Topology-based Hypothesis Generation on Causal Biological Networks using igraph

Robert Ness\textsuperscript{1}, Halima Bensmail\textsuperscript{3}, Olga Vitek\textsuperscript{1,2}

1. Department of Statistics, Purdue University, West Lafayette, Indiana, USA
2. Department of Computer Science, Purdue University, West Lafayette, Indiana, USA
3. Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar

\textsuperscript{*}Contact author: nessr@purdue.edu

\textbf{Keywords:} interaction networks, causal graph, proteomics, igraph

A causal biological network as a type of molecular interaction network constructed from biomedical literature. The nodes are molecular events, such as a change in abundance of a protein. Directed edges between two events imply a causal relationship between them. The edges have a directional attribute in terms of increase or decrease. For example, catalytic activity of enzyme A increases the abundance of protein B. Each edge is annotated with a reference to a publication from the biomedical literature that shows experimental validation for the edge. A key difference from similar network models such as Bayesian networks is that the directional edge attribute does not have a numerical magnitude (eg. a probability). This is because the edges each come from different experimental contexts, so it is difficult to establish a basis for numerical comparison.

Causal networks are used to generate and prioritize mechanistic hypotheses for what is observed in experimental data. Specifically, we look at genomics and proteomics data with treatments and a control. Statistics for each feature (eg. gene probe measurement or protein peptide spectra) in the data are mapped to nodes in the network. Typically, the statistic is a 1 if the features measurements changed significantly across treatments, and 0 otherwise. Other nodes in the network are evaluated as mechanistic hypotheses for what was observed in the data, based on the directed paths in the network. High ranked hypotheses can then be validated experimentally. This approach is particularly useful in drug discovery and drug repositioning, for example, because of the potential to identify molecular mechanisms susceptible to drug intervention.

The question we ask is what is the best path-based scheme for ranking nodes in the network. Common practice is to use the shortest-paths algorithm. A node is ranked highly if there is a large number of shortest paths from itself to the nodes with a 1. The problem with this approach is that fails to incorporate broader network topology, in terms of the total number of all directed paths between the source node and the target nodes, rather than just the number of shortest paths. We test an alternative approach that uses a Markov random walk-based algorithm similar to Pagerank. This approach counts the total number of paths between a source and target node using a weighting function that assigns higher weights to more direct paths. We compare the performance of this algorithm to shortest-paths and demonstrate its effectiveness for hypothesis generation in systems biology experiments.

We conduct our analysis using the R package \texttt{igraph}. \texttt{igraph} is an open source software package for creating and manipulating undirected and directed graphs with implementation as a library in R, as well as other programming languages. Though causal biological networks can be small (100s of nodes), we are only interested in situations when they are large (10,000+ nodes), such when knowledge in a knowledgebase is being used to analyze data. The size of the network significantly impacts the speed of most network analysis algorithms. \texttt{igraph} contains several functions implemented in C that allow for fast analysis of large networks, including shortest-paths and other network topology algorithms.

\textbf{Funding:} This work was supported by the Qatar Computational Research Institute, Qatar Foundation, Doha, Qatar