


ORIGINAL RESEARCH

Identifying driver modules based on multi-omics biological networks in prostate cancer

Zhongli Chen^{1,2,3} | Biting Liang³ | Yingfu Wu³ | Haoru Zhou³ | Yuchen Wang² | Hao Wu² 

¹Tibet Center for Disease Control and Prevention, Lhasa, China

²School of Software, Shandong University, Jinan, China

³School of Information Engineering, Northwest A&F University, Yangling, China

Correspondence

Hao Wu, School of Software, Shandong University, Jinan, Shandong, 250100, China.
Email: haowu@sdu.edu.cn

Funding information

National Key Research and Development Program, Grant/Award Number: 2021YFF0704103; Fundamental Research Funds of Shandong University; National Natural Science Foundation of China, Grant/Award Number: 61972322; Natural Science Foundation of Shaanxi Province, Grant/Award Number: 2021JM110

Abstract

The development of sequencing technology has promoted the expansion of cancer genome data. It is necessary to identify the pathogenesis of cancer at the molecular level and explore reliable treatment methods and precise drug targets in cancer by identifying carcinogenic functional modules in massive multi-omics data. However, there are still limitations to identifying carcinogenic driver modules by utilising genetic characteristics simply. Therefore, this study proposes a computational method, NetAP, to identify driver modules in prostate cancer. Firstly, high mutual exclusivity, high coverage, and high topological similarity between genes are integrated to construct a weight function, which calculates the weight of gene pairs in a biological network. Secondly, the random walk method is utilised to reevaluate the strength of interaction among genes. Finally, the optimal driver modules are identified by utilising the affinity propagation algorithm. According to the results, the authors' method identifies more validated driver genes and driver modules compared with the other previous methods. Thus, the proposed NetAP method can identify carcinogenic driver modules effectively and reliably, and the experimental results provide a powerful basis for cancer diagnosis, treatment and drug targets.

1 | INTRODUCTION

Cancer genomics research can reveal much unknown information about cancer [1–4]. It can not only explore the pathogenesis of cancer in-depth but also provide more drug targets for the clinical treatment of cancer by reasoning about the interaction between genes [5–7]. It also lays the foundation for the development of precision medicine [8, 9]. The in-depth development of cancer genomics research has enriched genomic data, thus enabling researchers to design efficient and effective computational methods to study cancer-related genomics data systematically [10–13].

Vandin et al. [14] proposed the Dentrix algorithm to identify oncogenic driver pathways by designing the greedy algorithm based on gene mutation data. However, the method can only

obtain local optimal solutions, and the number of genes in a driver pathway must be specified in advance. Therefore, the application of the Dentrix algorithm is limited to a certain extent. To solve the problems of the Dentrix algorithm, Zhao et al. proposed the MDPFinder algorithm [10] to identify oncogenic driver pathways in which a linear programming algorithm was applied to solve the maximum weight submatrix problem based on somatic mutation and gene expression data. Subsequently, Leisersen et al. proposed the Multi-Dentrix algorithm [11] in order to identify multiple driver pathways simultaneously that satisfy high mutual exclusivity and high coverage. However, some parameters of the algorithm need to be preset. Therefore, the algorithm lacks generality.

As previously identified, protein plays an important role in cell structure and cell cycle [15–18]. Disruption of cell structure

Abbreviations: CNV, Copy number variation; FN, False Negative; FP, False Positive; HPRD, Human protein reference database; JS, Jensen–Shannon; KL, Kullback–Leibler; NCG, A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens; PRAD, Prostate adenocarcinoma; PPI, Protein–protein interaction; TCGA, The cancer genome Atlas; TN, True Negative; TP, True Positive.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *IET Systems Biology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

and cell cycle results in uncontrolled cell proliferation. Therefore, normal cells may transform into cancer cells [19, 20]. Given that abundant protein–protein interaction (PPI) data have been augmented by proteomics experiments, it is better to identify oncogenic driver modules based on PPI networks [21–24].

MCODE is a typical clustering algorithm to identify cancer functional modules based on PPI networks [21]. Firstly, the algorithm calculates the density value of the node to get the weight of the node. Secondly, a cluster is expanded centred on the node with the highest weight value. Finally, the algorithm filters the non-density subgraphs to identify optimal functional modules in cancer. However, MCODE cannot efficiently calculate the interaction strength between proteins. The algorithm identifies passenger proteins associated with high densities of proteins in the network, which lead to a decrease in the accuracy of the method. Instead, ClusterONE [22] calculates the strength of protein–protein interactions by constructing a weighted network, and the greedy strategy is applied to the algorithm to partition a set of proteins with high cohesion into functional modules. However, the ClusterONE algorithm can be biased due to overreliance on cohesive formulations. For example, the addition of a node may lead to a decrease in the cohesion of a candidate functional module; however, the node should be classified as a member of a functional module.

It is difficult to accurately identify functional modules using only proteomic data or genomics data. In this study, a computational method, NetAP, is proposed to identify oncogenic driver modules based on a multi-omics biological network. Firstly, a biological network is constructed by integrating gene mutation, copy number variation (CNV), and PPI network data. The interaction weights among genes in the biological network are calculated using the weight function, which integrates the similarity of topology among gene nodes and the two characteristics of high coverage and mutual exclusivity. Secondly, the random walk method is used to reevaluate the strength of the interaction among genes to avoid adding unnecessary passenger genes to the driver modules. Finally, the affinity propagation algorithm is applied to obtain the optimal driver modules. Furthermore, we analyse the contribution of the similarity index and random walk method to measure the performance of the NetAP method. The accuracy, effectiveness, and ability of the NetAP method to identify validated driver genes are compared with other previous methods in prostate cancer. Meanwhile, the evidence that the optimal driver modules identified by NetAP play an important role in prostate cancer and rational speculation that facilitates the study of cancer pathogenesis and drug targets is elucidated.

2 | METHODS

2.1 | Mutual exclusivity and coverage

Previous studies have shown that the driver modules have two key characteristics, namely high coverage and high mutual exclusivity [14], which have been widely used in carcinogenic driver module identification [25–28]. The definitions of coverage and mutual exclusivity are described in this section.

In this study, gene mutation and CNV data are fused on the basis of the PPI network data, which are stored as a binary matrix $A_{m \times n}$ with m rows and n columns, where m rows represent m patient samples and n columns represent n genes. If there is a mutation or copy number variation in the j -th gene of the i -th sample, $A_{ij} = 1$, otherwise, $A_{ij} = 0$. In this paper, $G = (V, E)$ represents the PPI network. Each vertex $u_i \in V$ in G represents a protein, and each undirected edge $(u_i, u_j) \in E$ corresponds to the interaction among proteins. If the genes in $A_{m \times n}$ correspond to nodes u_i in the PPI network, the genes are mapped to the PPI network. Let S_i denote the set of samples in which gene g_i is mutated. V represents the set of mutated genes. $M \subseteq V$ is a subset of mutated genes. For any pair of genes, $g_i, g_j \in M, g_i \neq g_j$, if $S_i \cap S_j = \emptyset$, the genes in M are mutually exclusive.

The mutual exclusivity of gene subset M is represented as follows:

$$ED(M) = \frac{|\bigcup_{g_i \in M} S_i|}{\sum_{g_i \in M} |S_i|} \quad (1)$$

The coverage of gene subset M is represented as follows:

$$CD(M) = \frac{|\bigcup_{g_i \in M} S_i|}{|\bigcup_{g_i \in V} S_i|} \quad (2)$$

2.2 | Similarity definition

The local area network is composed of a gene node and their neighbour nodes in the biological network, and its abnormality may have an impact on biological functions. Therefore, the influence of gene nodes on biological functions can be analysed by evaluating the similarity between adjacent nodes in the local area network. In this paper, the gene node similarity metrics are constructed utilising the Jensen–Shannon (JS) divergence [29, 30], where the JS divergence is calculated based on the discrete probability set [28, 31].

Firstly, a local area network with the gene node as the centre and the one-step range of the centre node as the radius is constructed. Next, the discrete probability of each gene node in the local area network is calculated to construct a discrete probability set. In this step, let d_i be the degree of the i -th gene node and d_{\max} be the maximum node degree. Assume that there are N elements in the probability set of each node, $N = d_{\max} + 1$.

In the local network of gene node g_i , D_{g_i} is the sum of node degrees of g_i , and the calculation formula is as follows:

$$D_{g_i} = \sum_{j=1}^n d_j \quad (3)$$

where n represents the number of genes and d_j represents the degree of the j -th gene node.

