NETWORK TOPOLOGY AND FUNCTIONAL ENRICHMENT ANALYSIS OF GWAS-DERIVED GENES REVEAL HUB PROTEINS AND PATHWAY MODULES IN TYPE 1 DIABETES

SATYAVATHI DRONAMRAJU[1] AND M. SIVA PARVATHI [2]

Abstract. Genome-wide association studies (GWAS) have identified numerous loci associated with Type 1 Diabetes (T1D), yet the functional relationships among implicated genes remain incompletely understood. In this study, we applied a network-based systems biology framework to investigate the topological and functional organization of GWAS-derived T1D genes. Genes associated with significant GWAS SNPs were retrieved from the GWAS Catalog and mapped onto a protein–protein interaction (PPI) network using data from the STRING database. Graph-theoretic metrics—including degree centrality, betweenness centrality, and clustering coefficient—were used to identify hub proteins and key network modules. Functional enrichment analyses were conducted using Gene Ontology and pathway databases, revealing significant overrepresentation of immune response, cytokine signaling, and pancreatic β-cell–related processes. Several highly connected hub genes correspond to known T1D susceptibility genes, while others represent novel candidates potentially involved in disease pathogenesis. These findings demonstrate that network topology provides critical insight into the biological mechanisms underlying T1D and highlight candidate molecular targets for further experimental validation.

1. Introduction

Glioblastoma multiforme (GBM) represents the most malignant form of astrocytic tumors as it is characterized by rapid proliferation, and has a higher incidence observed in older patient populations. Despite extensive sequencing efforts, patient prognosis remains poor, highlighting the insufficiency of approaches that focus solely on individual genes or mutations. GBM is increasingly understood as a complex, emergent phenomenon arising from dysregulated interactions among genes, proteins, and pathways.

Graph theory provides a natural mathematical language for capturing such interactions. By representing genes as nodes and coexpression relationships as edges, gene coexpression networks encode collective behavior that is invisible to single-gene analyses. Differences between healthy and malignant tissues can thus be framed as differences in network topology. This perspective aligns with systems biology, where disease states correspond to altered network organization rather than isolated molecular defects.

This paper investigates whether malignancy leaves reproducible mathematical signatures in gene coexpression networks and whether such signatures can be quantified using established tools from graph theory and spectral analysis.

2. Application of Graph Networks

Mathematical and computational models are increasingly being used in various disciplines of science due to the limitations, both scientific and ethical, of in vivo and in vitro studies. In this study, research was conducted in silico, to avoid such scientific and ethical concerns, with glioblastoma multiforme and normal brain gene expression data. After copious research and testing, several key graph theoretic analytical signatures each revealing a unique characteristic of a gene coexpression network were identified to better elucidate the differences between a cancer and normal network.

2010 *Mathematics Subject Classification.* 05C69.

Key words and phrases. Gene coexpression, Pearson's Correlation Coefficient, Mann-Whitney U test.

[1] Corresponding author.

The first signature we will discuss is degree distribution and power law exponent. We hypothesize that cancer networks follow scale-free behavior. Scale-free behavior is that which follows the power-law degree distribution [1]. Scale-free networks also demonstrate preferential attachment growth models, meaning that as the power-law degree distribution is established, it only gets exacerbated in a positive feedback loop as dysregulation causes more dysregulation in the hub regulatory nodes [2]. According to our hypothesis, cancer-affected regulatory hub genes are overexpressed causing unchecked cellular growth culminating in cancer [3]. The power-law exponent gamma should also be lower in cancer networks as low gamma values mean heavier tails, more power-law tendencies, and presence of hub nodes [1].

The second signature we will discuss is clique cover number and Ramsey theory. Clique cover number is defined as the number of complete subgraphs to cover all vertices [4]. Ramsey theory reveals that in any sufficiently large graph, a large clique or independent set must be present [5], and furthermore that a predictable and specific substructure will emerge [6]. We hypothesize cancer networks have a higher clique cover number, insinuating dense but aberrant interconnectivity. This is due to increased gene expression and mutations aiding metabolic rewiring and immune system evasion archetypal of cancer. Integrating Ramsey theory, we further conclude that we will see a clear substructure in the graph of density versus clique size in the cancer network.

