Evaluation of data integrative approaches for network inference

Khalique Newaz and Tijana Milenković*

Department of Computer Science and Engineering, University of Notre Dame
Interdisciplinary Center for Network Science and Applications; ECK Institute for Global Health
*Corresponding Author (E-mail: tmilenko@nd.edu)

Inference of condition-specific subgraphs: shift from static to dynamic setting

Identification of perturbed molecular pathways under some biological condition is important for understanding cellular behavior under that condition. Typically, the knowledge about perturbed pathways is obtained by studying transcriptomics (gene expression) data to identify differentially expressed genes or to construct a gene co-expression network (by functionally linking genes whose expressions significantly "correlate" across different time points). However, these studies neglect physical interactions among the dysregulated genes (i.e., their protein products). And it is the proteins that carry out cellular function by interacting with each other. Hence, studies of the protein-protein interaction (PPI) network are promising. However, the current PPI data span multiple biological conditions. Clearly, using the whole interactome without considering other condition-specific biological data fails to capture any condition-specific knowledge. As a result, recent studies have integrated transcriptomics data with PPI network data by mapping the activity of dysregulated genes (as captured by the gene expression data) to their corresponding proteins in the PPI network, in order to assign activity weights to nodes (genes/proteins) and edges (PPIs) in the network. Then, these studies consider only the most active network parts as condition-specific dysregulated pathways. We study four such data-integrative network inference methods, i.e., RWR [1], ORIENT [2], HotNet2 [3], and NetWalk [4], and, we evaluate their performance on how correctly they can prioritize activity of nodes and edges.

Importantly, all of these approaches deal with the inference of static dysregulated pathways (or subgraphs). However, cellular functioning is dynamic. So, inferring evolving subgraph underlying a dynamic biological process is of importance. A straightforward way to infer a dynamic PPI subgraph is to map time-stamped "active" genes/proteins on the static PPI network while keeping all the edges (i.e., induced-subgraph) among them [5]. Since, not every edge is "active" in a given biological condition, prioritization of edges is important. To address this and also because studying aging can help us understand many aging-related diseases, we use NetWalk and HotNet2 methods (RWR and ORIENT do not prioritize edges) to extract weighted, dynamic age-specific PPI subgraphs. Then, we use these dynamic age-specific PPI subgraphs to predict aging-related genes.