The discrete probability of gene node g_i is calculated as follows:

$$p(i) = \frac{d_i}{D_{g_i}} \quad (4)$$

The discrete probability set $p(i)$ consists of the standardised discrete probabilities of all gene nodes in the local area network sorted from small to large, which is expressed as follows:

$$P(i) = (p(1), p(2), \dots, p(n), \dots, p(N)) \quad (5)$$

where $p(n)$ represents the discrete probability value of the n -th gene node ($n \leq N$).

Secondly, the Kullback–Leibler (KL) divergence between gene nodes is calculated using the constructed discrete probability set, which is a measure of the asymmetry of the difference between the probability distributions $p(x)$ and $Q(x)$. In this paper, for two adjacent nodes g_i and g_j , $p(x)$ and $Q(x)$ correspond to different probability sets $P(i)$ and $P(j)$ with the same number of elements, respectively. The KL divergence is calculated as follows:

$$D_{KL}(P(i)||P(j)) = \sum_{k=1}^N P_i(k) \log \frac{P_i(k)}{P_j(k)} \quad (6)$$

Thirdly, the JS divergence value between nodes g_i and g_j is calculated based on the KL divergence value. The JS divergence solves the asymmetry of the KL divergence, which is calculated as follows:

$$D_{JS}(P(i)||P(j)) = \frac{1}{2} D_{KL} \left(\left(P(i) \right) \left\| \frac{P(i) + P(j)}{2} \right. \right) + \frac{1}{2} D_{KL} \left(\left(P(j) \right) \left\| \frac{P(i) + P(j)}{2} \right. \right) \quad (7)$$

Finally, the similarity metric $SIM(g_i, g_j)$ is constructed to analyse the topological similarity between gene pair g_i and g_j , which is calculated as follows:

$$SIM(g_i, g_j) = 1 - D_{JS}(P(i)||P(j)) \quad (8)$$

The value range of $SIM(g_i, g_j)$ is (0,1), and $SIM(g_i, g_j) = 1$ means that two gene nodes in the network have the same topological structure.

In particular, we set a threshold θ to strengthen the influence of similarity on edge weights. If $SIM(g_i, g_j) < \theta$, the weight value of the edge is assigned as 0. We discuss the value of the threshold in the parameter setting section of this study.

2.3 | Carcinogenic driver module identification

2.3.1 | Edge weighted network construction

$G = (V, E)$ represents the PPI network, where $V = \{g_1, g_2, g_3, \dots, g_n\}$ represents the set of abnormal genes

corresponding to the vertices in the PPI network, and $E = \{e = (g_i, g_j)\}$ represents the set of protein–protein interaction relationships.

Create a weighted undirected network graph G_w . Set the weight $\omega(g_i) = CD(\{g_i\})$, for each vertex $g_i \in V$, and the larger the coverage value of the gene, the larger the weight value of the vertex.

Taking into account the increased chance of the coexistence of a gene and its surrounding genes, the set of gene node g_i and its surrounding gene nodes is taken as the local area network $Ne(g_i)$ in this study, which is expressed as follows:

$$Ne(g_i) = \{g_i\} \cup \left\{ \cup_{v(g_i, g_j) \in E} g_j \right\} \quad (9)$$

where the surrounding genes refer to the genes within one step of a core gene.

In order to balance the mutual exclusivity between genes, the mutual exclusivity of gene pairs $ED(g_i, g_j)$ is determined as the average of $ED(Ne(g_i))$ and $ED(Ne(g_j))$. The calculation formula is as follows:

$$ED(g_i, g_j) = \frac{ED(Ne(g_i)) + ED(Ne(g_j))}{2} \quad (10)$$

To reduce the chance that a single gene with larger coverage dominates the edge weights, the product of the two gene coverages determines the gene pair coverage $CD(g_i, g_j)$. The calculation formula is as follows:

$$CD(g_i, g_j) = CD(\{g_i\}) \times CD(\{g_j\}) \quad (11)$$

Based on the characteristics of high mutual exclusivity and high coverage and the assumption of high topological similarity within the same driver module, we have a trade-off between the characteristics of mutual exclusivity and coverage of genes within the same driver module. Therefore, the weights of gene pairs in the constructed network $\omega(g_i, g_j)$ are calculated by the product of similarity and the harmonic mean of mutual exclusivity and coverage. The calculation formula is as follows:

$$\omega(g_i, g_j) = \begin{cases} \frac{2 \times SIM(g_i, g_j)}{\frac{1}{ED(g_i, g_j)} + \frac{1}{CD(g_i, g_j)}}, & \begin{matrix} ED(g_i, g_j) \neq 0, \\ CD(g_i, g_j) \neq 0, \\ SIM(g_i, g_j) \geq \theta, \end{matrix} \\ 0, & otherwise \end{cases} \quad (12)$$

The pseudocode of edge weighted network construction is provided in Algorithm 1.

Algorithm 1 Edge weighted network construction

Input: $G(V, E), S_i, \theta$
Output: G_ω
Initialization: $j = |E|, i = 1$
 2: **for** i to j **do**
 3: compute $ED(g_i, g_j), CD(g_i, g_j), SIM(g_i, g_j)$
 4: **if** $ED(g_i, g_j) \neq 0 \cup CD(g_i, g_j) \neq 0 \cup SIM(g_i, g_j) \geq \theta$ **then**
 5: compute $\omega(g_i, g_j)$
 6: **else** $\omega(g_i, g_j) = 0$
 7: $genenetwork[g_i][g_j] = \omega(g_i, g_j)$
 8: $i = i + 1$

2.3.2 | Random walk process

The random walk method can further mine the relationship among nodes in networks [32, 33]. In biological networks, the strength of interactions among genes [34–36] can be reassessed by random walk methods. Once the weighted undirected network graph G_ω is constructed, the random walk method is executed on G_ω . The random walk can be described as a network propagation process. That is, at the beginning, the traverser walks from any vertex g_i on G_ω to the neighbour vertex g_j of g_i with the probability of α . Meanwhile, the traverser can also randomly

jump to any other vertex g_s in G_ω with probability $1 - \alpha$. The weight distribution of G_ω is updated after each random walk process. Let the weight distribution at time t be denoted by F_t . The updated weight distribution is used as input to the next random walk process, which iterates repeatedly until a steady state is reached. The iteration formula is as follows:

$$F_{t+1} = \left[\frac{1-\alpha}{n} \cdot I + \alpha F_0 \right] \cdot F_t \quad (13)$$

where F_0 is a matrix with initial weight values, that is, $F_0 = G_\omega$, I is the identity matrix, and n is the number of nodes in the network. α is the random walk probability, which is calculated as follows:

$$\alpha(g_i, g_j) = \begin{cases} \frac{\omega(g_i, g_j)}{\sum_k \omega(g_k, g_j)}, & (g_i, g_j) \in E \\ 0, & otherwise \end{cases} \quad (14)$$

The weight of each gene pair g_i and g_j ($i \neq j$) can be updated to F by using the random walk method. Finally, a weighted undirected graph G_d is constructed.

2.3.3 | Driver module set identification

Each gene in the constructed weighted undirected graph G_d has dual features. In other words, each gene can be regarded as a potential clustering centre of a cluster and can also be classified as a member of a cluster with another clustering centre. Responsibility and availability are described as the duality of genes in the clustering process [37]. That is, for each gene pair (g_i, g_k) , responsibility represents the degree that gene g_k is the clustering centre of gene g_i , and availability represents the degree that gene g_i supports gene g_k as the clustering centre of gene g_i . The calculation formulas are as follows, respectively:

$$R(g_i, g_k) = C(g_i, g_k) - \max_{g_{i'} \neq g_k} \{A(g_i, g_{i'}) + C(g_i, g_{i'})\} \quad (15)$$

$$A(g_i, g_k) = \begin{cases} \min\{0, R(g_k, g_k) + \sum_{g_{i'} \notin (g_i, g_k)} \max\{0, R(g_{i'}, g_k)\}\} & g_i \neq g_j \\ \sum_{g_{i'} \neq g_k} \max\{0, R(g_{i'}, g_k)\}, & g_i = g_j \end{cases} \quad (16)$$

According to Formula 15 and 16, the larger the value of $R(g_k, g_k) + A(g_k, g_k)$, the larger the possibility that gene g_k is the clustering centre of gene g_i and the larger the possibility that gene g_i is classified as a member of the clustering centre g_k .

