Graph Theory Applications in Cancer Research

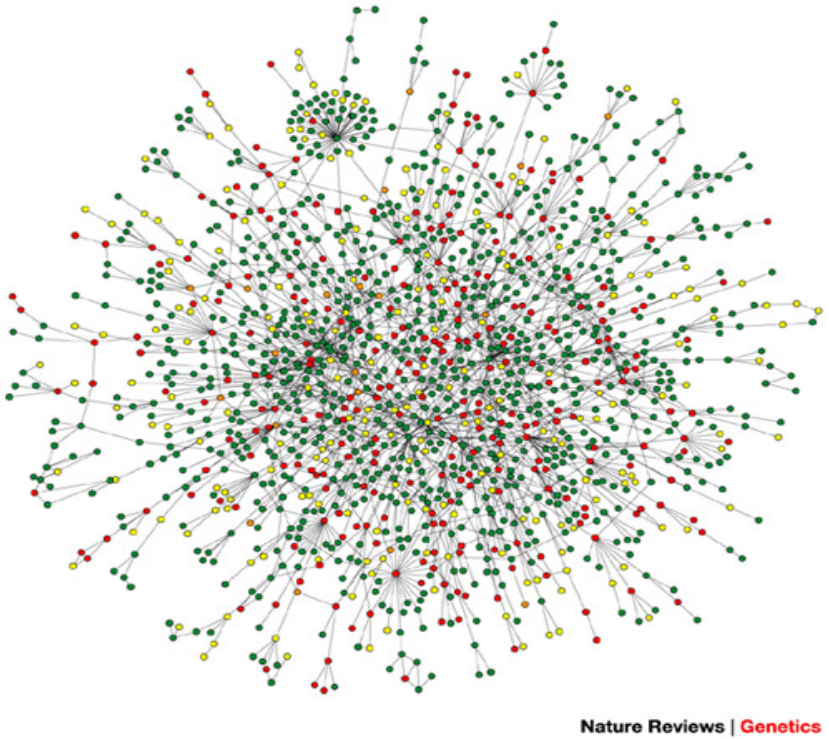
Cierra W. Zaslowe

December 15, 2015

Introduction

- Why cancer research?
- Why graph theory?
- Cross-disciplinary aspects.

Protein Protein Interaction Network of a Yeast Cell



The Data

- Micro-arrays

The Problem

- Missing Data

The Connection

- How Graph Theory uses a Protein Protein Interaction Network to infer information about genetic defects and information about different types of cancer.

Cut Based Approach - Karoaz *et al*

- Main Idea
  - Each node in the network can be one of three states:
    $+1$ if the protein is annotated with the function f
    -1 if the protein is annotated with another function
    0 if the proteins function is hypothetical
  - Each edge has real valued weight.
  - The goal of this method is to give an assignment of $+1$ or -1 to the unannotated proteins to see if they share the same function (an assignment of $+1$). Want two connected vertices to share the same function.

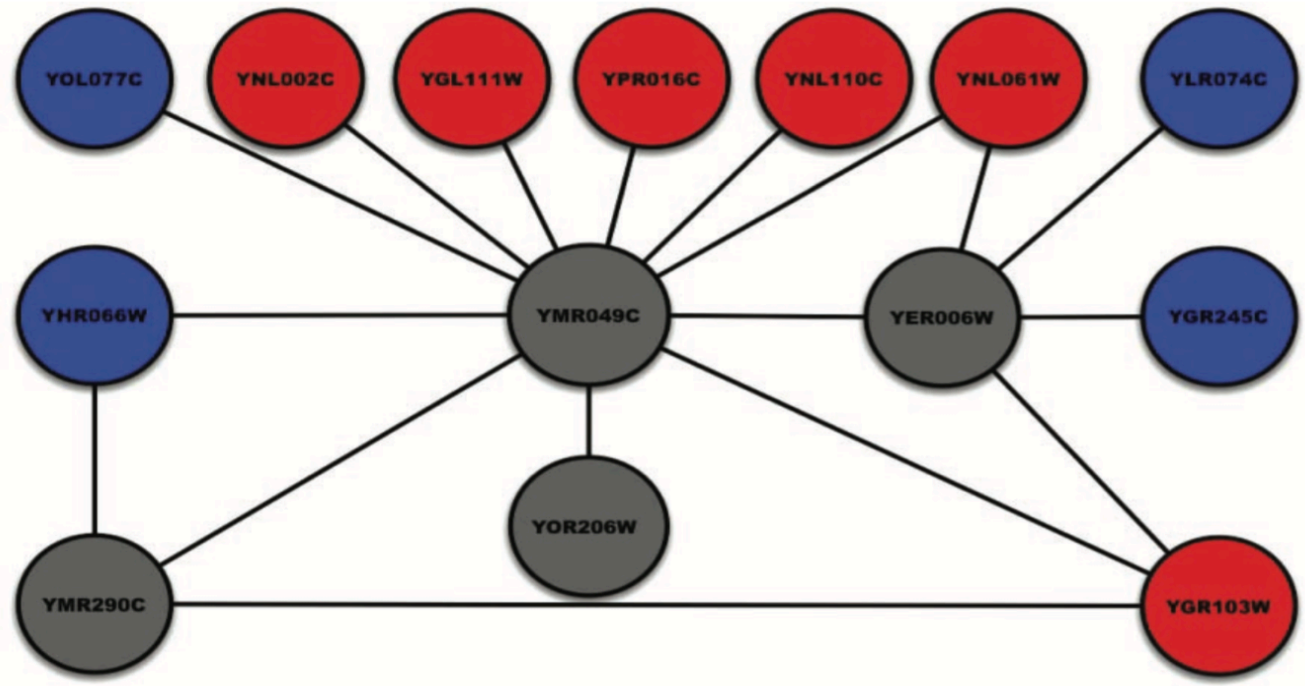
Cut Based Approach - Karoaz *et al cont.*

