JOURNAL OF
EMERGING INVESTIGATORS

Article

Applying centrality analysis on a protein interaction network to predict colorectal cancer driver genes

Pia Saha[1], Gabriel Cerono[2]

[1] Stephen F. Austin High School, Sugar Land, Texas

[2] Department of Neurology, University of California - San Francisco, San Francisco, California

SUMMARY

The colorectal cancer tumor microenvironment presents significant genetic heterogeneity with mutations in genes in several signaling pathways. Detecting these driver genes through wet lab experiments is costly and time-consuming. Computational models and bioinformatic tools have become a vital alternative in this effort. One of these novel computational methods, Centrality Analysis, models molecular functions, biological processes and biochemical pathways by creating and analyzing protein-protein interaction networks. Centrality Analysis is an approach to quantify node (in this case, protein) importance in these networks. Essential proteins play critical roles in cell function; therefore, centrality measures serve as a basis to study the relationship between lethality and essentiality by evaluating the topological features of the network. However, there is no established standard to determine the most appropriate centrality measure for analyzing a specific network. The choice of a suitable set is complicated by the impact of network topology because results vary based on network structure, correlation among the selected set of measures, and network data collection methods used. We hypothesized that centrality scores can be used to predict driver genes while statistical and machine learning analyses can identify the relevant centrality features for this task. We proposed different analyses to select a valid set of centrality algorithms to predict driver and non-driver genes. We first recreated a protein-protein interaction network for colorectal cancer featuring known driver and passenger genes, and then compared the centrality scores of eight different algorithms using statistical analysis. We further validated the results by implementing machine learning models. Both analyses identified betweenness and closeness centrality algorithms as most important to predict driver versus non-driver genes.

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer death in the United States (1). Early detection and treatment is essential for improving patient outcomes and CRC morbidity and mortality can be alleviated with the development of novel targeted therapies for CRC management and biomarkers for screening programs. Both are dependent on the accurate identification of driver genes, signaling pathways, and causal associations. Targeted therapies for CRC often fail clinical trials despite proving effective in preclinical studies (2). Resistance to targeted antitumor agents is believed to be due to the presence of several other mutations in the tumors. The CRC genetic landscape is highly heterogeneous due to multiple mutations accumulating over time (1). These mutations can involve multiple genes and can occur through a variety of mechanisms (2). For example, colorectal cancer development involves the contribution of multiple genes, Adenomatous Polyposis Coli, or *APC*, is a tumor suppressor gene that, when it loses function, can cause uncontrolled cell growth and the development of adenomatous polyps that can turn into cancerous growths. Tumor Protein 53 (*TP53*) plays pivotal roles in DNA repair and apoptosis, and its inactivation can lead to the multiplication of potentially cancerous cells with DNA abnormalities, facilitating tumor growth. *APC* and *TP53* are among the most common drivers of tumorigenesis (3). Additionally, the heterogeneity of colorectal tumors, with different regions of a single tumor having distinct genetic profiles, further contributes to the complexity of the mutations in this type of cancer.

Identifying key genes and mutations that drive tumorigenesis is therefore pivotal for developing new targeted therapies. To this end, sets of criteria have been developed to classify colorectal cancer into different subtypes based on their molecular characteristics. These criteria can include the presence or absence of certain gene mutations, the expression levels of certain genes, and the activity of certain signaling pathways. To date, the proposed molecular subtype classification criteria for CRC do not predict targets for therapeutic intervention comprehensively and further conventional experimental methods are cost prohibitive (4, 5). Subclassification, even when accounting for cancer pathways or driver-gene mutations, has not been able to predict drug responses. As a result, considerable scope remains in improving the accuracy of identifying key genes implicated in cell-cycle regulation and their influence at the cellular-network level.

Computational models and bioinformatics tools have rapidly evolved in their ability to analyze large and complex biological data sets as an alternate attempt to improve predictability (6-8). These computational methods use multiple features extracted from different sites such as function annotations and cellular localization among others to detect novel driver genes (9). These computational methods lack the ability to integrate orthogonal data from different sources and fall short in their modeling capabilities of highly complex networks of biochemical pathways. One model that can overcome these limitations is protein-protein interaction (PPI) networks. A PPI network simulates the tumor genetic structure by representing proteins as nodes and their functional relationships as edges. This framework is key to understanding how the CRC genetic landscape may function and respond to certain mutations in key genes.