The third signature we will discuss is k-core decomposition. K-core decomposition measures core depth, revealing insights on the balance of connectivity. K-core decomposition works by removing nodes with a degree lower than k until all remaining nodes are k or higher [7]. Hierarchical structure is also revealed as central and peripheral nodes are differentiated [8]. This can help identify networks that are deeply co-regulated [9]. We hypothesize that in cancer networks, global fragmentation is high due to mutations breaking interactions, hub gene dysregulation, and sparse rewiring of the network. Furthermore, there is noisier gene expression, antagonistic and spurious gene expression, and loss of functional redundancy due to reliance on few specific pathways.

The fourth and final signature to be discussed is Laplacian eigenvalue spectrum. The Laplacian matrix is found by the difference between the degree matrix and the adjacency matrix, and in our case we implemented the signed Laplacian due to the presence of negative weights [10]. The algebraic connectivity measures robustness and cohesion of the graph [11], while the multiplicity of 0 eigenvalue reveals the number of connected components. Negative eigenvalues imply presence of antagonistic modules and conflicting subgraphs [12]. Spectral moments reflect pathway redundancy [13]. We hypothesize that in cancer there will be a lower algebraic connectivity implying weak global connectivity. We further hypothesize that there will be more negative eigenvalues in cancer networks due to antagonistic modules and conflicting subgraphs.

The aim of our paper is to analyze these mathematical signatures and test significance through high-throughput analysis. If successful, it may be possible to predict and take preventative measures on people found to have cancer like gene-expression signatures. This has the potential to revolutionize genetic analysis and better the prognosis of millions of cancer patients across the world.

Table 1. Comparison of Hypothesized Graph Features

Metric	Cancer Network (Glioblastoma)	Normal Brain Network
Degree Distribution & Power-Law	Heavier tail; lower γ ($\gamma \approx 2.0$–2.5); more hubs	Lighter tail; higher γ ($\gamma \approx 2.5$–3.5); fewer hubs
Clique Cover & Ramsey Structure	Higher clique cover; overlapping, conflicting communities	Lower clique cover; modular but cleaner community boundaries
k-Core Decomposition	Higher max coreness; deeper, tightly nested cores	Lower coreness; flatter, less nested hierarchy
Laplacian (or Signed) Spectrum	Spread out, with possible **negative** eigenvalues (due to dysregulation)	Tightly clustered, all positive eigenvalues; highly cohesive

3. Materials and Methods

First, we obtain data from TCGA [14] and GTEX [15] data sources, the former referring to cancer gene expression data and the latter referring to normal gene expression data. This allows

for a comparative analysis between cancer and normal and better demonstrates the effectiveness of mathematical signatures in cancer diagnostics and prevention. Next, we use NetworkX and Python to run a gene co-expression analysis. The code will select random genes from the cancer data and find the same genes in the normal data. After filtering unwanted genes and doing relevant log transformations, it will then take the Pearson correlation coefficient of every co-expression pair. Correlations that are significant with an alpha of 0.05 will be added to the graph network. False discovery and outliers will be eliminated using Benjamini-Hochberg adjustment [16]. Finally, after the networks for both cancer and normal are finalized, then graph theory theorems will be used to analyze and compare the networks.

In this section we formally introduce four key graph-theoretic signatures—degree distribution and its power-law exponent, clique cover number and Ramsey theory, k–core decomposition, and the signed Laplacian eigenvalue spectrum—and present the main definitions and theorems that underpin their use in distinguishing cancer from normal gene coexpression networks.

3.1. **Degree Distribution in Networks.** The degree of a node i, denoted k_i, is the number of edges incident on that node. The degree distribution $P(k)$ is the probability that a randomly selected node has degree exactly k. In many real-world networks (biological, social, technological), the tail of $P(k)$ often follows a power law:

$$P(k) \sim C\, k^{-\alpha}, \quad k \geq k_{\min},$$

where $\alpha > 1$ is the power-law exponent, $k_{\min}$ is the lower bound above which the power law holds, and C is a normalization constant:

$$C = \Big(\sum_{k=k_{\min}}^{\infty} k^{-\alpha} \Big)^{-1}.$$

3.2. **Maximum-Likelihood Estimation (MLE) of α.** Given observed degrees $\{k_i\}$ with $k_i \geq k_{\min}$ for $i = 1, \ldots, n$, the MLE of α for discrete data is given by [17]:

$$\widehat{\alpha} = 1 + n \Bigg[\sum_{i=1}^{n} \ln\Big(\frac{k_i}{k_{\min} - 0.5} \Big) \Bigg]^{-1}.$$

This estimator minimizes the negative log-likelihood

$$\mathcal{L}(\alpha) = -\, n \ln C(\alpha, k_{\min}) + \alpha \sum_{i=1}^{n} \ln k_i.$$

In practice, one selects $k_{\min}$ by minimizing the Kolmogorov—Smirnov distance between the empirical cumulative distribution function (CDF) and the fitted power law CDF.

