that its performance strongly depends on the chosen null model (Yaveroğlu *et al.*, 2015). Importantly, we found that even when using the null-model resulting in its best performance, NetDis is inferior to the network distance measures that do not rely on a null model, and to GCD in particular (Yaveroğlu *et al.*, 2015). Note that GCD does not rely on a null model, but it accounts for the background graphlet counts in

the data by relying on the shared graphlet-count variance normalized by their total variance in the data. Similarly, RGFD and GDDA do not rely on a null model, but the distributions of graphlets and of their degrees in a network are normalized by their total numbers in the network. Hence, RGFD, GDDA and GCD also account for background graphlet counts in the data, as NetDis does, without relying on the challenging issue of choosing an appropriate null model, while at the same time they outperform NetDis in the task of AF network comparison (Yaveroğlu *et al.*, 2015).

2 Comparing apples and oranges

We restate our key point (Yaveroğlu *et al.*, 2015): Ali *et al.* (2014) should have evaluated their new AF measure, NetDis, against the existing AF measures, rather than against only one outdated AB method. This is because a new AF measure is usually introduced to outperform the existing AF measures, just as a new AB method is usually introduced to outperform the existing AB methods. Another reason for introducing a new AF measure may be a new idea that allows for filling a gap that the existing AF measures do not properly handle, and perhaps the reliance of NetDis on a null model to correct for background noise could be viewed as such. However, such reliance is not novel and it also introduces serious problems (see Section 1). In addition, it results in a lower accuracy and a higher computational complexity compared to the existing AF measures, as we demonstrated in (Yaveroğlu *et al.*, 2015) and as we further elaborate on in this letter. An additional reason for introducing a new AF measure, even though it might show inferior performance compared to the existing AF measures (as is the case with NetDis), may be its ability to capture novel insights that the existing AF measures cannot capture. However, since neither Ali *et al.* (2014) nor Ali *et al.* (2016) compared NetDis to any AF measure, they could not evaluate whether NetDis has this ability. Therefore, it remains an open question whether NetDis has an advantage over the state-of-the-art in AF network comparison. Importantly, nobody other than Ali *et al.* (2014), who compared NetDis only against one AB method, has ever mixed the two, as AF and AB methods differ substantially in what they are measuring and trying to achieve. Even Ali *et al.* (2014, 2016) do not deny that comparison of AF with AB methods is 'inherently ill-suited,' yet that is the only comparison they provide (Ali *et al.*, 2014).

Ali *et al.* (2016) argue that GCD was not published at the time of submission of their NetDis paper (Ali *et al.*, 2014), so they could not have compared NetDis against it. Yet, it remains unclear why Ali *et al.* (2014) did not compare NetDis against RGFD (Pržulj *et al.*, 2004) and GDDA (Pržulj, 2007), especially since the corresponding code has been publicly available as open source software since 2008 (Kuchaiev *et al.*, 2011; Milenković *et al.*, 2008). Also, why have Ali *et al.* not yet compared NetDis to RGFD, GDDA or GCD in their letter to the Editor (Ali *et al.*, 2016), but chose to speculate about its performance instead? Ali *et al.* (2016) claim that the GCD code (Yaveroğlu *et al.*, 2014) is not publicly available. It is available at http://www0.cs.ucl.ac.uk/staff/natasa/GCD/. In addition, the code for producing key components needed to compute GCD, namely 73-dimensional graphlet degree vectors, has been publicly available in open source software GraphCrunch since 2008 (Kuchaiev *et al.*, 2011; Milenković *et al.*, 2008). Given the graphlet degree vectors that GraphCrunch computes, all that is needed to compute GCD is to simply calculate Spearman's correlation coefficients between the vectors, as detailed in (Yaveroğlu *et al.*, 2014). In addition, the front page of Yaveroğlu *et al.* (2014) specifies that the GCD code, along with all other materials from the GCD paper (Yaveroğlu *et al.*, 2014), are available upon request, which is common practice and in full compliance with the requirements of the journal *Scientific Reports* where GCD was published.