There exist numerous methods and techniques that can be applied to PPI networks to gain insight into CRC pathophysiology (10, 11). One such method we apply is Centrality Analysis (CA). CA is a method used in network analysis to evaluate the importance or influence of individual nodes within a network, by measuring how central a node is in terms of its connections to other nodes. The relevance of CA in the context of biological network analysis rests on the "Centrality Lethality Hypothesis." It refers to the idea that highly connected proteins in a biological network are more critical or essential for the overall function of the network and therefore the removal of these central proteins, (their deletion), could have a more significant impact on the network's stability and overall functioning and even prove lethal, compared to less connected proteins (12). Moreover, studies have shown the importance of assessing the CA measures in totality to ascertain the essentiality of a particular node in a network (13). Therefore, centrality measures serve as a basis to study the relationship between lethality and essentiality by evaluating the topological features of the network.

In the context of PPI networks, CA ranks the nodes (proteins) in terms of their importance to the network structure and function, in an effort to identify key elements in the network (15, 16). Several different centrality metrics have been developed to capture different aspects of node importance; the mainstream measures include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality (12-14). Betweenness centrality and closeness centrality are two centrality algorithms that are used to measure the importance of nodes in a network. Both centrality algorithms are commonly used in network analysis to identify key nodes in a network (**Figure 1**).

Closeness centrality measures the average distance between a node and all other nodes in the network. It calculates the sum of the shortest paths between a specific node and all other nodes in the network (12, 14). Nodes with a high closeness centrality score are considered to be well-connected to other nodes in the network and are able to access information quickly. In the example of the *TP53*, the gene has high closeness centrality in the protein-protein interaction network that regulates cell growth and division. This means that *TP53* is able to quickly communicate with other proteins in the network. This is important because *TP53* needs to be able to respond quickly to DNA damage and other cellular stresses. If *TP53* is deleted from a cell, the other proteins in the network will be unable to communicate with each other as quickly. This can disrupt cellular processes and lead to cancer development (3).

Similarly, betweenness centrality measures the number of times a node acts as a bridge, or a "broker", between other nodes in the network. It calculates the number of shortest paths between all pairs of nodes that pass through a specific node. Nodes with a high betweenness centrality score are considered to be important in the network as they control the flow of information and have a high degree of control over the network (13). Degree centrality measures node importance using the number of direct connections it has in a network. Eigenvector centrality, however, introduces a greater level of complexity by considering both the connectivity of a given node and the influence, or centrality value, of the nodes to which it is connected. Nodes with high degree and eigenvector centralities are recognized as significant to the network due to their extensive, influential connections.

The selection of appropriate centrality measures is vital for determining key functional features of a biological network. In several studies, influential nodes in biological networks have been identified using classic centrality measures like degree, closeness, and betweenness centralities. Jeong, et. al. conducted pioneering work that found nodes with high degree centrality (hubs) among yeast PPI networks as likely to be associated with essential proteins (17). A node's degree centrality alone is insufficient to assess if it is crucial to the network as a whole and that many CA criteria should be taken into account while conducting network analysis.
Furthermore, due to the existence of dense clusters of interactions (modules) among a set of proteins in a network, rather than a random distribution of interactions, Joy et al. concluded that betweenness centrality is more likely to be important than degree centrality (18). On the other hand, a study by He and Zhang showed that the relationship between hub nodes and essentiality is not related to the network architecture (19).

Due to the varied outcome of different centrality studies of PPI networks, one can conclude that the most appropriate set of centrality measures for a particular network depends on the research question and the type of relationships being studied. The selection of an appropriate set of centrality measures to use in a study is complex because of the correlation between the centrality measures, the network data collection methods apart from the influence of network topology on the outcome. Different centrality measures yield different results based on the structure of the network and which centrality measure is appropriate for a particular network is contingent upon the purpose of the study and the type of connections being analyzed. Under the hypothesis that some algorithms are better predictors for detecting driver genes, we created a framework to analyze and select top centrality algorithms for the specific function of classifying genes as driver and non-drivers in CRC.