In this study, the idea of affinity propagation clustering is used to identify driver modules in prostate cancer in the following three steps:

A. Input matrix construction. A matrix consisting of correlations among gene nodes is created as the input. The correlation value is determined by the negative Euclidean distance between the gene nodes g_i and g_k in G_d [37]. The calculation formula is as follows:

$$C(g_i, g_k) = -\|g_i - g_k\|^2 \quad (17)$$

According to Formula 17, the larger the value of $C(g_i, g_k)$, the smaller the distance between genes, and the stronger the correlation between g_i and g_k . In this case, g_i and g_k tend to be in the same driver module.

B. Initialisation. The responsibility matrix and availability matrix are created using Formulas 15 and 16, which contain the responsibility and availability information of genes in the iterative process, respectively. Given that the iterative process can only be started from zero matrices, the two matrices are initialised to zero. Shocks are easily generated in the next iterative process, thus the damping coefficient λ is set for the convergence effect of the control method [37]. The iteration formula is as follows:

$$R_{t+1}(g_i, g_k) = \lambda \times R_t(g_i, g_k) + (1 - \lambda) \times R_{t+1}(g_i, g_k) \quad (18)$$

$$A_{t+1}(g_i, g_k) = \lambda \times A_t(g_i, g_k) + (1 - \lambda) \times A_{t+1}(g_i, g_k) \quad (19)$$

The responsibility matrix and the availability matrix will be updated after each iteration. When the value of $R(g_i, g_k) + A(g_i, g_k)$ is the maximum, the clustering centre of gene g_i can be decided to be gene g_k . If $g_i = g_k$, then g_i is the clustering centre of the module to which it belongs. Otherwise, g_i is the member gene of the cluster where g_k is the clustering centre.

C. Formulas 18 and 19 are executed alternately. Gene nodes will be divided into corresponding clustering centres until the modules no longer change.

The pseudocode of edge weighted network construction is provided in Algorithm 2.

Algorithm 2 Driver module set identification

Input: $C(g_i, g_k)_{g_i, g_k \in \{1, 2, 3, \dots, N\}}$, λ , P
Output: clustering results C
1: *Initialisation:* $R(g_i, g_k) = 0$, $A(g_i, g_k) = 0$
2: **repeat**
3: $R_{t+1}(g_i, g_k) = \lambda \times R_t(g_i, g_k) + (1 - \lambda) \times R_{t+1}(g_i, g_k)$
4: $A_{t+1}(g_i, g_k) = \lambda \times A_t(g_i, g_k) + (1 - \lambda) \times A_{t+1}(g_i, g_k)$
5: **until** $R_{t+1}(g_i, g_k) = R_t(g_i, g_k)$ and $A_{t+1}(g_i, g_k) = A_t(g_i, g_k)$
6: $C_i = \text{argmax}_k [A_{t+1}(g_i, g_k) + R_{t+1}(g_i, g_k)]$
7: $C = (C_1, C_2, C_3, \dots, C_i, \dots, C_N)$

3 | DATA PREPROCESSING

This study integrates multiple omics data, including gene mutation data, CNV data of prostate adenocarcinoma (PRAD) from TCGA and human PPI network data from the human protein reference database (HPRD) [38].

3.1 | PRAD data preprocessing

Some passenger genes in the PRAD database are filtered to simplify the gene network. The specific screening operations are as follows: Delete the low-mutation genes that are not higher than the mutation in one sample from the genetic mutation data of 495 PRAD patient samples containing 40,543 genes. Delete the low-variable genes that are less than 15% of the samples from the CNV data of 493 PRAD patient samples containing 24,776 genes. Finally, 489 samples containing 507 candidate driver genes are obtained. The results of PRAD data preprocessing are shown in Table 1.

3.2 | The combination of genes and the HPRD network

It is difficult to accurately identify functional modules using only proteomic or genomic data. Therefore, candidate driver genes are mapped to the human PPI network from HPRD, which consists of 14,213 proteins and 173,231 interaction relationships. Finally, a gene network consisting of 178 proteins and 251 interaction relationships is obtained.

4 | RESULTS

4.1 | Evaluation metrics

The three evaluation metrics of Calinski–Harabaz Score, accuracy and F-measure are introduced to evaluate the results of the NetAP method. The Calinski–Harabaz Score is a reference for evaluating the quality of the clustering method [39]. The higher the Calinski–Harabaz Score, the better the clustering results. Accuracy and F-measure are the most widely used to evaluate the clustering effect. Generally, the higher the accuracy, the better the clustering method. F-measure balances the precision rate and recall rate to evaluate the effectiveness of the method scientifically. The higher the F-measure, the stronger the ability of the identified driver modules to be enriched into

TABLE 1 Results of prostate adenocarcinoma (PRAD) data preprocessing

Type of data	Before filtering		After filtering		Results	
	Number of samples	Number of genes	Number of samples	Number of genes	Number of samples	Number of genes
Gene mutation data	495	40,543	495	4618	489	507
Copy number variation data	493	24,776	493	6139		

known biological pathways. Accuracy and F-measure are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (23)$$

In this study, the validated driver gene sets in the driver modules are set as ground truth. The validated driver genes are obtained from the tumour driver gene database, NCG 7.0 [43]. The enrichment results of predicted driver modules and the ‘ground truth’ of driver modules are obtained by the functional enrichment analysis tool, DAVID. Driver modules with p -value < 0.05 are set as positive classes. True Negative (TN) indicates the number of modules in which negative classes are predicted to be negative classes; True Positive (TP) indicates the number of modules in which positive classes are predicted to be positive classes; False Negative (FN) indicates the number of modules in which positive classes are predicted to be negative classes; and False Positive (FP) indicates the number of modules in which negative classes are predicted to be positive classes.

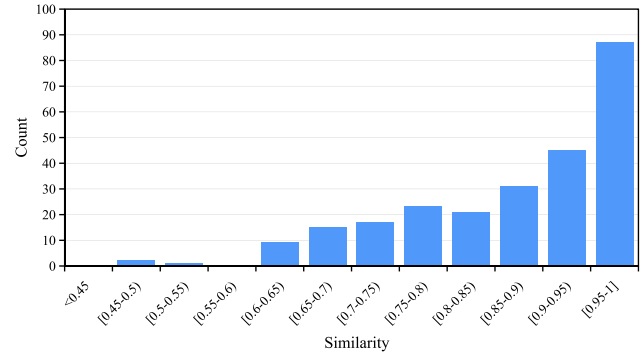
4.2 | Parameter settings

For the NetAP method, three important parameters of θ , p and λ need to be set. To determine the value range of the similarity threshold θ , we check the similarity values of all edges and plot the distribution of the similarity values of all edges shown in Figure 1a. It can be seen from the figure that the similarity value of each edge is greater than 0.45. Therefore, the values of θ are set to be $\{0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$.

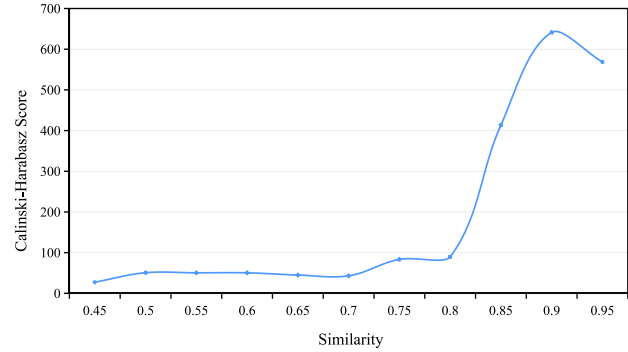
Figure 1b shows the Calinski–Harabaz score with different θ after running NetAP. We observe that the Calinski–Harabaz score is the maximum when $\theta = 0.9$. Therefore, $\theta = 0.9$.

The threshold p plays a crucial role in generating an appropriate number of clusters. In general, the median value of the correlation degree is selected as the value of p . To ensure that the value of p is set reasonably, the range of p is set to be $\{median/2, median, 2 \times median\}$. Additionally, the convergence effect of the NetAP method is controlled by λ , and the range of λ is $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ [37].