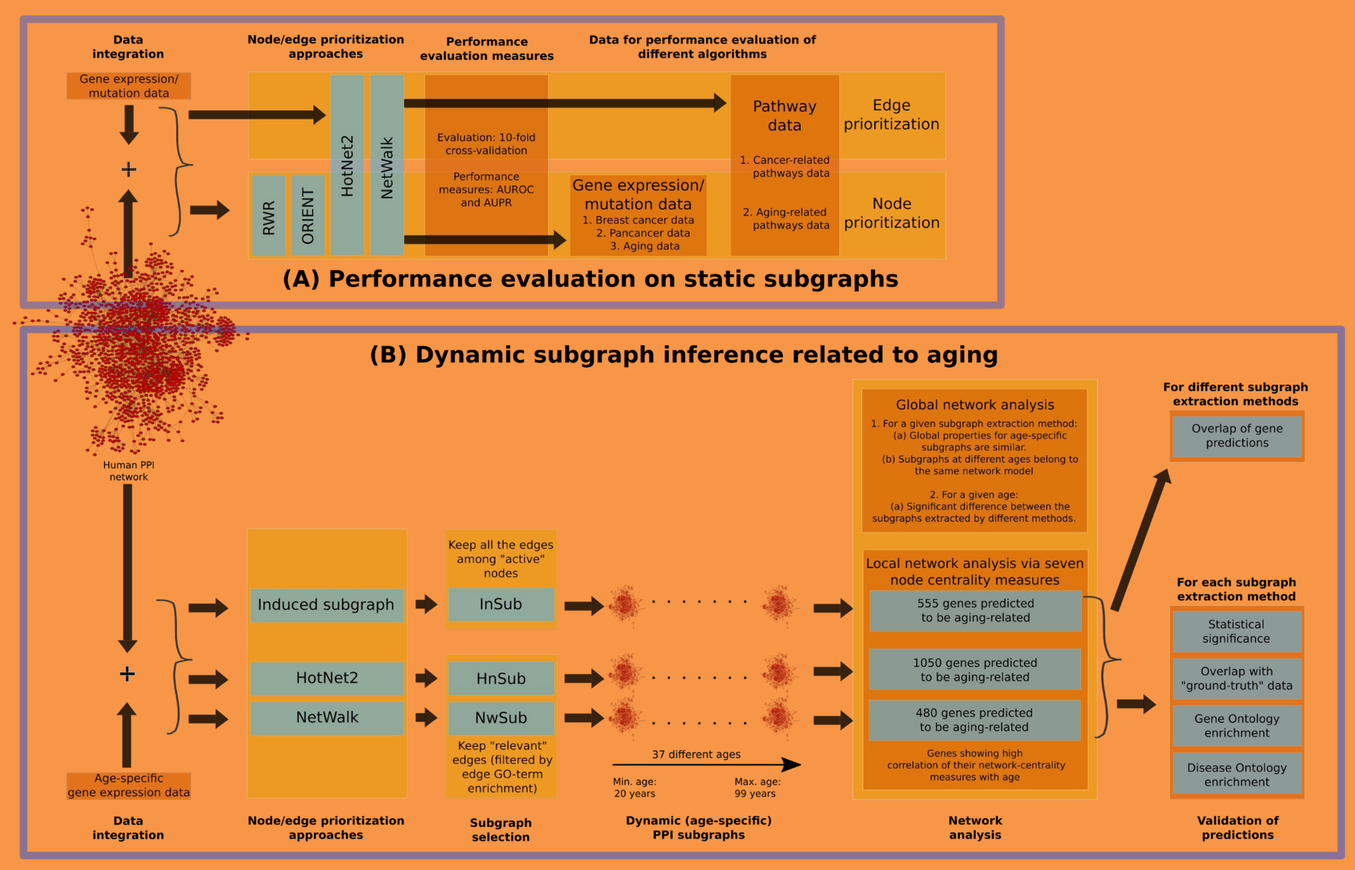


Figure 1: Summary of this study. (A) We evaluate the node/edge prioritization performance of four algorithms, i.e., RWR, ORIENT, HotNet2, and NetWalk, on static subgraphs. (B) We integrate static PPI network with age-specific gene-expression data and obtain age-specific PPI subgraphs using three algorithms, i.e., induced subgraph method, HotNet2, and NetWalk. For each algorithm, we analyze the global and local network properties with age, predict aging-related genes, and validate the predictions.

Results

The node/edge prioritization performance of the algorithms is dependent upon the ground-truth data that we use for the evaluation, and also on the data used by the algorithms to prioritize nodes/edges in the network. In general, the use of condition-specific biological data improves the "quality" of the inferred subgraph.

What is the effect of using condition-specific biological data on the "quality" of the inferred subgraph?

- Since a good measure to quantify the "quality" of an inferred subgraph is currently not available, we use the node and the edge prioritization performance as a proxy to measure the "quality" of the inferred subgraph.
- Figure 2 (A) and 2 (D) shows better performance for the cases where we use actual values for the mutated/expressed genes (in case of NetWalk and HotNet2) in comparison to the cases in which we treat all mutated/expressed genes equally (RWR and ORIENT).
- Use of condition-specific data gives a chance to prioritize edges in the PPI network. HotNet2 performs better than NetWalk when we aim to identify edges among the mutated nodes in the PPI network (figure 4 (A))

Which algorithm works best for which kind of data and for what kind of goals?

- Since NetWalk and HotNet2 is biased towards condition-specific data, they should perform better when the goal is to extract a condition-specific subgraph which primarily consists of "active" nodes/genes/proteins (figure 2(A) and 2(D).)
- Since ORIENT is biased towards the neighborhood of the "active" nodes/genes, it performs better when the goal is to identify genes which are topologically (in terms of PPI network) close to the "active" set of genes (figure 3 (A).)

Does the "prediction-quality" of aging-related genes increase when we use subgraph-inference methods instead of just using the induced subgraphs?

- Statistically, NetWalk* produces better results (figure 6) than other methods.