3.3. **Goodness of Fit p-Value via Bootstrapping.** To test whether the power law is a plausible fit:

(1) Fit $\widehat{\alpha}$ and record the Kolmogorov–Smirnov statistic D_{obs}.
(2) Generate many synthetic datasets of size n drawn from a power law with exponent $\widehat{\alpha}$ and the same $k_{\min}$.
(3) Refit each synthetic dataset, computing its KS statistic D_{synth}.
(4) The p-value is the fraction of synthetic D_{synth} that exceed D_{obs}:

$$p = \frac{\#\{D_{\text{synth}} \geq D_{\text{obs}}\}}{\text{number of simulations}}.$$

If $p > 0.1$, the power law is considered a plausible model; if $p < 0.1$, it is rejected.

3.4. **Log-Likelihood Ratio Tests.** To compare the power-law model to an alternative heavy-tailed distribution (e.g., exponential or log-normal), consider two competing PDFs $p_1(k)$ and $p_2(k)$. Compute their log-likelihoods on the same data $\{k_i\}$:

$$\mathcal{L}_1 = \sum_{i=1}^{n} \ln p_1(k_i), \quad \mathcal{L}_2 = \sum_{i=1}^{n} \ln p_2(k_i).$$

The log-likelihood ratio is

$$R = \mathcal{L}_1 - \mathcal{L}_2.$$

A positive R indicates support for model 1 (e.g., power law), while $R < 0$ favors model 2. Significance can be assessed via Vuong's method or by bootstrap sampling under the null model [18].

3.5. **Clique Cover Number.** For a graph $G = (V, E)$, a *clique cover* is a collection of cliques $\{C_1, \ldots, C_k\}$ such that every vertex $v \in V$ lies in at least one C_i. The **clique cover number** $\theta(G)$ is the minimum size of such a collection:

$$\theta(G) = \min\Big\{k : \exists\, C_1, \ldots, C_k \text{ all cliques, } V = \bigcup_{i=1}^{k} C_i\Big\}.$$

Since covering V by cliques in G corresponds to coloring the complement graph $\overline{G}$, one has

$$\theta(G) = \chi\big(\overline{G}\big),$$

where χ denotes the chromatic number. If $\omega(G)$ is the size of a maximum clique in G, then

$$\omega(G) \le \theta(G) \le |V|.$$

3.6. **Ramsey Theory.** The *Ramsey number* $R(s, t)$ is the smallest integer N such that every graph on N vertices contains either a clique of size s or an independent set of size t:

$$R(s,t) = \min\{N : \forall\, G \text{ on } N \text{ vertices}, \omega(G) \ge s \ \vee\ \alpha(G) \ge t\}.$$

Erdős and Szekeres proved the classical bound:

$$R(s,t) \le \binom{s+t-2}{s-1}, \quad \text{and in particular} \quad R(s,s) \le \binom{2s-2}{s-1} \sim \frac{4^{s-1}}{\sqrt{\pi(s-1)}}.$$

3.7. **K-core Decomposition.** Let $G = (V, E)$ be an undirected graph. The *k-core* of G is the maximal subgraph in which every vertex has degree at least k. Formally, define

$$C_k(G) \ = \ \max\big\{\, H = (V_H, E_H) \subseteq G : \delta(H) \ge k\big\},$$

where $\delta(H)$ denotes the minimum degree of subgraph H.

Each node $v \in V$ is assigned a *core number* $\phi(v)$, given by

$$\phi(v) \ = \ \max\big\{\, k : v \in C_k(G)\big\}.$$

Equivalently, $\phi(v)$ is the largest k such that v remains after iteratively pruning all nodes of degree less than k.

3.8. **Signed Graph and Laplacian: Theory and Definitions.** Let $G = (V, E^+, E^-)$ be a signed graph, where V is the set of vertices, and the edge set is partitioned into positive edges E^+ (e.g. activating interactions) and negative edges E^- (e.g. inhibitory interactions).