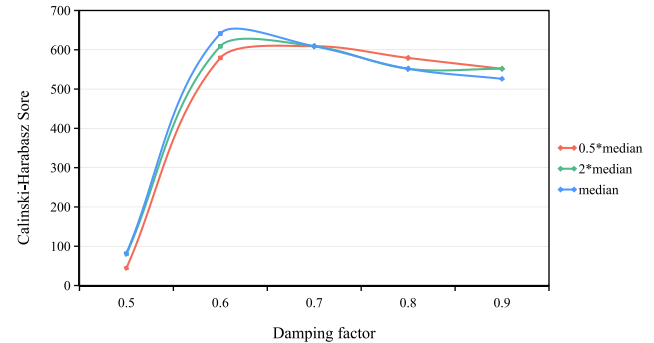
To obtain the optimal results of the NetAP method, the grid optimisation strategy is used to obtain the appropriate values of parameters p and λ . It can be seen from Figure 1c that the NetAP method obtains the maximum Calinski–Harabaz score with $p = median$ and $\lambda = 0.6$. It means that



(a) The distribution of the similarity values of all edges



(b) The Calinski-Harabaz score for NetAP with different θ



(c) The result of NetAP by using the grid optimization strategy

FIGURE 1 Parameter settings

NetAP has the best clustering effect in this case. Therefore, we set $p = median$ and $\lambda = 0.6$.

4.3 | Influence of the initial input gene set

To investigate the effect of different choices of the initial input gene set on module identification, we applied the NetAP method to gene mutation data, CNV data and combined data in prostate cancer. Subsequently, the results are evaluated using the Calinski–Harabaz score. It can be seen from Table 2 that the NetAP method achieves a better Calinski–Harabaz score using the combined data than using the single gene mutation data or the single CNV data. The results show that the module

TABLE 2 Results of the NetAP algorithm with different initial input data sets

	Gene mutation data	CNV data	Gene mutation U CNV data
Calinski–Harabaz score	1.33	230.12	641.31

recognition effect of the NetAP method is higher on combined data than on single data. We can conclude that the gene mutation data and CNV data complement each other and both contribute to the reconstruction of the biological network. Therefore, multi-omics data is more helpful to improve the effect of module identification than single data.

4.4 | Contribution of random walk and similarity

The method framework of the similarity metric (Similarity) and non-similarity metric (Non-Similarity) are introduced to illustrate the contribution of similarity to the NetAP method. In addition, the random walk (RW) method and the RW and restart random walk (RRW) methods are compared to illustrate the contribution of the random walk method to the NetAP method. The restart probability of the RRW method with $\beta \in (0,1)$, the three representative values of β are set as 0.1, 0.5 and 0.9, respectively. In this study, the method that uses the random walk is represented by RW, and the method that does not use the random walk is represented by NRW. According to different values of β , the methods that use the restart random walk are represented by RRW_ β 01, RRW_ β 05, and RRW_ β 09, respectively. The results using the above methods are shown in Figure 2.

The Calinski–Harabaz scores of the five methods are utilised to evaluate the contribution of random walk, and the results are shown in Figure 2. Within the framework of the similarity metric, the scores of the RW and NRW are 641.31 and 0, respectively. The score of the RW method is higher than that of the NRW method significantly. In addition, the score of the RW method is 91 times, 471.4 times, and 616.6 times higher than that of RRW_ β 01, RRW_ β 05, and RRW_ β 09, respectively. Similarly, the scores of the five methods are also compared within the framework of non-similarity. The RW method achieves the highest score of 97.40. The scores of the RRW_ β 05 and the RRW_ β 09 methods are 0, and the score of the RW method is 22.8 times and 62 times higher than that of the RRW_ β 01 method and NRW method, respectively. Therefore, the random walk method has a significant promotion on the performance within the two frameworks, and it is better than the restart random walk method.

The contribution of the similarity metric is shown in Figure 2. The scores of the five methods within the similarity metric framework and the non-similarity framework are compared. The scores of the RW method and RRW_ β 01 method increase by 6.6 times and 1.7 times, respectively, and the scores of the RRW_ β 05 and RRW_ β 09 methods increase from 0 to 1.36 and 1.04, respectively. Although the framework of the similarity metric does not perform well in the NRW method, the similarity metric improves the performance of the

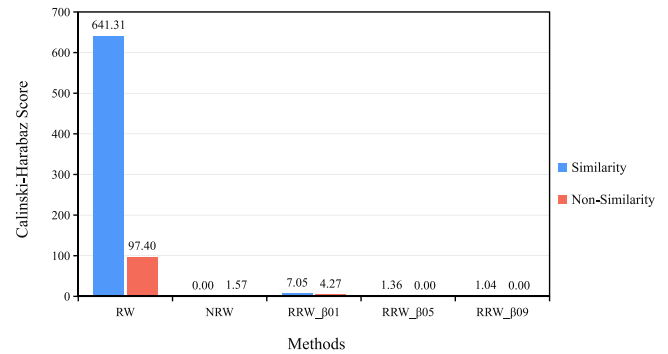


FIGURE 2 The results of the five methods

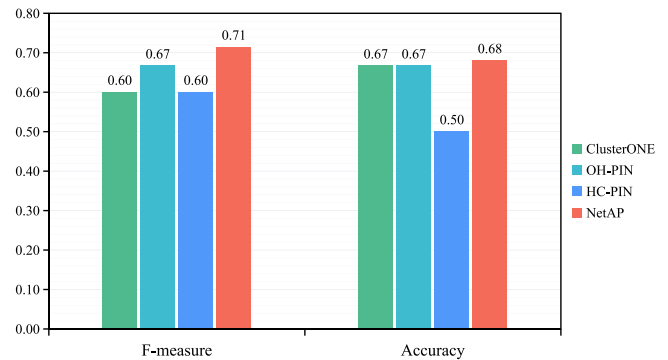


FIGURE 3 Accuracy and F-measure of the NetAP method

NetAP method to a great extent. In conclusion, ignoring either similarity or random walk may adversely affect the performance of the NetAP method. We can conclude that both the similarity metric and the random walk method are helpful to improve the performance of the NetAP method. It shows that it is reasonable to consider the similarity metric and random walk method in the NetAP method.

4.5 | Identification proficiency verification

We run the MCODE [21], ClusterONE [22], OH-PIN [40], HC-PIN [41] and NetAP methods in prostate cancer data from TCGA. MCODE identifies only one effective driver module. It means that the MCODE method does not have good data compatibility and effectiveness. Therefore, the results of ClusterONE, OH-PIN, HC-PIN and NetAP are compared in Figure 3. It can be seen that the accuracy of the NetAP method is 11%, 4% and 11% higher than that of ClusterONE, OH-PIN and HC-PIN, respectively. The F-measure of the NetAP method is 1%, 1% and 18% higher than that of the other three methods, respectively. The experimental results show that the NetAP method is more accurate and effective

than MCODE, ClusterONE, OH-PIN and HC-PIN. Therefore, the NetAP method shows a better ability to identify driver modules in prostate cancer compared with other previous competitive methods.

4.6 | Comparison of ability to identify validated driver genes

In order to evaluate the ability of different methods to identify driver genes and driver modules, we compare the gene sets identified by different algorithms with a list of validated cancer driver genes [42]. Firstly, the list of cancer driver genes is downloaded from the commonly used tumour driver gene database, NCG 7.0, including 3177 validated cancer driver genes [43]. Secondly, 55 validated prostate cancer driver genes are screened out from NCG as a benchmark to analyse cancer driver genes identified by the five algorithms [42]. The comparison of the identification results with the validated driver genes from NCG is shown in Table 3. The results show that the NetAP and HC-PIN identify the same number of validated cancer driver genes, which is higher than the number of driver genes identified by the MCODE, ClusterONE and OH-PIN methods. The driver genes identified by NetAP have the same enrichment in prostate cancer as those by the HC-PIN and have higher enrichment in prostate cancer than those by the other three methods. Therefore, the NetAP has a strong ability to identify driver genes.