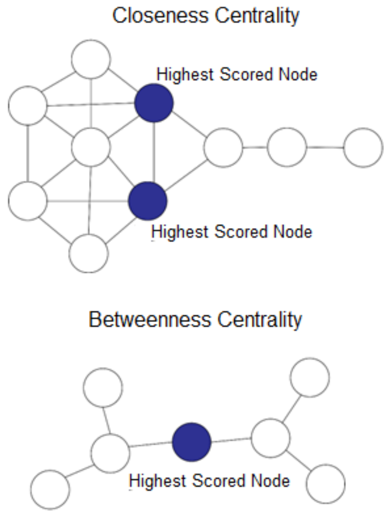


Figure 1: The two most significant Centrality measures identified in the network analysis of CRC. The circles indicate nodes, the lines indicate a functional relation between two nodes, and the node highlighted in blue has the highest centrality value in the network, indicating greater significance.

The task of classification is a popular application for machine learning (ML) methods (20). “Random Forest” (RF) is a popular ML algorithm that can be used for classification. It is an ensemble method that combines multiple decision trees to make predictions. In an RF, many decision trees are built on random samples of the data, and their predictions are combined to make a final prediction. The idea behind this is that, if each decision tree is trained on a different sample of the data, the overall prediction will be more robust and less prone to overfitting. This approach often results in more accurate predictions compared to using a single decision tree. RF also inputs a feature importance score to each node according to the Gini impurity index, making it a great tool for analyzing feature importance. The Gini impurity index measures the probability of misclassifying an observation. It’s used in decision tree algorithms to quantify a dataset’s impurity level or disorder.

In this work, we recreated a PPI network in Cytoscape, by querying genes associated with CRC in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (21, 22). STRING is a database that provides information about protein-protein interactions, including direct (physical) and indirect (functional) associations, for a large number of organisms, including humans (**Figure 2**). We pulled the top 50 most mutated genes in colorectal cancer from cBioPortal, a web-based interface for genetic analysis of cancer (23). We used information from Intogen to further classify these genes in driver (n=32) versus non-driver genes (n=18) and matched them to the ones we possessed in our original Cytoscape’s network (24).

We applied eight different centrality algorithms to score each protein; we further analyzed each algorithm according to its ability to discern between driver genes and non-driver genes. We first performed a non-parametric statistical analysis to rank our algorithms according to statistical significance. Later, we ran an RF classifier and extracted the most important features to predict driver genes. We found in both analyses that betweenness and closeness algorithms are the most important when trying to detect driver genes in CRC.

RESULTS

Of the original list of 50 genes extracted from cBioPortal, only 32 appeared in StringApp’s most important 1000-proteins network. We believe the lack of match is due to established findings that most mutated genes in cancer are of biological and clinical unimportance, but we were able to show this as most of the lacking genes are non-driver genes (25).

We extracted several properties from the nodes, including the already-computed centrality scores and standard functional annotations from StringApp such cell localization and enzyme properties. Descriptive statistics showed that driver genes have a higher centrality score for all algorithms and a higher standard deviation than the median, which showed a high dispersion of the data (**Table 1**).

The values of centrality were not normally distributed and therefore we used the non-parametric statistical test, the Mann-Whitney U test, to perform statistical analysis for the different centrality scores in driver and non-driver genes. We corrected for multiple-hypothesis testing bias, using a Bonferroni correction of n=8; only betweenness and closeness were statistically significant (**Table 1**).

This limited test suggests that there is a statistical significance between driver and non-driver genes, even with reduced power. We found that the two features with the lowest p-value were betweenness and closeness algorithms. As such, we hypothesized that these algorithms are the most important when classifying genes into driver versus non-drivers. We further validated our initial results using an ML classifier to predict driver and passenger genes. To do so, we built two different RF models. The first RF was trained on functional features like cell localization, expression in different tissue, and disease score derived from their presence together in scientific literature. The second RF was trained only on centralities scores to emphasize the predicting power of network-based properties. Our first model based on functional features was used as a benchmark, on the premise that already well-known features and properties of genes are a good starting point as demonstrated by previous studies (26). We bootstrapped the results by generating 1000 different iterations of train/test split sets and re-training the model at each iteration.

We used both F1 score and Receiver Operating Characteristic, Area Under the Curve (ROC AUC) to evaluate our models. F1 score, ROC and AUC are commonly used metrics to evaluate the performance of a classifier such as RF in a binary classification problem. The F1 score combines precision and recall to provide a single metric that reflects the balance between false positives and false negatives. A higher F1 score indicates better performance. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings, while the AUC is the area under the ROC curve. The AUC provides a single number that summarizes the performance of a classifier across all threshold settings. A higher AUC indicates better performance.

