Detection of optimized subnetwork based on node scores and edge scores in metabolic network

Yubing Yao, Denise Scholtens,Michael Nodzenski, Raji Balasubramanian

University of Massachusetts, Amherst

July 29, 2016

1. Goal
2. Background and Data
3. Our proposed method
4. Simulation Results
5. Summary and Future work
6. Reference

Goal

To develop an efficient algorithm to detect optimized subnetwork based on node scores and edge scores in metabolic network.

Goal

To develop an efficient algorithm to detect optimized subnetwork based on node scores and edge scores in metabolic network.

1. **Data type**:Case control studies with metabolic data
2. Optimized subnetwork is to maximize the node scores and edge scores.

Goal

To develop an efficient algorithm to detect optimized subnetwork based on node scores and edge scores in metabolic network.

1. **Data type**:Case control studies with metabolic data
2. Optimized subnetwork is to maximize the node scores and edge scores.
3. Node score is proportional to the strength of its association with an outcome of interest.

Goal

To develop an efficient algorithm to detect optimized subnetwork based on node scores and edge scores in metabolic network.

1. **Data type**:Case control studies with metabolic data
2. Optimized subnetwork is to maximize the node scores and edge scores.
3. Node score is proportional to the strength of its association with an outcome of interest.
4. Edge score reflects the statistical strength of data driven connections (correlations) between metabolite pairs.

The data feature of metabolic network

1. Metabolic networks, in which nodes represent distinct metabolites and undirected edges between nodes are weighted by the magnitude of the correlation between the pair of nodes (metabolites).

The data feature of metabolic network

1. Metabolic networks, in which nodes represent distinct metabolites and undirected edges between nodes are weighted by the magnitude of the correlation between the pair of nodes (metabolites).
2. Metabolic networks different from genomic or proteomic networks in which edges denote pathway relationships or physical interactions (e.g. PPI).

The data feature of metabolic network

1. Metabolic networks, in which nodes represent distinct metabolites and undirected edges between nodes are weighted by the magnitude of the correlation between the pair of nodes (metabolites).
2. Metabolic networks different from genomic or proteomic networks in which edges denote pathway relationships or physical interactions (e.g. PPI).
3. The only information we have about the association among the nodes in the metabolic network is the correlation between the nodes(metabolites).

Methods for identification of functional modules in biological network

1. Various robust and/or efficient algorithms have been proposed to identify the functional modules in protein-protein interaction networks or integrated network incorporating gene expression profiles, transcriptome, and proteome data.

Methods for identification of functional modules in biological network

1. Various robust and/or efficient algorithms have been proposed to identify the functional modules in protein-protein interaction networks or integrated network incorporating gene expression profiles, transcriptome, and proteome data.
2. Of particular interest to our research are algorithms to identify optimal connected subnetworks based on maximizing a scoring function from nodes or edges of a network (Ideker et al. 2002, Guo et al. 2007).

Methods for identification of functional modules in biological network

1. Various robust and/or efficient algorithms have been proposed to identify the functional modules in protein-protein interaction networks or integrated network incorporating gene expression profiles, transcriptome, and proteome data.
2. Of particular interest to our research are algorithms to identify optimal connected subnetworks based on maximizing a scoring function from nodes or edges of a network (Ideker et al. 2002, Guo et al. 2007).
3. Dittrich et al. (2008) proposed the exact solution to find the maximally node scoring subnetwork through interger-linear programming.

Derivation of node scores in the metabolic network

1. Assume that the case control data with n subjects with p metabolites measured on each subject. p p-values-p_j from some statistical test on p metabolites.

Derivation of node scores in the metabolic network

1. Assume that the case control data with n subjects with p metabolites measured on each subject. p p-values-p_j from some statistical test on p metabolites.
2. $p_j \sim \lambda B(a, 1) + (1 - \lambda)B(1, 1), 0 < a < 1, 0 \leq \lambda \leq 1, j = 1, \ldots, p$