4.7 | Comparison of the results of five clustering methods

The results of applying the MCODE, ClusterONE, HC-PIN, OH-PIN and NetAP methods to prostate cancer data from

TCGA are shown in Table 4, where the effective modules are driver modules with p -value <0.05 and the module size is ≥ 3 [27, 28]. It can be seen from Table 4 that both the HC-PIN and NetAP methods identify more candidate driver genes than the other three methods, which indicates that the HC-PIN and NetAP methods have a better ability to capture driver genes. However, HC-PIN identifies fewer candidate driver modules than the NetAP method, because there are large driver modules identified by the HC-PIN method. The large driver modules are not conducive to clinical drug target experiments. It can be seen that the OH-PIN method has the same situation as the HC-PIN method. Meanwhile the NetAP method identifies a large number of candidate driver modules, which is different from the HC-PIN and OH-PIN methods. Therefore, the average size of the modules identified by the NetAP method is more appropriate [27]. In addition, the NetAP and ClusterONE methods identify a larger number of effective driver modules compared with the other three methods. However, the effective driver modules identified by the NetAP method have a smaller p -value compared with the ClusterONE method. It shows that the driver modules identified by the NetAP method are more statistically significant. In summary, the NetAP method identifies more candidate driver genes, candidate driver modules and effective driver modules compared with other previous competitive methods. Moreover, the driver modules identified by the NetAP method have an appropriate size and are more statistically significant than other previous competitive methods.

4.8 | Module analysis

Seven driver modules with statistical significance are identified by the NetAP method in prostate cancer. They are analysed in

TABLE 3 Results of different methods to identify validated driver genes

Methods	The number of genes	Validated driver genes in prostate cancer	p -value
MCODE	1	TP53	N/A
ClusterONE	3	CHD1 NCOR1 TMRSS2	N/A
OH-PIN	8	APC ATM BRAF BRCA2 EPHA7 NCOR1 PTEN TP53	2.09E-03
HC-PIN	11	APC ATM BRAF BRCA2 EPHA7 NCOR1 PIK3CA PTEN RB1 TMRSS2 TP53	5.30E-08
NetAP	11	APC ATM BRAF BRCA2 CDKN1B EPHA7 NCOR1 PTEN RB1 TMRSS2 TP53	5.30E-08

TABLE 4 Results of different methods in prostate cancer

Methods	Number of candidate driver genes	Number of candidate driver modules	Average size of a module	Number of effective driver modules	Optimal p -value for effective driver module
MCODE	8	3	2.67	1	2.60E-04
ClusterONE	50	12	4.17	7	1.10E-06
OH-PIN	54	3	18.00	3	4.10E-15
HC-PIN	128	8	16.13	6	4.80E-15
NetAP	118	25	4.72	7	4.10E-15

detail and their roles in the development and progression of prostate cancer are elucidated.

4.8.1 | TP53 module

The TP53 module containing 49 genes is a critical driver module, and its network structure is shown in Figure 4a. The optimal *p*-value of the TP53 module is 4.1E-15 by using DAVID, which shows that the internal genes of the TP53

module have strong biological relevance and statistical significance. 10 statistically significant biological pathways closely related to the development and progression of prostate cancer are shown in Table 5. Many driver genes in the TP53 module are included in each biological pathway, showing that the TP53 module plays an important role in prostate cancer.

We analyse three signalling pathways, including the PI3K-Akt signalling pathway, FoxO signalling pathway and Wnt signalling pathway, in which the TP53 module has a significant enrichment. It can be seen from Table 5 that the *p*-value of the

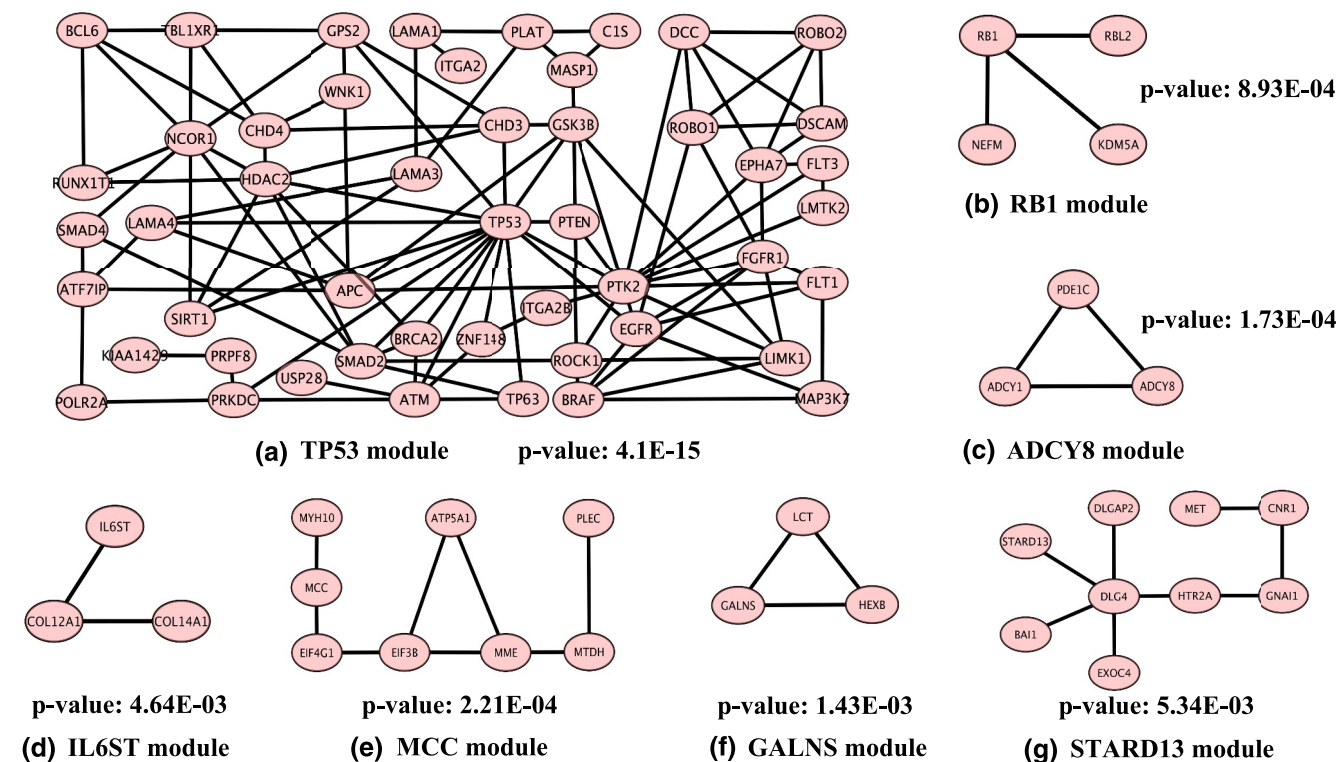


FIGURE 4 Optimal driver modules

TABLE 5 Biological pathways significantly enriched for TP53 modules

Term	Count	<i>p</i> -value	Genes
PI3K-Akt signalling pathway	12	2.68E-06	GSK3B, FLT1, LAMA1, LAMA4, ITGA2, LAMA3, ITGA2B, PTEN, TP53, PTK2, EGFR, FGFR1
Protein kinase activity	9	6.95E-06	GSK3B, ROCK1, WNK1, PRKDC, LIMK1, BRAF, MAP3K7, PTK2, EGFR
Histone deacetylation	5	6.95E-06	HDAC2, TBL1XR1, CHD4, CHD3, SIRT1
Protine serine/threonine kinase activity	9	9.74E-06	GSK3B, ROCK1, WNK1, PRKDC, LIMK1, LMTK2, BRAF, ATM, MAP3K7
FoxO signalling pathway	8	9.92E-06	SMAD2, SMAD4, BCL6, PTEN, BRAF, ATM, SIRT1, EGFR
Regulation of actin cytoskeleton	9	2.19E-05	APC, ROCK1, ITGA2, LIMK1, ITGA2B, BRAF, PTK2, EGFR, FGFR1
Prostate cancer	6	1.41E-04	GSK3B, PTEN, BRAF, TP53, EGFR, FGFR1
Histone deacetylase activity	4	1.83E-04	HDAC2, CHD4, CHD3, SIRT1
Protein tyrosine kinase activity	5	5.48E-04	EPHA7, FLT1, PTK2, EGFR, FGFR1
Wnt signalling pathway	6	1.13E-03	GSK3B, SMAD4, APC, TBL1XR1, MAP3K7, TP53

TP53 module is $2.68E-06$ in the PI3K-Akt signalling pathway. The PI3K-Akt signalling pathway is involved in and controls cell proliferation, apoptosis and tumourigenesis. Regulation of the PI3K-Akt signalling pathway affects the occurrence and development of prostate cancer. It has been reported that the PI3K-Akt signalling pathway is one of the most important ways to promote the development of prostate cancer, and its abnormal activation may promote cell invasiveness and promote the development of prostate cancer [44]. Similarly, the FoxO signalling pathway is involved in many cellular physiological events, such as apoptosis and cell cycle control. The p -value of the TP53 module is $9.92E-06$ in the FoxO signalling pathway. It can be seen from Table 5 that the eight members of the TP53 module participate in the FoxO signalling pathway. Regulation of the FoxO signalling pathway inhibits the functional reversal of the PC-3 cell viability, thereby achieving the purpose of treating prostate cancer [45]. The p -value of the TP53 module is $1.13E-03$ in the Wnt signal pathway. It can be seen from Table 5 that the six members of the TP53 module are involved in the Wnt signalling pathway. The Wnt signalling pathway plays an important role in prostate cancer, affects cell proliferation and polarity, and regulates the expression of factors related to tumour metastasis and development. It has been reported that gene mutations or expression changes in the Wnt signalling pathway are associated with prostate tumours [46]. Therefore, targeting the driver genes in the TP53 module has a better effect on the treatment of prostate cancer.