The *signed adjacency matrix* $A^s \in \mathbb{R}^{n \times n}$ is defined entrywise by

$$A^s_{ij} \ = \ \begin{cases} +1, & (i,j) \in E^+, \\ -1, & (i,j) \in E^-, \\ 0, & \text{otherwise.} \end{cases}$$

The *signed degree matrix* D^s is diagonal, with entries

$$D^s_{ii} \;=\; \sum_{j=1}^{n} \big|A^s_{ij}\big|,$$

so that each diagonal entry counts the sum of the absolute weights of edges incident on node i.

The *signed Laplacian* is then given by

$$L^s \;=\; D^s \;-\; A^s,$$

which is symmetric and positive semi-definite.

The eigenvalues of L^s encode structural balance properties of the signed graph:

$$0 \;=\; \lambda_1 \;\le\; \lambda_2 \;\le\; \cdots \;\le\; \lambda_n,$$

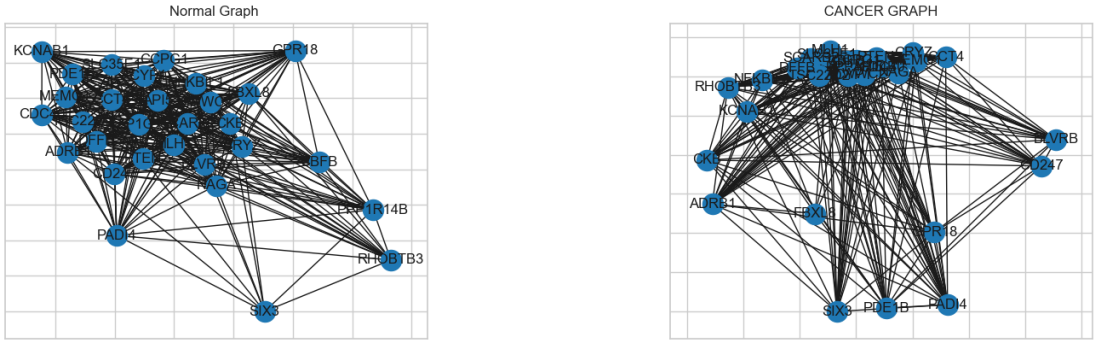
where each λ_i satisfies the eigenvalue equation

$$L^s v_i \;=\; \lambda_i v_i.$$

The multiplicity of the zero eigenvalue corresponds to the number of connected balanced components in G.

4. Results and Conclusion

Our results are as follow.

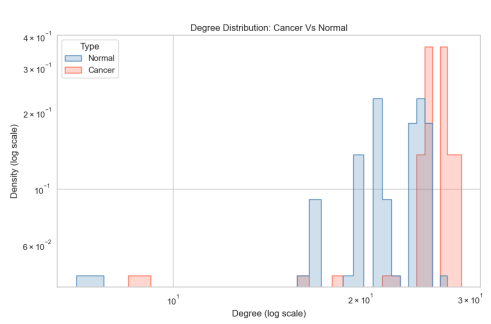


(A) Normal Network (B) Cancer Network

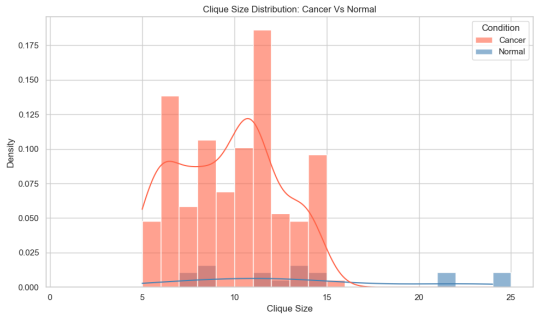
Figure 1. Comparison Between Two Gene Coexpression Networks.

4.1. **Analysis, Interpretations, and Conclusions.** In order to validate the aforementioned four key mathematical signatures of malignancy, we can first start with the figures, and then follow up with a statistical and tabular data analysis. The first set of figures is a comparison between cancer network and normal network.