Derivation of node scores in the metabolic network

1. Assume that the case control data with n subjects with p metabolites measured on each subject. p p-values-p_j from some statistical test on p metabolites.
2. $p_j \sim \lambda B(a,1) + (1-\lambda)B(1,1), 0 < a < 1, 0 \leq \lambda \leq 1, j = 1, \ldots, p$
3. An adjusted log likelihood ratio node score with a threshold P-value-$\tau_1(FDR_1)$:

$$\begin{aligned} S^{FDR_1}(x) &= \log\left(\frac{ax^{a-1}}{a\tau_1^{a-1}}\right) \\ &= (a-1)(\log(x) - \log(\tau_1(FDR_1))) \end{aligned} \tag{1}$$

Derivation of edge scores in the metabolic network-1

1. Permutation test on $H_0 : \rho_{ij} = 0, H_1 : \rho_{ij} \neq 0, i \neq j = 1, 2, \ldots, p$

Derivation of edge scores in the metabolic network-1

1. Permutation test on $H_0 : \rho_{ij} = 0, H_1 : \rho_{ij} \neq 0, i \neq j = 1, 2, \ldots, p$
2. The exact Monte Carlo p-value under null distribution of the permutation test (Gordon and Phipson 2010):

$$P(|\rho_{ij}^{perm}| \geq |\rho_{ij}^{obs}|) = \frac{b+1}{m+1} \tag{2}$$

where m is the number of permutation samples and b is the number of times out of m that $|\rho_{ij}^{perm}| \geq |\rho_{ij}^{obs}|$.

Derivation of edge scores in the metabolic network-2

1. The aggregated(ordered) $\frac{p(p-1)}{2}$ p-values based on permutation test of the correlations in the metabolic network are $P_{(1)}, P_{(2)}, \dots, P_{(\frac{p(p-1)}{2})}$

Derivation of edge scores in the metabolic network-2

1. The aggregated(ordered) $\frac{p(p-1)}{2}$ p-values based on permutation test of the correlations in the metabolic network are $P_{(1)}, P_{(2)}, \ldots, P_{(\frac{p(p-1)}{2})}$
2. $P_i \sim \lambda B(b,1) + (1-\lambda)B(1,1), 0 < b < 1, 0 \leq \lambda \leq 1, i = 1, \ldots, \frac{p(p-1)}{2}$

Derivation of edge scores in the metabolic network-2

1. The aggregated(ordered) $\frac{p(p-1)}{2}$ p-values based on permutation test of the correlations in the metabolic network are $P_{(1)}, P_{(2)}, \ldots, P_{(\frac{p(p-1)}{2})}$
2. $P_i \sim \lambda B(b,1) + (1-\lambda)B(1,1), 0 < b < 1, 0 \leq \lambda \leq 1, i = 1, \ldots, \frac{p(p-1)}{2}$
3. An adjusted log likelihood ratio edge score with a threshold P-value-$\tau_2(FDR_2)$:

$$\begin{aligned} S^{FDR_2}(z) &= \log(\frac{bz^{b-1}}{b\tau_2^{b-1}}) \\ &= (b-1)(\log(z) - \log(\tau_2(FDR_2))) \end{aligned} \tag{3}$$

Multiplicity in multiple testing of the nodes and edges in metabolic network

1. In order to adjust the multiplicity in multiple testing of both the nodes and edges in metabolic network, we set overall FDR level in the multiple testing in all the nodes and the edges-$FDR = \alpha$.

Multiplicity in multiple testing of the nodes and edges in metabolic network

1. In order to adjust the multiplicity in multiple testing of both the nodes and edges in metabolic network, we set overall FDR level in the multiple testing in all the nodes and the edges-$FDR = \alpha$.
2. Equally split the overall FDR level α to the nodes and the edges such that $FDR_1 = FDR_2 = \alpha/2$

Optimization of both node scores and edge scores in metabolic network

Our objective is to identify optimized sub-network(potential functional module) through maximizing the sum of both node scores and edge scores in the network:

Optimization of sum of node scores and edge scores

$$S_{subnet}^{FDR} = S_{subnet}^{FDR_1} + S_{subnet}^{FDR_2} = \Sigma_{x_i \in subnet} S^{FDR_1}(x_i) + \Sigma_{z_{ij} \in subnet} S^{FDR_2}(z_{ij}) \tag{4}$$

An introduction of heuristic optimization algorithm

1. We propose an algorithm that is adapted from the BioNet library implemented in R (Dittrich et al. 2010).

An introduction of heuristic optimization algorithm

1. We propose an algorithm that is adapted from the BioNet library implemented in R (Dittrich et al. 2010).
2. The revised algorithm is developed to identify an optimized subnetwork by maximizing the sum of both node scores and edge scores in the network.