4.8.2 | RB1 module

The network structure of the RB1 module is shown in Figure 4b. The RB1 module contains only four driver genes but is altered in 73.01% of samples, and it means that the RB1 module has extremely high coverage. The p -value of The RB1 module is $8.93E-04$ in the regulation of lipid kinase activity, thus it has strong statistical significance. It has been reported that targeting the lipid kinase PIKfyve can inhibit autophagy and further affect metabolism and cell death [47]. Advanced prostate cancer is sensitive to immunotherapy by targeting lipid kinases. Therefore, it is speculated that the RB1 module is of great significance in the targeted therapy of prostate cancer.

4.8.3 | ADCY8 module

The network structure of the ADCY8 module is shown in Figure 4c. The p -value of the ADCY8 module is $6.7E-04$ in the Calcium signalling pathway, which indicates that the ADCY8 module has significant enrichment in the Calcium signalling pathway. It has been reported that the Calcium signalling pathway is involved in the malignant progression of prostate cancer cells mediated by androgens [48]. The incidence of prostate cancer can be reduced by using calcium signalling pathway blockers. Therefore, the ADCY8 module can affect the occurrence and development of prostate cancer by regulating the Calcium signalling pathway, and it may have a better

guiding significance for the clinical treatment of prostate cancer by targeting the ADCY8 module.

4.8.4 | IL6ST module

The network structure of the IL6ST module is shown in Figure 4d. The IL6ST module participates in the activity of collagen fibre tissue cells, and its p -value is $4.64E-03$. Collagen is composed of collagen fibrous tissue and plays an important role in bone metastasis of prostate cancer. It has been reported that bone metastasis of prostate cancer can cause degradation of only collagen in bone tissue, which leads to the further development of prostate cancer [49]. Bone metastasis is the main factor affecting the prognosis of prostate cancer. Therefore, the IL6ST module may have a great impact on the bone metastasis of prostate cancer.

4.8.5 | MCC module

The network structure of the MCC module is shown in Figure 4e. Seven driver genes in the MCC module are involved in cytoplasmic cell activities, and the p -value of the MCC module is $2.91E-03$. It has been reported that the PTOV1 protein in prostate tumour cells is located in the cytoplasm, and its overexpression can promote the proliferation of prostate tumour cells [50]. Therefore, the cytoplasmic life activities regulated by the MCC module are related to the occurrence and development of prostate cancer.

4.8.6 | GALNS module

It can be seen from Figure 4f that the GALNS module contains three interacting genes. The GALNS module participates in the activity of hydrolase active cells, and its p -value is $3.55E-03$. It has been reported that the hydrolase PSA can activate the growth factors VEGF-C and VEGF-D, thereby promoting the metastasis of prostate cancer [51]. Therefore, the GALNS module can affect the development of prostate cancer by regulating the activity of hydrolase.

4.8.7 | STARD13 module

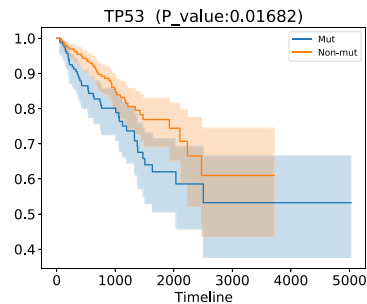
It can be seen from Figure 4g that the STARD13 module is composed of nine driver genes. Six of them are involved in the life activities of the cell membrane, and the p -value of the STARD13 module is $8.18E-03$. It has been reported that the related membrane proteins of the targeted protein adipocytes can be found by extracting the subcellular components on the membrane of prostate cancer cells. Membrane protein expression is increased in the presence of cholesterol, thereby promoting the development of prostate cancer [52]. Therefore, the STARD13 module has a strong driving effect on prostate cancer.

4.9 | Survival analysis

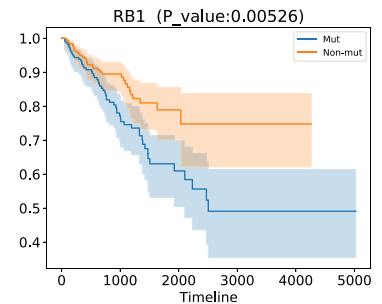
To verify the prognostic values of the identified driver modules for prostate cancer, we test the association of the major genes in the identified driver modules with patient survival. It can be seen from Figure 5 that the samples

with TP53, RB1, ADCY8, IL6ST, MCC, GALNS and STARD13 gene mutations have significantly shorter survival times than normal samples. Therefore, TP53, RB1, ADCY8, IL6ST, MCC, GALNS and STARD13 genes play a vital role in the occurrence and development of prostate cancer.

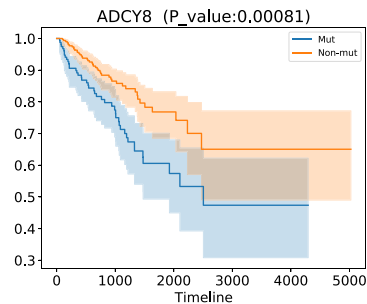
FIGURE 5 Survival analysis of major genes in identified modules



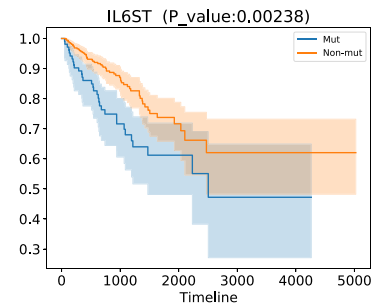
(a) TP53 survival analysis



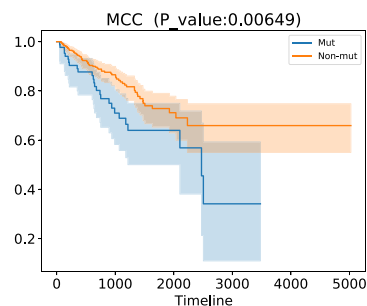
(b) RB1 survival analysis



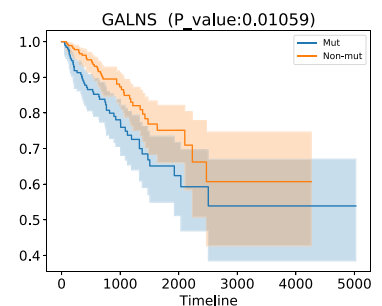
(c) ADCY8 survival analysis



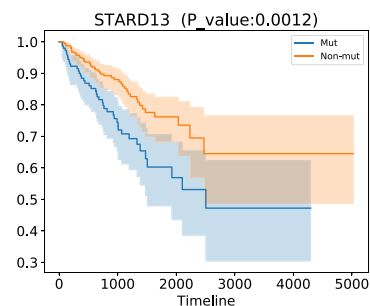
(d) IL6ST survival analysis



(e) MCC survival analysis



(f) GALNS survival analysis



(g) STARD13 survival analysis

4.10 | Application in gliomatosis

Gliomatosis is a complex disease that is difficult to treat and prone to relapse. It is of great significance to deeply explore the molecular mechanisms in the occurrence and development of gliomatosis. The findings can guide molecular typing and drug targets in the development of gliomatosis. Therefore, the NetAP algorithm is applied to the gliomatosis data from TCGA, and we obtain 1564 candidate driver genes and 88 driver modules. After DAVID enrichment analysis, we obtain 83 driver modules with high statistical significance and biological relevance. Most notably, we find that the EGFR gene-centred module is statistically significant, and its p -value is $8.7E-57$. The p -value of the EGFR module in gliomatosis is $3.8E-10$, which indicates that the EGFR module is highly enriched in gliomatosis. 22 known genes closely related to gliomatosis are included in the EGFR module, and they are involved in the occurrence and development of gliomatosis. Indeed, more than half of GBM patients have mutations in the EGFR gene, and anti-EGFR drugs have been used for GBM treatment [53]. In addition, a gliomatosis treatment programme targeting the BRAF gene has been formed clinically [54]. Mutations in PTEN and TP53 can lead to the formation of gliomatosis [55]. The PIK3CA mutation has been identified as a novel prognostic marker in gliomatosis [56]. However, resistance to these agents is a major problem clinically, thus it might benefit from targeting multiple genes in this module. Therefore, the EGFR modules identified by the NetAP algorithm may play an important role in the diagnosis and treatment of gliomatosis, and the NetAP algorithm also has a high ability to identify driver modules from gliomatosis data.