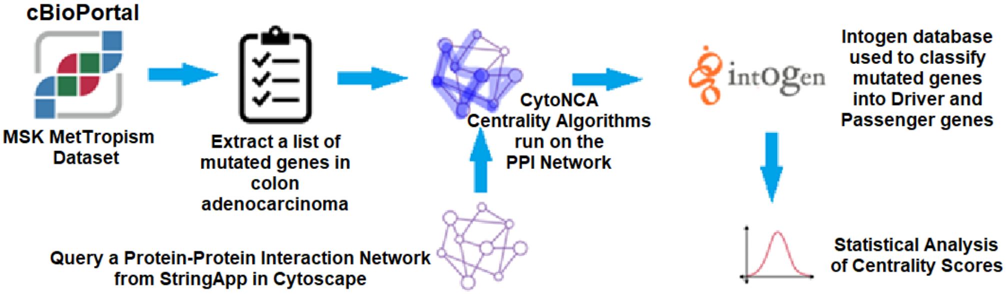

cBioPortal
MSK MetTropism Dataset
Extract a list of mutated genes in colon adenocarcinoma
CytoNCA Centrality Algorithms run on the PPI Network
intOgen
Intogen database used to classify mutated genes into Driver and Passenger genes
Query a Protein-Protein Interaction Network from StringApp in Cytoscape
Statistical Analysis of Centrality Scores

Figure 2: Schematic presentation of workflow to study Centrality measures in CRC. The diagram illustrates inflows of data from the MSK MetTropism Dataset, StringApp, and Intogen, summarizing the usage of this data to derive statistical results.

https://doi.org/10.59720/23-025

Centrality Algorithms	CRC Driver Gene	CRC Non- Driver Gene	P-Value
Degree (Weight)	65.221 ± 64.22	28.531 ± 33.46	0.010663
Eigenvector (Weight)	0.042 ± 0.04	0.016 ± 0.02	0.008682
Information (Weight)	5.33 ± 0.42	4.88 ± 0.54	0.007819
LACW (Weight)	19.61 ± 7.72	13.43 ± 5.96	0.032757
Network (Weight)	33.72 ± 46.41	17.44 ± 18.44	0.020997
Subgragh (Weight)	1.45e24 ± 5.01e24	2.22e23 ± 1.88e24	0.008682
Betweenness (Weight)	1858 ± 15615	169 ± 1480	0.000483
Closeness (Weight)	0.00811 ± 1.7e-5	0.00810 ± 1.8e-5	0.004539

Table 1: Centrality features weights for driver and passenger genes on a reconstructed CRC network using CytoNCA. Median ± standard deviation. P-value measured using with Mann-Whitney U Test (** statistically significant).

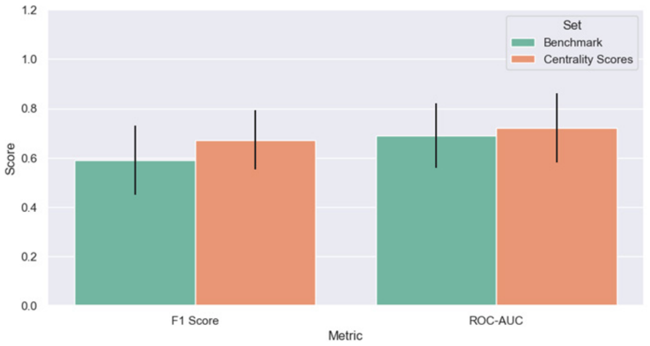


Figure 3. Mean F1 and ROC AUC scores out of 1000 iterations. Error bars represent standard deviation.

On both metrics, the centrality scores model outperformed the model trained on regular parameters (**Figure 3**). For the model trained on functional features, the results were 69.32 (CI 95% 45.12-90.32) AUC and F1 score of 0.59 (CI 95% 0.44-0.91). For the centrality scores model the results were 72.17 (CI 95% 47.11-89.54) AUC and F1 score of 0.67 (CI 95% 0.53-0.92).

The model trained on centrality scores, with an F1 score of 0.67, was better able to identify both positive and negative examples than the model trained on regular parameters given its F1 score of 0.59. Furthermore, the model trained on centrality scores, with an AUC of 72.17, was better able to rank positive examples above negative examples than the model trained on regular parameters, as its AUC was 69.32. The confidence intervals for the F1 score and AUC metrics both overlap, which means that there is no statistically significant difference in performance between the two models. These results suggest that using centrality scores as features for a random forest classifier can improve its performance in a binary classification problem.