An introduction of heuristic optimization algorithm

1. We propose an algorithm that is adapted from the BioNet library implemented in R (Dittrich et al. 2010).
2. The revised algorithm is developed to identify an optimized subnetwork by maximizing the sum of both node scores and edge scores in the network.
3. Key changes in our proposed algorithm when compared to that implemented in Dittrich et al. (2010) is that our methods are based on a combination of both node and edge scores.

Simulation scenarios

Simulate the network data with two random graph models with the number of nodes same as the number of nodes from a real metabolic network data-472:

1. Erdös-Rényi random graph (ER) model (Erdös, P., Rényi, A. 1959) with different connecting probabilities-$p = 0.1, 0.3$.

Simulation scenarios

Simulate the network data with two random graph models with the number of nodes same as the number of nodes from a real metabolic network data-472:

1. Erdös-Rényi random graph (ER) model (Erdös, P., Rényi, A. 1959) with different connecting probabilities-$p = 0.1, 0.3$.
2. Barabási-Albert random graph (BA) model (Barabási A., Albert R. 1999) with degrees of power-$power = 1, 3$ with approximately same corresponding number of edges as Erdös-Rényi random graph model with $p = 0.1, 0.3$.

Comparison of other score based optimization methods to detect optimized subnetwork

Our proposed heuristic optimization method will compare to other two similar methods:

1. the heuristic optimized subnetwork detection algorithm(Dittrich-node) based only on the node scores of the network from Ditrrich et. al. (2008, 2010)

Comparison of other score based optimization methods to detect optimized subnetwork

Our proposed heuristic optimization method will compare to other two similar methods:

1. the heuristic optimized subnetwork detection algorithm(Dittrich-node) based only on the node scores of the network from Ditrrich et. al. (2008, 2010)
2. the exact optimized subnetwork detection algorithm(Dittrich-nodeedge) based on both node scores and edge scores of the network proposed in Ditrrich et.al. (2012).

Consistency of our optimization method to detect optimized subnetwork

Node scores$\sim$ *Uniform*$(-2, 2)$,edge scores$\sim$ *Uniform*$(-4, 4)$.
Percentage: average percentage of size of detected optimized subnetwork versus the total number of nodes

Methods/Percentage	ER,p=0.1	ER,p=0.3	BA,power=1($p \approx 0.1$)	BA,power=1($p \approx 0.3$)	BA,power=3($p \approx 0.1$)	BA,power=3($p \approx 0.3$)
Our proposed method	0.024	0.028	0.026	0.029	0.032	0.036
Dittrich-node	0.013	0.017	0.014	0.026	0.024	0.027
Dittrich-nodeedge	0.021	0.024	0.023	0.027	0.029	0.031

Efficiency of our optimization method to detect optimized subnetwork

Repeated simulation number$=$ 100,
3 positive scoring cluster with the size-50
First cluster-node scores$\sim$ $Uniform(0.1, 2)$,edge scores$\sim$ $Uniform(0.1, 4)$;
Second cluster-node scores$\sim$ $Uniform(0.1, 2)$,edge scores$\sim$ $Uniform(-4, 4)$;
Third cluster-node scores$\sim$ $Uniform(0.1, 2)$,edge scores$\sim$ $Uniform(-4, -0.1)$;
Other nodes with scores$\sim$ $Uniform(-2, -0.1)$, edges connecting any two nodes in 3 positive scoring clusters$\sim$ $Uniform(-4, -2.1)$, other edges with scores$\sim$ $Uniform(-4, 4)$
Percentage1: average percentage of the number of nodes within detected optimized subnetwork in 1st positive scoring cluster versus the size of 1st positive scoring cluster
Percentage2: average percentage of the number of nodes NOT within detected optimized subnetwork but in other two positive scoring cluster versus the total size of other two positive scoring cluster

Accuracy and specificity of our optimization method to detect optimized subnetwork