5 | CONCLUSIONS

In this study, we propose a computational method, NetAP, to identify oncogenic driver modules based on multi-omics biological networks. Firstly, this study integrates gene mutation, CNV, and PPI network data into the biological network, and then a weight function created based on high coverage, high mutual exclusivity, and high topological similarity among genes is used to calculate the interaction weights among genes in biological networks. Secondly, a random walk method is used to reassess the strength of interactions among genes. Finally, the affinity propagation algorithm is used to identify the optimal driver modules. It is experimentally verified that it is reasonable to consider the similarity index and random walk method in the NetAP method, as they improve the performance of the NetAP method. The experimental results show that the NetAP algorithm has higher accuracy, effectiveness and ability to identify the validated driver genes compared with the other four previous competing algorithms, and the driver modules identified by the NetAP method are of suitable size and have great statistical significance. Moreover, the biological functions of the driver modules identified in this study are analysed in detail, and it is confirmed that they are directly or

indirectly related to the occurrence and development of prostate cancer. Therefore, multi-omics-based biological networks are helpful in accurately identifying cancer driver modules. The results of the study help to explore the pathogenesis of cancer, and it is conducive to the diagnosis and drug targets of clinical cancer. Gene expression reflects the abundance of mRNA of gene transcription products measured directly or indirectly in cells, and methylation is an important modification of proteins and nucleic acids, which regulates the expression and closure of genes, and is closely related to many diseases, such as cancer, ageing, Alzheimer's disease and so on. Therefore, it is very meaningful to identify driver modules by integrating gene expression, gene mutation, methylation and PPI network.

ACKNOWLEDGEMENTS

We thank Jihua Dong for her careful proofreading and also thank Bing Zhou, Zhaoheng Ai, Mengdi Liu and Pengyu Zhang for their helpful advice and discussions. The scientific calculations in this paper have been done on the HPC Cloud Platform of Shandong University. This work is supported by the National Natural Science Foundation of China (Grant No. 61972322), the Natural Science Foundation of Shaanxi Province (Grant No. 2021JM110), the National Key Research and Development Program (Grant No. 2021YFF0704103) and the Fundamental Research Funds of Shandong University. The funders did not play any role in the design of the study, the collection, analysis, and interpretation of data, or the writing of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

This study used the gene mutation data and copy number variation data of prostate adenocarcinoma from TCGA (<https://portal.gdc.cancer.gov>), and the human protein-protein interaction network data from the Human Protein Reference database (<http://www.hprd.org>).

ORCID

Hao Wu  <https://orcid.org/0000-0003-2340-9258>

REFERENCES

1. Khurana, E., et al.: Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342(6154), 1235587 (2013). <https://doi.org/10.1126/science.1235587>
2. Kuen, K.L.A., Lai, F., Pycab, C.: Advances in colorectal cancer genomics and transcriptomics drive early detection and prevention. *Int. J. Biochem. Cell Biol.* 137, 106032 (2021). <https://doi.org/10.1016/j.biocel.2021.106032>
3. Hu, J., Chen, M., Zhou, X.: Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res.* 50(4), e21 (2022). <https://doi.org/10.1093/nar/gkab1147>
4. Hu, J., Zhong, Y., Shang, X.: A versatile and scalable single-cell data integration algorithm based on domain-adversarial and variational approximation. *Briefings Bioinf.* 23(1), bbab400 (2022). <https://doi.org/10.1093/bib/bbab400>