- Methodology and Results

$$E = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} w_{ij} s_i s_j$$

Where n is the number of vertices in the graph.
w_{ij} is the weight of the edge connecting proteins i and j
s_i is the state assigned to protein i.

Cut Based Approach - Karoaz *et al cont.*



Cut Based Approach - Karoaz *et al cont.*

- The algorithm used requires an "activation rule" iteratively until until further application makes no changes (convergence).
- $s_i = sgn\left(\sum_{1 \leq j \leq n_i} w_{ij} s_j - \theta \right)$
  Where n_i is the number of neighbors of i
  θ is the activation threshold.
- Flaws and Advantages

Nabieva *et al*

- Main Idea
  Undirected graph: $G = (V, E)$, where there is a vertex $v \in V$ for each protein, and an edge between vertices u and v if the corresponding proteins are known to interact.
- FunctionalFlow
  - To Obtain Functional Score
  - Run this process for each biological function in turn, and obtain for each protein, its functional score.
  - The highest score indicates the function of that protein. Note that functional score is the amount of flow in each reservoir.

Nabieva *et al cont.*

- For each protein u define $R_t^a(u)$, which is the amount in the reservoir for function a at time t.
- When $t = 0$

$$R_0^a(u) = \begin{cases} \infty & \text{If } u \text{ is annotated by } a \\ 0 & \text{Otherwise} \end{cases}$$

- When $t > 0$

$R_t^a(u) = R_{t-1}^a + \sum_{v:(u,v)\in E}(g_t^a(v,u) - g_t^a(u,v))$ Where $g_t^a(u,v)$ is the amount of flow on the edges.

- Initially at $t = 0$, $g_0^a(u,v) = 0$
- When $t > 0$

$$g_t^a(u,v) = \begin{cases} 0 & \text{If } R_{t-1}^a(u) < R_{t-1}^a(v) \\ min(w_{u,v}, \frac{w_{u,v}}{\sum_{(u,y)\in E} w_{u,y}}) & \text{Otherwise} \end{cases}$$

Nabieva *et al cont.*

- The functional score for vertex u and function a over d iterations is calculated as the total amount of flow that has entered the vertex.

$$f_a(u) = \sum_{t=1}^{d} \sum_{v:(u,v)\in E} g_t^a(v,u)$$

Nabieva *et al cont.*

- Results
- Flaws and Advantages

The Goal of Md Jamiul Jahid and Jianhua Ruan

- Find a series of intermediate proteins (biomarkers)
- Others can see if these biomarkers have influence on cancer growth.
- Note: Proteins signify gene expression; the Protein Protein Interaction network still applies.

Steiner Trees

- Steiner Trees: A minimum weight spanning tree that connects a set of vertices. The trees may include other vertices outside this set, known as Steiner Vertices.

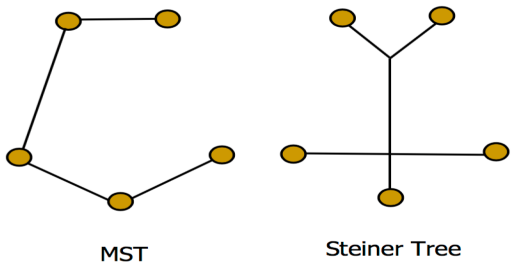


Figure: A Minimum Spanning Tree (MST) and a Steiner Tree

Randomized Steiner Tree Based Approach

- $G = (V, E, w)$
- Subset of vertices $R \subseteq V$
- Vertices $U \subseteq V$, and edges $S \subset E$.
- Here the vertices in R are known as the terminal vertices (necessary vertices) and $U \backslash R$ as Steiner vertices.
- Forrest T' comprising of the terminal vertices R.
- Each iteration the algorithm finds the two vertices in T' that are closest in distance and adds the intermediate vertices to T'.
- Repeated until T' becomes connected.
- Minimum spanning tree of those selected vertices are built and all leaf vertices that are not in R are removed.