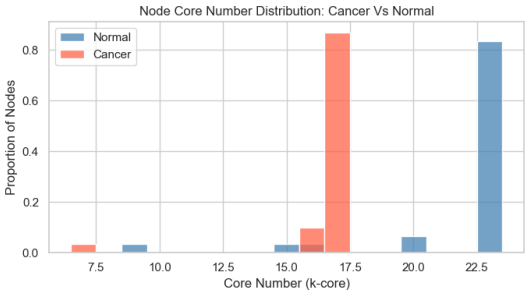
From this comparison, a few visual observations can be gleaned. The normal network is more balanced in edge weights than the cancer network. This is due to the fact that cancerous physiology at the cellular level involves overexpressed and mutations in central key regulatory genes. This overexpression prevents the cell cycle from working properly resulting in such malignancies. Furthermore, there are more overlapping and conflicting communities in the cancer network. This again, is related to graph theory. Clique Cover and Ramsey Theory state



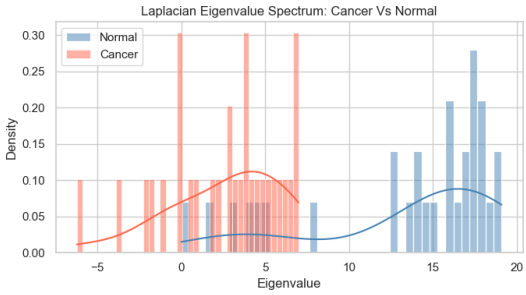
(A) Degree Distribution and Power Law



(B) Clique Cover Number and Ramsey Theory



(C) K-core Decomposition



(D) Laplacian Eigenvalue Spectrum

FIGURE 2. Four Mathematical Signatures of Malignancy.

REFERENCES

[1] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509.

[2] M. E. J. Newman. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics* 46.5 (2005), pp. 323–351. DOI: 10.1080/00107510500052444.

[3] Hawoong Jeong et al. "Lethality and centrality in protein networks". In: *Nature* 411.6833 (2001), pp. 41–42. DOI: 10.1038/35075138.

[4] Martin Charles Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Elsevier, 2004. DOI: 10.1016/B978-0-444-51595-5.X5000-8.

[5] Ronald L. Graham, Bruce L. Rothschild, and Joel H. Spencer. *Ramsey Theory*. 2nd. Wiley-Interscience, 1990. DOI: 10.1002/9781118030894.

[6] P. Erdös and G. Szekeres. "A combinatorial problem in geometry". In: *Compositio Mathematica* 2 (1935), pp. 463–470. URL: http://eudml.org/doc/88611.

[7] Stephen B. Seidman. "Network structure and minimum degree". In: *Social Networks* 5.3 (1983), pp. 269–287. DOI: 10.1016/0378-8733(83)90028-X.

[8] Vladimir Batagelj and Matjaž Zaveršnik. *Fast algorithms for determining (generalized) core groups in social networks*. 2003. arXiv: cs/0310049 [cs.SI]. URL: https://arxiv.org/abs/cs/0310049.

[9] Stefan Wuchty, Zoltan N. Oltvai, and Albert-László Barabási. "Evolutionary conservation of motif constituents in the yeast protein interaction network". In: *Nature Genetics* 35.2 (2003), pp. 176–179. DOI: 10.1038/ng1242.

[10] Jérôme Kunegis et al. "Spectral analysis of signed graphs for clustering, prediction and visualization". In: *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 2010, pp. 559–570. DOI: 10.1137/1.9781611972801.48.

[11] Miroslav Fiedler. "Algebraic connectivity of graphs". In: *Czechoslovak Mathematical Journal* 23.2 (1973), pp. 298–305. DOI: 10.1007/BF01608787.

[12] Yongtang Hou. "Bounds for the least Laplacian eigenvalue of a signed graph". In: *Czechoslovak Mathematical Journal* 55.3 (2005), pp. 701–710. DOI: 10.1007/s10587-005-0083-2.

[13] Fan R. K. Chung. *Spectral Graph Theory*. Vol. 92. CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997. DOI: 10.1090/cbms/092.

[14] Suhas V. Vasaikar et al. "LinkedOmics: Analyzing Multi-Omics Data Within and Across 32 Cancer Types". In: *Nucleic Acids Research* 46.D1 (2018), pp. D956–D963. DOI: 10.1093/nar/gkx1090.

[15] GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (2020), pp. 1318–1330. DOI: 10.1126/science.aaz1776.

[16] Yoav Benjamini and Yosef Hochberg. "Power-Law Distributions in Empirical Data". In: *Royal Statistical Society* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[17] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111.

[18] Quang H. Vuong. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". In: *Econometrica* 57.2 (1989), pp. 307–333.

Department of Applied Mathematics, Sri Padmavati Mahila Visvavidyalayam, Tirupati –517502, Andhra Pradesh, India

Email address: [1]satyadronamraju11@gmail.com

Email address: [2]parvatimani2008@gmail.com