5. Zhao, X., et al.: Combinatorial CRISPR/Cas9 screening reveals epistatic networks of interacting tumor suppressor genes and therapeutic targets in human breast cancer. *Cancer Res.* 81(24), 6090–6105 (2021). <https://doi.org/10.1158/0008-5472.can-21-2555>
6. Choi, J., et al.: Evaluation of postmortem microarray data in bipolar disorder using traditional data comparison and artificial intelligence reveals novel gene targets. *J. Psychiatr. Res.* 141, 328–336 (2021). <https://doi.org/10.1016/j.jpsychires.2021.08.011>
7. Manoochehri, H., et al.: Identification of key gene targets for sensitizing colorectal cancer to chemoradiation: an integrative network analysis on multiple transcriptomics data. *J. Gastrointest. Cancer* 12, 1–20 (2021). <https://doi.org/10.1007/s12029-021-00690-2>
8. George, S., Ragin, C., Ashing, K.T.: Black is diverse: the untapped beauty and benefit of cancer genomics and precision medicine. *JCO oncology practice* 17(5), 279–283 (2021). <https://doi.org/10.1200/op.21.00236>
9. Miyabayashi, K., Nakagawa, H., Koike, K.: Molecular and phenotypic profiling for precision medicine in pancreatic cancer: current advances and future perspectives. *Front. Oncol.* 11, 682872 (2021). <https://doi.org/10.3389/fonc.2021.682872>
10. Zhao, J.F., et al.: Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* 28(22), 2940–7 (2012). <https://doi.org/10.1093/bioinformatics/bts564>
11. Leiserson, M.D., et al.: Simultaneous identification of multiple driver pathways in cancer. *Bioinformatics* 9(5), e1003054 (2014). <https://doi.org/10.1371/journal.pcbi.1003054>
12. Wu, J., et al.: Identifying mutated driver pathways in cancer by integrating multi-omics data. *Comput. Biol. Chem.* 80, 159–167 (2019). <https://doi.org/10.1016/j.compbiolchem.2019.03.019>
13. Wang, J., et al.: Cooperative driver pathway discovery via fusion of multi-relational data of genes, miRNAs and pathways. *Briefings Bioinf.* 22(2), 1984–1999 (2021). <https://doi.org/10.1093/bib/bbz167>
14. Vandin, F., Upfal, E., Raphael, B.J.: De novo discovery of mutated driver pathways in cancer. *Res. Comput. Mol. Biol.* 6577, 499–500 (2011)
15. Hicks, J.L., Liu, X., Williams, D.S.: Role of the ninac proteins in photoreceptor cell structure: ultrastructure of ninac deletion mutants and binding to actin filaments. *Cell Motil Cytoskeleton* 35(4), 367–379 (1996). [https://doi.org/10.1002/\(sici\)1097-0169\(1996\)35:4<367::aid-cm8>3.0.co;2-3](https://doi.org/10.1002/(sici)1097-0169(1996)35:4<367::aid-cm8>3.0.co;2-3)
16. Lee, J.H., et al.: Energy-dependent regulation of cell structure by AMP-activated protein kinase. *Nature* 477(7147), 1017–1020 (2011). <https://doi.org/10.1038/nature05828>
17. Chen, Y., et al.: Plac1 affects cell to cell communication by interacting with the desmosome complex. *Placenta* 110, 39–45 (2021). <https://doi.org/10.1016/j.placenta.2021.06.001>
18. Wang, X., et al.: Mutual dependency between lncRNA LETN and protein NPM1 in controlling the nucleolar structure and functions sustaining cell proliferation. *Cell Res.* 31(6), 664–683 (2021). <https://doi.org/10.1038/s41422-020-00458-6>
19. Freeman-Cook, K., et al.: Expanding control of the tumor cell cycle with a cdk2/4/6 inhibitor. *Cell Res.* 39(10), 1404–1421 (2021). <https://doi.org/10.1016/j.ccell.2021.08.009>
20. Beveridge, D.J., et al.: The tumor suppressor miR-642a-5p targets Wilms Tumor 1 gene and cell-cycle progression in prostate cancer. *Sci. Rep.* 11, 18003 (2021). <https://doi.org/10.1038/s41598-021-97190-x>
21. BaDer, G.D., Hogue, C.: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* 4(1), 2 (2003). <https://doi.org/10.1186/1471-2105-4-2>
22. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9(5), 471–472 (2012). <https://doi.org/10.1038/nmeth.1938>
23. Yu, W., Lin, G.: Detecting protein complexes by an improved affinity propagation algorithm in protein-protein interaction networks. *J. Comput.* 7, 1761–1768 (2012)
24. Wu, H., et al.: Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks. *PLoS One* 9(3), e91856 (2014). <https://doi.org/10.1371/journal.pone.0091856>
25. Leiserson, M., et al.: Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47(2), 106–114 (2015). <https://doi.org/10.1038/ng.3168>
26. Wu, H., Dong, J., Wei, J.: Network-based method for detecting dysregulated pathways in glioblastoma cancer. *IET Syst. Biol.* 12(1), 39–44 (2018). <https://doi.org/10.1049/iet-syb.2017.0033>
27. Ahmed, R., et al.: MEXCOWalk: mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics* 36, 872–879 (2020)
28. Wu, H., et al.: Integrating protein-protein interaction networks and somatic mutation data to detect driver modules in pan-cancer. *Interdiscipl. Sci. Comput. Life Sci.* 14, 1–17 (2021). <https://doi.org/10.1007/s12539-021-00475-y>
29. Virosztek, D.: The metric property of the quantum jensen-shannon divergence. *Adv. Math.* 380, 107595 (2021). <https://doi.org/10.1016/j.aim.2021.107595>
30. Guo, X.: Js-ma: a jensen-shannon divergence based method for mapping genome-wide associations on multiple diseases. *Front. Genet.* 11 (2020). <https://doi.org/10.3389/fgene.2020.507038>
31. Qi, Z., Li, M., Yong, D.: A new structure entropy of complex networks based on nonextensive statistical mechanics. *Int. J. Mod. Phys. C* 27(10), 1650118 (2016). <https://doi.org/10.1142/s0129183116501187>
32. Bahadori, S., Moradi, P., Zare, H.: An improved limited random walk approach for identification of overlapping communities in complex networks. *Appl. Intell.* 51(6), 1–20 (2021). <https://doi.org/10.1007/s10489-020-01999-4>
33. Anton, E., et al.: How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun. Phys.* 4(1), 133 (2021). <https://doi.org/10.1038/s42005-021-00634-z>
34. Nies, H.W., et al.: Enhanced directed random walk for the identification of breast cancer prognostic markers from multiclass expression data. *Entropy* 23(9), 1232 (2021). <https://doi.org/10.3390/e23091232>
35. Yeon, K.S., et al.: Multi-layered network-based pathway activity inference using directed random walks: application to predicting clinical outcomes in urologic cancer. *Bioinformatics* 37(16), 2405–13 (2021). <https://doi.org/10.1093/bioinformatics/btab086>
36. Yao, Y., et al.: Predicting lncRNA–disease association by a random walk with restart on multiplex and heterogeneous networks. *Front. Genet.* 12, 712170 (2021). <https://doi.org/10.3389/fgene.2021.712170>
37. Brendan, J.F., Delbert, D.: Response to comment on "clustering by passing messages between data points. *Science* 319(5864), 726 (2008). <https://doi.org/10.1126/science.1151268>
38. Peri, S., et al.: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13(10), 2363–2371 (2003). <https://doi.org/10.1101/gr.1680803>
39. Calinski, T., Harabasz, J.: A dendrite method for cluster Analysis. *Commun. Stat. Simulat. Comput.* 3, 1–27 (1974). <https://doi.org/10.1080/03610927408827101>
40. Wang, J., et al.: Identification of hierarchical and overlapping functional modules in PPI networks. *IEEE Trans. NanoBioscience* 11(4), 386–393 (2012). <https://doi.org/10.1109/tnb.2012.2210907>
41. Wang, J., et al.: A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE ACM Trans. Comput. Biol. Bioinf* 8(3), 607–620 (2011). <https://doi.org/10.1109/tcb.2010.75>
42. Gao, B., et al.: Prediction of driver modules via balancing exclusive coverages of mutations in cancer samples. *Adv. Sci.* 6(4), 1801384 (2019). <https://doi.org/10.1002/advs.201801384>
43. Dressler, L., et al.: Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. *Genome Biol.* 23(1), 35 (2022). <https://doi.org/10.1186/s13059-022-02607-z>
44. Shukla, S., et al.: Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion. *Int. J. Cancer* 121(7), 1424–1432 (2007). <https://doi.org/10.1002/ijc.22862>
45. Zhang, T., et al.: Proteomics reveals the function reverse of mpss-treated prostate cancer" associated fibroblasts to suppress pccell viability via the foxo pathway. *Cancer Med.* 10(7), 2509–2522 (2021). <https://doi.org/10.1002/cam4.3825>

46. Robinson, D.R., Zylstra, C.R., Williams, B.O.: Wnt signaling and prostate cancer. *Curr. Drug Targets* 9(7), 571–80 (2008). <https://doi.org/10.2174/138945008784911831>
47. Qiao, Y., et al.: Autophagy inhibition by targeting PIKfyve potentiates response to immune checkpoint blockade in prostate cancer. *Nat. Can. (Que.)* 2(9), 978–993 (2021). <https://doi.org/10.1038/s43018-021-00237-1>
48. Murtha, P.E., et al.: Effects of Ca⁺⁺ mobilization on expression of androgen-regulated genes: interference with androgen receptor-mediated transactivation by AP-1 proteins. *Prostate* 33(4), 264–270 (1997). [https://doi.org/10.1002/\(sici\)1097-0045\(19971201\)33:4<264::aid-pros7>3.0.co;2-h](https://doi.org/10.1002/(sici)1097-0045(19971201)33:4<264::aid-pros7>3.0.co;2-h)
49. Yu, L., et al.: Exosomes derived from osteogenic tumor activate osteoclast differentiation and concurrently inhibit osteogenesis by transferring COL1A1-targeting miRNA-92a-1-5p. *J. Extracell. Vesicles* 10(3), e12056 (2021). <https://doi.org/10.1002/jev2.12056>
50. Santamaría, A., et al.: PTOV-1, a novel protein overexpressed in prostate cancer, shuttles between the cytoplasm and the nucleus and promotes entry into the S phase of the cell division cycle. *Am. J. Pathol.* 162(3), 897–905 (2003). [https://doi.org/10.1016/s0002-9440\(10\)63885-0](https://doi.org/10.1016/s0002-9440(10)63885-0)
51. Jha, S.K., et al.: KLK3/PSA and cathepsin D activate VEGF-C and VEGF-D. *Elife* 17, e44478 (2019). <https://doi.org/10.7554/elife.44478>
52. Jiang, S., et al.: Cholesterol induces epithelial-to-mesenchymal transition of prostate cancer cells by suppressing degradation of EGFR through APMAP. *Cancer Res.* 79(12), 3063–3075 (2019). <https://doi.org/10.1158/0008-5472.can-18-3295>
53. Chong, D.Q., et al.: Combined treatment of Nimotuzumab and rapamycin is effective against temozolomide-resistant human gliomas regardless of the EGFR mutation status. *BMC Cancer* 15(1), 255 (2015). <https://doi.org/10.1186/s12885-015-1191-3>
54. Patrick, Y.W., et al.: Dabrafenib plus trametinib in patients with BRAF^{V600E}-mutant low-grade and high-grade glioma (ROAR): a multicentre, open-label, single-arm, phase 2, basket trial. *Lancet Oncol.* 23(1), 53–64 (2022). [https://doi.org/10.1016/s1470-2045\(21\)00578-7](https://doi.org/10.1016/s1470-2045(21)00578-7)
55. Verma, R., Lu, R.: Loss of PTEN and TRP53 in oligodendrocyte progenitors leads to glioma formation. *Neuro Oncol.* 21(Supplement_6), vi267 (2019). <https://doi.org/10.1093/neuonc/noz175.1121>
56. Draaisma, K., et al.: PI3 kinase mutations and mutational load as poor prognostic markers in diffuse glioma patients. *Acta Neuropathol. Commun.* 3(1), 88 (2015). <https://doi.org/10.1186/s40478-015-0265-4>

How to cite this article: Chen, Z., et al.: Identifying driver modules based on multi-omics biological networks in prostate cancer. *IET Syst. Biol.* 16(6), 187–200 (2022). <https://doi.org/10.1049/syb2.12050>