Pseudo Code

Steiner tree algorithm
Input: Weighted PPI network, $G = (V, E, w)$; DE genes (proteins), R
Output: A minimum spanning tree, T, that spans R.

1. Start with a forest (T') comprising the DE genes, notated by R, but no edges.
2. While $T'$is not a tree do connect two shortest- distance disconnected vertices $u, v \in T'$ and add vertices on the path to T'
3. Build a minimum spanning tree (T) with the subgraph of G induced by the vertices in T'
4. Delete any leaf node in T that is not in R

Any intermediate vertices are considered potential biomarkers for breast cancer metastasis.

Conclusions on the Method by Md Jamiul Jahid and Jianhua Ruan

- Results
  - A new way to identify biomarkers in breast cancer.
  - Can be applied to other genetic diseases.
- Flaws and Advantages
  - Reproducibility

Known Breast Cancer mutations discovered by the Steiner Tree based Algorithm

Table 3 Breast cancer genes in STMs

Breast cancer known genes in STMs (van de Vijver dataset)	Breast cancer known genes in STMs (Wang dataset)
RAD54L	*HRAS*
HRAS	*ITGA2*
ERBB2	*BRCA2*
BRCA2	*BRCA1*
PGR	*APC*
XRCC1	*KRAS*
BRCA1	*ITGB3*
PHB	*ESR1*
TYMS	*TP53*
TNF	*TGFB1*
APC	*VDR*
ESR1	*AR*
TP53	*RAD51*
PIK3CA	*TSG101*
TGFB1	*CDH1*
GSTP1	
GSTT1	
LOC651610	
ATM	
RAD51	
TK1	
CYP1A1	
TSG101	
CHD1	

Conclusion

- What this means for cancer research
- Final Thoughts

Sources

References

[1] "Dictionary.com Unabridged". In: (2015). URL: http://dictionary.reference.com/browse/biomarker.

[2] Amit Agarwal Bernard Chazelle Elena Nabieva Kam Jim and Mona Singh. "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps." In: *Bioinformatics* 21 (2005), pp. i302 –i310. DOI: 10.1093/bioinformatics/bti1054.

[3] Valery Trifonov Haiyuan Yu Alberto Paccanaro and Mark Gerstein. "Predicting interactions in protein networks by completeing defective cliques." In: *Bioinformatics* 22.7 (2006), pp. 823–829. DOI: 10.1093/bioinformatics/btl014.

[4] Md Jamiul Jahid and Jianhua Ruan. "A Steiner tree-based method for biomarker discovery and classification in breast cancer metastasis." In: *BMC Genomics* 13.S6:S8 (2011). DOI: http://www.biomedcentral.com/1471-2164/13/S6/S8.

[5] Pamela K. Kreeger and Douglas A. Lauffenburger. "Cancer Systems Biology: a network modeling perspective." In: *Carcinogenesis* 31.1 (2010), pp. 2–8. DOI: 10.10933/carcin/bgp261.

[6] Ricki Lewis. *Human Genetics, Concepts and Applications*. New York, NY: McGraw Hill Education, 2015.

[7] Avi Ma'ayan. "Introduction to Network Analysis in Systems Biology." In: *NIH Public Access* 190.4 (), pp. 100–100.

[8] Nature.com. *Nature.com subject areas*. URL: http://www.nature.com/subjects/protein-protein-interaction-networks.

[9] Igor Ulitsky Roded Sharan and Ron Shamir. "Network-based prediction of protein function." In: *Molecular Systems Biology* 3.88 (2007). DOI: 10.1038/msb4100129.

[10] Uconn. *Steiner Trees*. URL: http://www.engr.uconn.edu/~tehrani/teaching/cad/07_steiner.pdf.

[11] Stan Letovsky Yu Zheng Chunming Ding Charles R. Cantor Ulas Karaoz T. M. Murali and Simon Kasif. "Whole-genome annotation by using evidence integration in functional-linkage networks." In: *PNAS* ().

[12] Wikipedia. *Crosstalk (biology) — Wikipedia, The Free Encyclopedia*. [Online; accessed 28-November-2015]. 2015. URL: \url{https://en.wikipedia.org/w/index.php?title=Crosstalk_(biology)&oldid=679077216}.

[13] Wikipedia. *Steiner Tree Problem*. URL: https://en.wikipedia.org/wiki/Steiner_tree_problem#Generalization_of_minimum_Steiner_tree.