Table: Accuracy of our method comparing to other two methods

Methods/Percentage1	ER,p=0.1	ER,p=0.3	BA,power=1($p \approx 0.1$)	BA,power=1($p \approx 0.3$)	BA,power=3($p \approx 0.1$)	BA,power=3($p \approx 0.3$)
Our proposed method	0.896	1	1	1	1	1
Dittrich-node	1	1	1	1	1	1
Dittrich-nodeedge	1	1	1	1	1	1

Table: Specificity of our method comparing to other two methods

Methods/Percentage2	ER,p=0.1	ER,p=0.3	BA,power=1($p \approx 0.1$)	BA,power=1($p \approx 0.3$)	BA,power=3($p \approx 0.1$)	BA,power=3($p \approx 0.3$)
Our proposed method	0.658	0.359	0.647	0.434	0.617	0.376
Dittrich-node	0.255	0.25	0.215	0.195	0.267	0.261
Dittrich-nodeedge	0.758	0.513	0.714	0.627	0.782	0.543

Summary

1. Our proposed optimized method accounts for statistical strength of both association of metabolites(nodes) with the outcome and also uncertainties of the correlations(edges) in the process of deriving the optimal subnetwork.

Summary

1. Our proposed optimized method accounts for statistical strength of both association of metabolites(nodes) with the outcome and also uncertainties of the correlations(edges) in the process of deriving the optimal subnetwork.
2. An heuristic optimized algorithm in R is developed and is compared to two closely related algorithms in the simulation.

Summary

1. Our proposed optimized method accounts for statistical strength of both association of metabolites(nodes) with the outcome and also uncertainties of the correlations(edges) in the process of deriving the optimal subnetwork.
2. An heuristic optimized algorithm in R is developed and is compared to two closely related algorithms in the simulation.
3. Based on simulation results, our proposed algorithm is valid and fairly close to the exact optimized algorithm based on both node scores and edge scores in specificity of detecting optimized subnetwork.

Summary

1. Our proposed optimized method accounts for statistical strength of both association of metabolites(nodes) with the outcome and also uncertainties of the correlations(edges) in the process of deriving the optimal subnetwork.
2. An heuristic optimized algorithm in R is developed and is compared to two closely related algorithms in the simulation.
3. Based on simulation results, our proposed algorithm is valid and fairly close to the exact optimized algorithm based on both node scores and edge scores in specificity of detecting optimized subnetwork.
4. Based on simulation results, our proposed algorithm have higher specificity in detecting optimized subnetwork than the one based on node scores only.

Future work

1. Run simulation compared our algorithm to other score based optimization methods and other types of the algorithms to detect potential functional module in the network

Future work

1. Run simulation compared our algorithm to other score based optimization methods and other types of the algorithms to detect potential functional module in the network
2. Run simulation under other random graphical models such as Gaussian graphical model and stochastic blockmodel

Future work

1. Run simulation compared our algorithm to other score based optimization methods and other types of the algorithms to detect potential functional module in the network
2. Run simulation under other random graphical models such as Gaussian graphical model and stochastic blockmodel
3. Apply our proposed algorithm to real datasets such as cardiovascular disease metabolomics study from Women's Health Initiative cohort study.

References-1

- Barabási, Albert-László; Albert, Réka
  Emergence of scaling in random networks.
  Science, 286 (5439): 509-512, 1999.

- Beisser,D., Dittrich MT et al.
  Bionet: an R-package for the functional analysis of biological networks
  Bioinformatics, 26, 1129-1130, 2010.

- Beisser,D., Dittrich MT et al.
  Robustness and accuracy of functional modules in integrated network analysis.
  Bioinformatics, 28, 1887-1894, 2012.

- Dittrich MT et al.
  Identifying functional modules in protein-protein interaction networks: an integrated exact approach.
  Bioinformatics, 24, i223-i231, 2008.

References-2

- Erdös, P., Rényi, A.
  On Random Graphs.
  Publicationes Mathematicae, 6: 290-297,1959

- Phipson B, Smyth GK.
  Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn.
  Stat Appl Genet Mol Biol, 9,1544-6115,2010

- Guo,Z. et al.
  Edge-based scoring and searching method for identifying condition responsive protein-protein interaction sub-network.
  Bioinformatics, 23, 2121-2128,2007

- Ideker,T. et al.
  Discovering regulatory and signalling circuits in molecular interaction networks.
  Bioinformatics, 18(Suppl. 1), S233-S240,2002