FUNDING: NSF CAREER CCF-1452795

Node inference performance

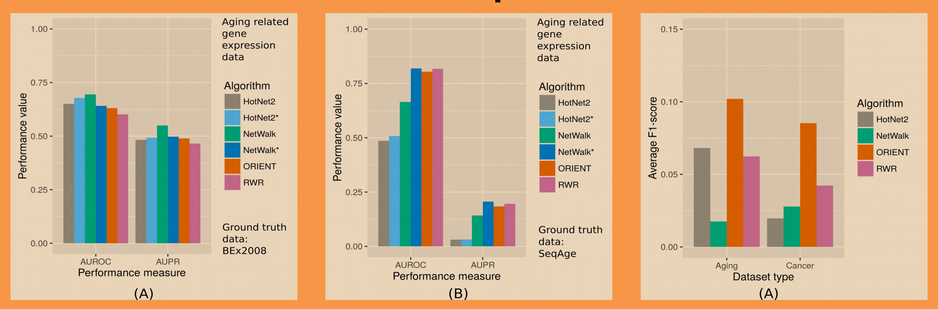


Figure 2: Node performance on real-world (gene expression/mutation) data

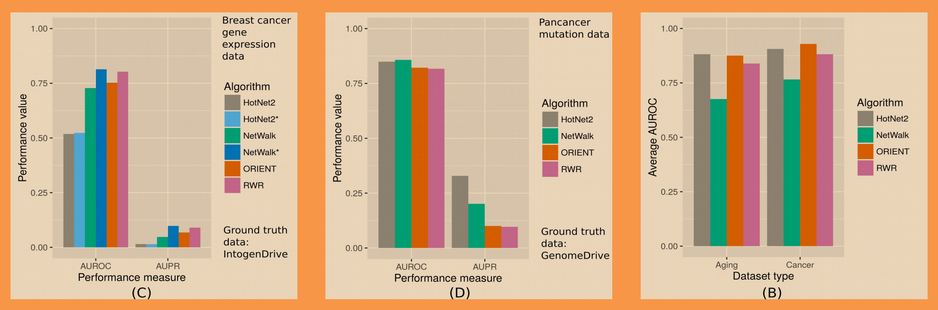


Figure 3: Node performance on biological pathways data

Edge inference performance

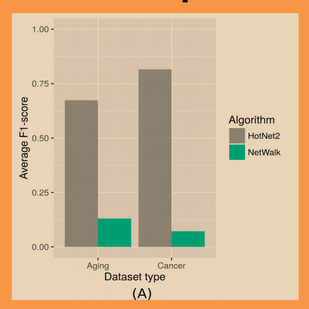
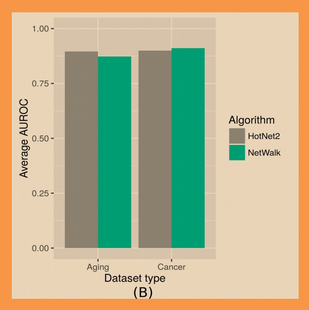


Figure 4: Edge performance on biological pathways data



Aging-related gene predictions

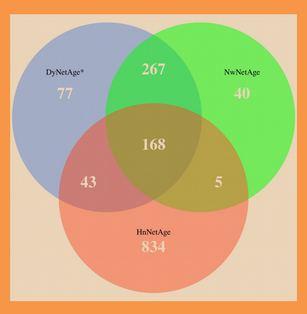


Figure 5: Overlap of predicted genes by different methods

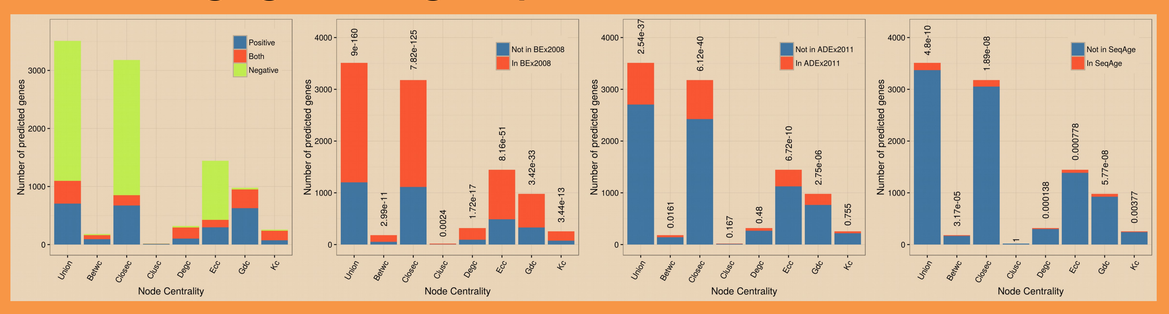


Figure 6: Aging-related gene predictions when we use NetWalk* for subgraph inference.

References

[1] S. Köhler et. al. (2008), American Journal of Human Genetics, 82(4), 949-958, doi: 10.1016/j.ajhg.2008.02.013.
[2] Duc-Hau Le et. al. (2013), Computational biology and chemistry, 44, 1-8, doi: 10.1016/j.compbiolchem.2013.01.001.
[3] Mark D. M. Leiserson et. al. (2015), Nature Genetics, 47,106-114, doi:10.1038/ng.3168.
[4] Kakajan Komurov et. al. (2010), PloS Computational biology, 6(8): e1000889, doi: 10.1371/journal.pcbi.1000889.
[5] Fazle E. Faisal and Tijana Milenković (2014), Bioinformatics, 30 (12), 1721-1729, doi: 10.1093/bioinformatics/btu089.