We used our ML model to further enhance our understanding of centrality scores and their predictions capabilities. We extracted the feature importances from our RF models at each iteration and computed the mean across the 1,000 iterations. The feature importances are computed using the Gini impurity across each column. Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

At each iteration we extracted the feature importances score for each centrality algorithm and we computed the mean. We further ranked the centrality algorithms according to how important they are for the RF classifier. The two features that were previously marked as statistically significant, betweenness and closeness, were also ranked top in our importance of ML analysis, further validating our previous results (**Table 2**).

DISCUSSION

In this study we have shown that centrality algorithms, specifically betweenness centrality and closeness centrality, can be effectively used to predict CRC driver genes. Out of the original eight centrality algorithms applied, only closeness centrality and betweenness centrality reached statistical significance after a Bonferroni correction. The Bonferroni correction adjusts probability values because of the increased risk of a type I error when making multiple statistical tests. The method which multiplies the raw probability values by the number of tests.

We further used ML to improve our understanding of these algorithms as features to identify cancer driver genes. We implemented an RF classifier to classify driver genes from non-driver genes, and we further extracted the feature importance vector. We found that the key predictors for our ML models are betweenness and closeness, further validating our initial results.

Next we compared our RF classifier trained on a matrix of centrality scores, against a RF trained on a matrix of standard features like cell location, differential expression in tissue, and disease score. We demonstrated that the RF model that used centrality scores was better than the one with standard features. By incorporating network topological information from a PPI network, the proposed centrality-based approach was able to identify putative driver genes with higher accuracy. Furthermore, the results of this study demonstrate that integrating network topological information can enhance the performance of cancer driver gene prediction methods. The proposed centrality-based approach can be applied to other types of cancers, to uncover novel cancer driver genes.

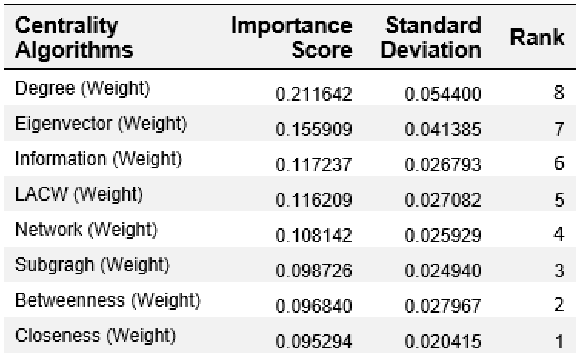
However, it's important to note that the proposed approach is not a substitute for experimental validation as limitations still exist in our understanding of the underlying biology and the challenge of incorporating complex relationships and interactions into the network representation. Further studies are needed to confirm the identified genes as drivers, as knowledge bias in our PPI networks could have limited the external validity of our work. These results can instead provide additional validation to experimental studies with the goal of improving early detection of colorectal cancer and opening up doors for novel diagnostic tools and treatment plans, thus further improving patient outcomes like it has been studied in hematopoietic cancers (27). In summary, this research offers new insights on how network topology can be utilized in the identification of CRC driver genes and opens the door for future developments in this field.

MATERIALS AND METHODS

Accessing Genomic Data

Genomic data was collected from the MSK MetTropism dataset in the cBioPortal for the most commonly implicated genes in colorectal cancer. cBioPortal is a cancer genomic

https://doi.org/10.59720/23-025



Centrality Algorithms	Importance Score	Standard Deviation	Rank
Degree (Weight)	0.211642	0.054400	8
Eigenvector (Weight)	0.155909	0.041385	7
Information (Weight)	0.117237	0.026793	6
LACW (Weight)	0.116209	0.027082	5
Network (Weight)	0.108142	0.025929	4
Subgragh (Weight)	0.098726	0.024940	3
Betweenness (Weight)	0.096840	0.027967	2
Closeness (Weight)	0.095294	0.020415	1

Table 2: Feature Importance Ranking by Random Forest Classifier. A lower importance score and lower standard deviation indicates higher accuracy and therefore a higher rank.

dataset analysis tool. MSK MetTropism is a PanCancer study that collects clinical and genomic data from 25,000 patients with more than 25 different types of cancer with the goal of discovering connections between genetic mutations and patterns of metastasis. The 50 genes with the most frequent nonsynonymous mutations in CRC were selected.