Furthermore, Ali *et al.* (2016) claim that MI-GRAAL was the only available method that was used to produce phylogenetic trees based on subraph counts. However, other AB methods, such as GRAAL (Kuchaiev *et al.*, 2010) and H-GRAAL (Kuchaiev *et al.*, 2010), were also used for this purpose. Furthermore, since a phylogenetic tree is constructed based on the level of similarity between molecular networks of species in question, any AF network distance measure (and not just AB methods) could have been used for that purpose. Also, since Ali *et al.* (2014) decided that MI-GRAAL, an AB method from 2011, could be used to construct phylogenetic trees, then clearly they could have also used any newer AB method, so Yaveroğlu *et al.* (2015) suggested three such methods, GHOST, NETAL and MAGNA (and several newer methods have appeared since). Alas, Ali *et al.* (2016) have again decided to speculatively object to our study, focusing their objections on two of the suggested methods that were not published at the time of NetDis's submission (GCD and MAGNA), instead of conducting a proper evaluation that would have supported or refuted their arguments.

3 Proper AF network comparison

Ali *et al.* (2014) used NetDis to compare protein–protein interaction (PPI) networks of five species (*Helicobacter pylori*, *Escherichia coli*, *Drosophila melanogaster*, *Homo sapiens sapiens* and *Saccharomyces cerevisiae*) and then reconstructed the phylogenetic tree of these species based on the resulting NetDis distances. We argued in (Yaveroğlu *et al.*, 2015) that the application of NetDis to phylogeny reconstruction, as designed and carried out by Ali *et al.* (2014), is scientifically inaccurate. We stated that for the following reasons. (i) Currently available PPI network data are incomplete, with many labs throughout the world continuously contributing additional PPI data, so these datasets grow and change very quickly; that makes null model-based AF comparisons extremely biased due to quickly changing null-model of the data. (ii) The same phylogenetic tree cannot be obtained by NetDis when it uses PPI networks of the above species that come from different databases. (iii) Different input parameters for NetDis (i.e. different null models and graphlet sizes) result in different phylogenetic trees for the same input data, so the reconstructed phylogenetic tree reported by Ali *et al.* (2014) is a cherry-picked case out of many possible outcomes. (iv) Phylogenetic trees similar to the one reported by Ali *et al.* (2014) can be partially produced by using trivial network properties as network distances, such as network density. (v) Relying on only five networks to reconstruct phylogeny might not give enough statistical power to properly evaluate significance of the resulting tree. For experimental evidence that supports all five of the above points, see (Yaveroğlu *et al.*, 2015) and its Supplementary Materials. Here, we discuss in more detail the first two points, as these are relevant for rebutting the claims of Ali *et al.* (2016).

Namely, an appropriate null model should fit well the given real-world data. If the data are incomplete and evolve quickly, as is the case with the current PPI data, then the null model should be revised in the light of new, changed data. For this reason, regarding the application of NetDis to phylogenetic tree reconstruction (Ali *et al.*, 2014), we argued that a null model-based approach, such as NetDis, should not be used to reconstruct phylogeny from quickly evolving PPI network data (see Section 3.5 and Supplementary Section 3 of Yaveroğlu *et al.* (2015)). Namely, by using the newest PPI data at

the time of our study, we observed that NetDis could not reconstruct correctly the phylogenetic trees, as claimed by Ali *et al.* (2014) who used older and thus obsolete PPI data. In their Letter to the Editor, Ali *et al.* (2016) state that this observation of ours was flawed because we did not use the same, obsolete PPI data that they had used (which we actually *did* consider, in addition to the newest data, as discussed in Supplementary Section 3 of (Yaveroğlu *et al.*, 2015)). By stating this, Ali *et al.* (2016) admit that the claimed ability of NetDis to correctly reconstruct phylogenetic trees is *not* a generic property of NetDis, but is dataset-dependent. This, in turn, invalidates any general conclusions about NetDis's ability to reconstruct phylogeny that is claimed by Ali *et al.* (2014). We conclude the phylogeny reconstruction discussion by noting that it had been argued in the literature well prior to the NetDis study (e.g. in the GRAAL paper (Kuchaiev *et al.*, 2010)), that PPI networks are an inappropriate choice for reconstructing a phylogenetic tree for as distant species as those analyzed by Ali *et al.* (2014), which is why unlike Ali *et al.* (2014) who used PPI data, Kuchaiev *et al.* (2010) instead used metabolic networks.

We are confused by Ali *et al.*'s (2016) comment regarding the graphlet size choice: we never claimed that using different graphlet sizes would lead to the same results, as Ali *et al.* (2016) have stated. Actually, it is the opposite: since using larger graphlets sometimes helps and sometimes does not (Ali *et al.*, 2014; Yaveroğlu *et al.*, 2014), we varied graphlet sizes within both NetDis and GCD to give each method the best case advantage (Yaveroğlu *et al.*, 2015).

