Fair Evaluation of Global Network Aligners

Joseph Crawford*, Yihan Sun† and Tijana Milenkovíc*  
*Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications, and ECK Institute for Global Health, University of Notre Dame  
†Department of Computer Science and Technology, Tsinghua University and Computer Science Department, Carnegie Mellon University  
Email: jcrawfo7@nd.edu, yihans@cs.cmu.edu, tmilenko@nd.edu

Abstract—Analogous to genomic sequence alignment, biological network alignment identifies conserved regions between networks of different species. Then, function can be transferred from well- to poorly-annotated species between aligned network regions. Network alignment encompasses two algorithmic components: node cost function (NCF), which measures similarities between nodes in different networks, and alignment strategy (AS), which uses these similarities to rapidly identify high-scoring alignments. Different methods use both different NCFs and different ASs. Thus, it is unclear whether the superiority of a method comes from its NCF, its AS, or both. We already showed on state-of-the-art methods, MI-GRAAL and IsoRankN, that combining NCF of one method and AS of another method can lead to a new superior method. Here, we evaluate MI-GRAAL against a newer approach, GHOST, by mixing-and-matching the methods’ NCFs and ASs to potentially further improve alignment quality. While doing so, we approach several important questions that have not been asked systematically thus far. First, we ask how much of the node similarity information within NCF should come from protein sequence data compared to network topology data. Existing methods determine this parameter more-less arbitrarily, which could significantly affect the resulting alignment(s). Second, when topological information is used in NCF, we ask how large the size of the neighborhoods of the compared nodes should be. Existing methods assume that the larger the neighborhood size, the better. We find that MI-GRAAL’s NCF is superior to GHOST’s NCF, while the performance of the methods’ ASs is data-dependent. Thus, for data on which GHOST’s AS is superior to MI-GRAAL’s AS, the combination of MI-GRAAL’s NCF and GHOST’s AS represents a new superior method. Also, we find that which amount of sequence information is used within NCF does not affect alignment quality, while the inclusion of topological information is crucial for producing good alignments. Finally, we find that larger neighborhood sizes are preferred, but often, it is the second largest size that is superior. Using this size instead of the largest one would decrease computational complexity.

Taken together, our results lead to several general recommendations for a fair evaluation of network alignment methods.

Index Terms—Across-species protein function prediction, Network alignment; Network similarity; Protein-protein interaction networks;

I. Big Picture

Biological network alignment identifies topologically and functionally conserved regions between networks of different species [1], [2]. Then, biological function can be transferred from well- to poorly-annotated species between aligned network regions [3]. Network alignment has two algorithmic components: node cost function (NCF) and alignment strategy (AS). NCF captures pairwise similarities between nodes of different networks. AS uses these similarities to find high-scoring alignments (i.e., node mappings). Since different existing methods use both different NCFs and ASs, it is not clear whether the superiority of a method comes from its NSF, AS, or both [3]. Thus, here we fairly evaluate MI-GRAAL [1] and GHOST [2] state-of-the-art approaches by mixing and matching their NSFs and ASs (Q1). While doing so, we approach additional research questions that have not been asked in the context of network alignment thus far: 1) how much of the NCF information should come from protein sequence data compared to network topology data (Q2), and 2) how much of a nodes network neighborhood should be considered within NCF (Q3). For more details on this study, please refer to our article in Algorithms for Molecular Biology [4].

Aligners resulting from combining existing NCFs and ASs, and their parameters

| Table 1 | The three aligners considered in this study. The first letter in the aligner represents NCF of the aligner, while the second letter represents AS of the aligner. |
| Aligner | Node Cost Function | Alignment Strategy |
|---------|-------------------|-------------------|
| M-M | MI-GRAAL | MI-GRAAL |
| G-M | GHOST | MI-GRAAL |
| G-G | GHOST | GHOST |

Varying the amount of sequence versus topological information within NCF. For each aligner, we generate NCFs with varying amounts of sequence and topology information, as $\alpha T + (1 - \alpha)S$, where $T$ represents topological similarity score (e.g. node-GDV-similarity) and $S$ represents sequence similarity score. We vary $\alpha$ from 0 to 1 in increments of 0.1.
Varying the size of network neighborhood within NCF. For each aligner (and for each value of \( \alpha \)), we consider four different neighborhood sizes, as described in Table II.

| Size | Graphlet size (used by MI-GRAAL's NCF) | \( k \)-hop neighborhood (used by GHOST's NCF) |
|------|----------------------------------------|------------------------------------------------|
| T1   | 2-node graphlets                       | 1-hop neighborhood                              |
| T2   | 2-3-node graphlets                     | 2-hop neighborhood                              |
| T3   | 2-4-node graphlets                     | 3-hop neighborhood                              |
| T4   | 2-5-node graphlets                     | 4-hop neighborhood                              |

Network alignment quality measures. To measure the quality of alignment, we use the five following alignment quality measures: node correctness (NC), edge correctness (EC), induced conserved structure (ICS), symmetric substructure score (S^3), the size of the largest common connected component, GO correctness, and EXP GO correctness. NC is a measure of how close the alignment is to the true node mapping. EC, ICS, and S^3 measure the amount of conserved edges in the alignment. And lastly, GO/EXP GO correctness measures how many of the aligned nodes share at least one GO/EXP GO term.

II. RESULTS AND DISCUSSION

A. Q1: What is the best NCF and the best AS?

Alignment quality of the three aligners (over all values of \( \alpha \) and all neighborhood sizes) with respect to seven alignment quality measures, for) synthetic yeast-yeast 20% (Figure 1(a)) and real H-W (Figure 1(b)) alignment. M-M is comparable or superior to G-M, indicating that MI-GRAAL's NCF is comparable or superior to GHOST's NCF. In Figure 1(a), G-G outperforms G-M, indicating that GHOST's AS is superior to MI-GRAALs AS, but in Figure 1(b), the two ASs are comparable. Hence, MI-GRAALs NCF is always superior to GHOSTs, while AS superiority is data-dependent. Hence, we propose the combination of MI-GRAALs NCF and GHOSTs AS as a new superior method for certain data.
B. Q2: The amount of sequence versus topological information within NCF?

Figure 2 shows the ranking of the 11 values of (from 0 to 1 in increments in 0.1) over all alignments of real networks (across all aligners, all neighborhood sizes, and all topological alignment quality measures). At least some amount of topological information should be included within NCF, as this results in topologically (and biologically) good alignments, while sequence information alone should not be used.

C. Q3: The size of nodes’ neighborhoods within NCF?

The ranking of the four neighborhood sizes, with respect to all topological alignment quality measures for synthetic (Figure 3(a)) and real network (Figure 3(b)) sets. The larger (but not necessarily the largest) the neighborhood size is used within NCF, the higher the alignment quality. In some cases, using T3 will not only increase alignment quality but also reduce the computational complexity compared to using T4.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation CCF-1319469 and EAGER CCF-1243295 grants.

REFERENCES

[1] O. Kuchaiev and N. Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[2] R. Patro and C. Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.

[3] F. Faisal, H. Zhao, and T. Milenkovic. Global network alignment in the context of aging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(1):40–52, 2014.

[4] J. Crawford, Y. Sun, and T. Milenković. Fair evaluation of global network aligners. *Algorithms for Molecular Biology*, 10(19), 2014.