Identification of CRC driver genes based on mutation profiles

The selected genes were assigned the label of driver or non-driver genes using Intogen. Intogen (Integrative Oncogenomics) is a database that collects data from cancer genomes in order to identify cancer driver genes.

One network containing 1000 proteins was chosen. The size was selected in keeping with the complexity of cancer biology, while ensuring it remained computationally feasible to analyze. Many of the genes involved in colorectal cancer development and progression, encompassing both well-studied and less well-studied genes, are likely to be included in this type of network.

The network was imported to Cytoscape, through the plugin StringApp. We used the query term “colorectal cancer” to include only proteins already associated with this disease. A tool for the analysis and visualization of biological networks is provided by Cytoscape.. In our case, PPI networks were used, where gene proteins are represented by nodes, and signaling pathways are represented by edges. At least one other node is connected to each node in the network, and the weight of the edge (ranging from 0 to 1) represents the strength of this relationship.

Applying CA for Independent Identification

Betweenness, closeness, degree, eigenvector, information, LACW, network, and subgraph centralities were calculated through the Cytoscape Network’s CytoNCA app, which uses algorithms to calculate centralities of both weighted and unweighted networks (28). The weighted algorithms were used to model the strength of the association between proteins.

Confirming the Significance of CA Measures

A non-parametric T test (Mann Whitney U test) was used to compute the p values, and the alpha threshold was adjusted to 0.00625 following Bonferroni Correction (n=8) for multiple hypothesis testing (29).

Predicting driver genes with ML

Driver genes were predicted using both functional features (cell localization, expression in different tissues, and disease score from STRING), and centrality scores. An RF classifier was constructed using Python’s SkLearn library, and random training-test splits of 70-30 were created to train and predict these driver genes. Due to the small dataset size, this method was repeated 1,000 times in a bootstrapping fashion to reduce variance. The most important features for the ML models were extracted for further analysis of the most important centrality scores.

Received: February 6, 2023
Accepted: August 30, 2023
Published: November 18, 2023

REFERENCES

1. Li, Jiexi, et al. “Genetic and Biological Hallmarks of Colorectal Cancer.” *Genes & Development*, no. 11–12, June 2021, pp. 787–820. https://doi.org/10.1101/gad.348226.120.
2. Porta-Pardo, Eduard, et al. “Understanding Oncogenicity of Cancer Driver Genes and Mutations in the Cancer Genomics Era.” *FEBS Letters*, no. 24, Apr. 2020, pp. 4233–46. https://doi.org/10.1002/1873-3468.13781.
3. Vogelstein B, Kinzler KW. “The genetic basis of colorectal tumors.” *N Engl J Med*. 1993;328(6):1500-1508. https://doi.org/10.1056/NEJM199303183280607.
4. Hu, Fangjie, et al. “Comprehensive Analysis of Subtype-Specific Molecular Characteristics of Colon Cancer: Specific Genes, Driver Genes, Signaling Pathways, and Immunotherapy Responses.” *Frontiers in Cell and Developmental Biology*, Nov. 2021. https://doi.org/10.3389/fcell.2021.758776.
5. Guinney, Justin et al. “The consensus molecular subtypes of colorectal cancer.” *Nat Med.* 2015 Nov;21(11):1350-6. https://doi.org/10.1038/nm.3967.
6. Althubaiti, Sara, et al. “Ontology-Based Prediction of Cancer Driver Genes.” *Scientific Reports*, no. 1, Nov. 2019. https://doi.org/10.1038/s41598-019-53454-1.
7. Hu, Fuyan, et al. “Network-Based Identification of Biomarkers for Colon Adenocarcinoma.” *BMC Cancer*, no. 1, July 2020. https://doi.org/10.1186/s12885-020-07157-w.
8. Liu, Xiaoqun, et al. “Identification of Crucial Genes and Pathways Associated with Colorectal Cancer by Bioinformatics Analysis.*” Oncology Letters*, Jan. 2020. https://doi.org/10.3892/ol.2020.11278.
9. Mansouri, Vahid, et al. “Gene Screening of Colorectal Cancers via Network Analysis.” Gastroenterology and Hepatology from Bed to Bench Vol. 12,2 (2019): 149-154.
10. Nikiforova, Victoria J., and Lothar Willmitzer. “Network Visualization and Network Analysis.” *Experientia Supplementum*, Birkhäuser Basel, pp. 245–75. https://doi.org/10.1007/978-3-7643-7439-6_11.
11. Sharma, Pooja, et al. “Centrality Analysis in PPI Networks.” 2016 International Conference on Accessibility to Digital World (ICADW), IEEE, Dec. 2016. https://doi.org/10.1109/icadw.2016.7942528.
12. Hahn, Matthew W., and Andrew D. Kern. “Comparative

Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks." *Molecular Biology and Evolution*, no. 4, Dec. 2004, pp. 803–06. https://doi.org/10.1093/molbev/msi072.

13. Koschützki, Dirk, and Falk Schreiber. "Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks." *Gene Regulation and Systems Biology*, Jan. 2008, p. GRSB.S702. https://doi.org/10.4137/grsb.s702.
14. Ghasemi, Mahdieh, et al. "Centrality Measures in Biological Networks." *Current Bioinformatics*, no. 4, Aug. 2014, pp. 426–41. https://doi.org/10.2174/15748936113086660013.
15. Barabási, Albert-László, and Zoltán N. Oltvai. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics*, no. 2, Feb. 2004, pp. 101–13. https://doi.org/10.1038/nrg1272.
16. Fell, David A., and Andreas Wagner. "The Small World of Metabolism." *Nature Biotechnology*, no. 11, Nov. 2000, pp. 1121–22. https://doi.org/10.1038/81025.
17. Jeong, H., et al. "Lethality and Centrality in Protein Networks." *Nature*, no. 6833, May 2001, pp. 41–42. https://doi.org/10.1038/35075138.
18. Joy, Maliackal Poulo, et al. "High-Betweenness Proteins in the Yeast Protein Interaction Network." *Journal of Biomedicine and Biotechnology*, no. 2, 2005, pp. 96–103. https://doi.org/10.1155/jbb.2005.96.
19. He, Xionglei, and Zhang, Jianzhi. "Why Do Hubs Tend to Be Essential in Protein Networks?" *PLoS Genetics*, no. 6, June 2006, p. e88. https://doi.org/10.1371/journal.pgen.0020088.
20. Andrades, Renan, and Mariana Recamonde-Mendoza. "Machine Learning Methods for Prediction of Cancer Driver Genes: A Survey Paper." *Briefings in Bioinformatics*, no. 3, Mar. 2022. https://doi.org/10.1093/bib/bbac062.
21. Doncheva, Nadezhda T., et al. "Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data." *Journal of Proteome Research*, no. 2, Nov. 2018, pp. 623–32. https://doi.org/10.1021/acs.jproteome.8b00702.
22. Shannon, Paul, et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research*, no. 11, Nov. 2003, pp. 2498–504. https://doi.org/10.1101/gr.1239303.
23. Gao, Jianjiong, et al. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the CBioPortal." Science Signaling, no. 269, American Association for the Advancement of Science (AAAS), Apr. 2013. https://doi.org/10.1126/scisignal.2004088.
24. Gonzalez-Perez, Abel et al. "IntOGen-mutations identifies cancer drivers across tumor types." *Nature Methods* vol. 10,11 (2013): 1081-2. https://doi.org/10.1038/nmeth.2642,
25. Vogelstein, Bert et al. "Cancer genome landscapes." *Science* vol. 339,6127 (2013): 1546-58. https://doi.org/10.1126/science.1235122.
26. Luo, Ping et al. "deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks." *Frontiers in Genetics* vol. 10 13. 29 Jan. 2019, https://doi.org/10.3389/fgene.2019.00013.
27. Kar, Siddhartha P et al. "Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis." *Nature Genetics* vol. 54,8 (2022): 1155-1166. https://doi.org/10.1038/s41588-022-01121-z.
28. Tang, Yu, et al. "CytoNCA: A Cytoscape Plugin for Centrality Analysis and Evaluation of Protein Interaction Networks." *Biosystems*, Jan. 2015, pp. 67–72. https://doi.org/10.1016/j.biosystems.2014.11.005.
29. Nachar, Nadim. "The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution." Tutorials in Quantitative Methods for Psychology, no. 1, The Quantitative Methods for Psychology, Mar. 2008, pp. 13–20. https://doi.org/10.20982/tqmp.04.1.p013.