Then, Ali *et al.* (2016) question our evaluation framework, but they do so with flawed arguments. A good network distance measure should yield smaller distances between similar networks (e.g. those from the same random network model) than between dissimilar ones (e.g. those from different random network models). And this is exactly what our evaluation framework measures by relying on ROC and precision-recall (PR) curve analyses, *which are standard and widely adopted ways of doing this*. The suggestion of Ali *et al.* (2014, 2016) to instead use Rand Index is flawed. Namely, Rand Index measures the agreements between two different clusterings of networks: (1) the given gold standard clusters and (2) clusters of the networks constructed based on their pairwise distances. However, a key question here is how to obtain the distance-based clusters in point 2 above? This requires choosing an appropriate clustering method (out of a multitude of available clustering methods) and its typically many parameters, adding an unnecessary and complex step on top of our straight-forward evaluation framework; importantly, this additional step could substantially affect the results. Note that our ROC-based evaluation directly uses the pairwise distances, without requiring any clustering method (i.e. it does not require point 2 of Rand Index described above); it does rely on the given gold standard clusters (as does Rand Index, point 1 described above). Furthermore, Rand Index and our ROC analysis rely on the same principle: both classify pairs of networks as true-positives (tp), true-negatives (tn), false-positives (fp) and false-negatives (fn), with respect to belonging to a cluster from the gold standard. The only difference is that in our ROC analysis, determining a tp, tn, fp and fn is based on the networks having or not their pairwise distance smaller than a threshold (and we consider distances between all pairs of networks as thresholds, without any sampling), while in Rand Index it is based on the networks being or not in the same cluster that is obtained from the chosen distance-based clustering method (step 2 in Rand Index described above). Once it chooses a clustering method to make clusters, the formula for computing Rand Index is the same as for Accuracy in ROC analysis, both being $(tp + tn)/(tp + tn + fp + fn)$. It would have been interesting if in their letter to the Editor, Ali *et al.* (2016) actually evaluated their Rand Index-based evaluation framework against ours, since they would likely produce identical rankings of network distances, in which case their whole argument would have been moot.

Finally, Ali *et al.* (2016) suggest that we mis-computed areas under the ROC and PR curves, or that our computations may not be accurate enough for comparing the performances of distance measures. In our computations (Yaveroğlu *et al.*, 2015), we used all values of thresholds that arise from the data (we did not use any sampling). Given the large numbers of these values, the differences between lower- and upper-bounds on the approximations of the areas under the curves (that are necessary since we are dealing with large but discrete numbers of points) are orders of magnitude smaller than the observed differences between the areas under the curves from different distance measures, so our comparisons are robust in this respect.

4 Conclusion

We made an important step towards a proper evaluation of the current AF network comparison methods (Yaveroğlu *et al.*, 2015). Since the problem of network comparison is computationally hard, meaning that all existing methods are heuristic, the network comparison research will continue to evolve. The same holds for the biological network data, which will continue to grow in size and complexity, so the methods for their analyses will keep needing to be improved. Hence, when a new method is proposed, it needs to be compared against the latest and appropriate methods, and tested on the most recent data, which Ali *et al.* (2014) failed to do when they introduced NetDis.

Conflict of Interest: none declared.

References

Ali,W. *et al.* (2014) Alignment-free protein interaction network comparison. *Bioinformatics*, **30**, i430–i437.

Ali,W. *et al.* (2016) Letter to the editor in response to the paper: Proper evaluation of alignment-free network comparison methods. *Bioinformatics*.

Artzy-Randrup,Y. *et al.* (2004) Comment on "network motifs: simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science*, **305**, 1107–1107.

Cook,S.A. (1971). The complexity of theorem-proving procedures. In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing, STOC '71*. ACM, New York, NY, USA, pp. 151–158.

Kuchaiev,O. *et al.* (2010) Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interface*, **7**, 1341–1354.

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

Milenković,T. *et al.* (2008) GraphCrunch: a tool for large network analyses. *BMC Bioinformatics*, **9**.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Pržulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.

Pržulj,N. *et al.* (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

Yaveroğlu,ÖN. *et al.* (2014) Revealing the hidden language of complex networks. *Sci. Rep.*, **4**.

Yaveroğlu,ÖN. *et al.* (2015) Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, **31**, 2697–2704.