Computational and Structural Biotechnology Journal 27 (2025) 4332–4339

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

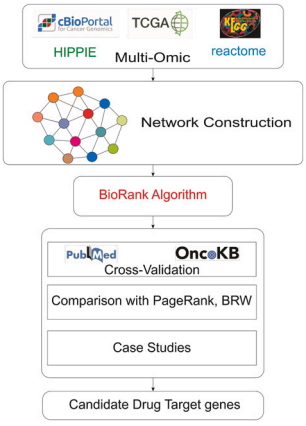
Research article

Prioritizing cancer therapeutic genes using BioRank: A biologically-informed PageRank framework

Duc-Tinh Pham [a,*], Huu-Tam Nguyen [b], Van-Hai Pham [b], Van-Thanh Le [c,d]

[a] School of Information and Communications Technology, Hanoi University of Industry, 298 Cau Dien street, Bac Tu Liem District, Hanoi, Viet nam
[b] School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi, Viet nam
[c] Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Viet nam
[d] Cyber School, Vinh University, Nghean, Viet nam

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
BioRank
Cancer
Drug target gene
Biomolecular networks
Gene expression information

ABSTRACT

The identification of therapeutic target genes constitutes a critical yet challenging aspect of cancer research, primarily due to the inherent complexities of biological systems and the heterogeneity of molecular data. This study introduces BioRank, an innovative gene prioritization methodology that extends the traditional PageRank algorithm by integrating biological insights through a custom-designed vector. This vector synthesizes differential gene expression, functional annotations (derived from GO, KEGG, and Reactome), and coexpression similarity to achieve a classification of enhanced biological significance. BioRank was validated using RNA sequencing data from The Cancer Genome Atlas (TCGA), alongside protein–protein interaction networks from HIPPIE across seven cancer datasets. Experimental results illustrate that BioRank effectively facilitates the identification and prioritization of therapeutic target genes. Comparative analysis with previous methodologies indicates that BioRank achieves superior predictive performance concerning both the number of target genes in OncoKB, as well as Recall@ and nDCG@ metrics. BioRank operates as a research instrument designed for hypothesis generation, prioritizing candidate therapeutic target genes based on a specified cancer type and standard molecular/network inputs. This empowers researchers to prioritize genes for subsequent biological validation, such as functional assays, while simultaneously retrieving known targets and identifying under-explored candidates.

* Corresponding author.
Email address: tinhpd@haui.edu.vn (D.-T. Pham).

https://doi.org/10.1016/j.csbj.2025.09.032
Received 2 July 2025; Received in revised form 22 September 2025; Accepted 22 September 2025
Available online 1 October 2025

1. Introduction

The advent of high-throughput technologies has enabled the representation of biological data as complex networks, where nodes denote genes or proteins and edges signify their interactions [1,2]. Employing this network paradigm, numerous computational models have been developed to assess the significance of nodes and edges. Genes or proteins that receive high rankings from these models frequently align with known disease targets, while those without prior evidence may emerge as novel candidates for experimental validation, consequently reducing the time and costs associated with wet-lab experiments.

Within the realm of network-based analytical methodologies, the PageRank algorithm [3] is employed to evaluate the prominence of individual nodes (genes/proteins within a biological network) grounded on the inherent connectivity framework of the network. Nodes (genes) that attain higher rankings are postulated to exert a substantial influence on pathological processes, thereby facilitating the identification of disease-associated genes and the exploration of potential therapeutic targets in cancer research. Nevertheless, conventional implementations of the PageRank algorithm predominantly focus on network topology, to the exclusion of critical biological attributes such as gene expression levels, functional annotations, and the biological similarities among genes [4]. Numerous studies have endeavored to address this limitation by incorporating biological information within the PageRank framework. Yet, these methodologies have frequently failed to fully exploit the synergy between interaction networks and gene-level biological features, resulting in suboptimal accuracy for the prioritization of cancer genes [5]. For instance, [6] employed a weighted PageRank algorithm to identify disease-associated genes utilizing PPI data, where edge weights represent the confidence scores of protein interactions. Although this approach improved prioritization by taking into account interaction reliability, it neglected other crucial biological data such as gene expression profiles, annotations, or pathway information, thus limiting its efficacy, particularly for complex diseases such as cancer, where the integration of multimodal data is essential. In another study, [7] compared various network propagation methods and examined multilayer data integration by fusing both PPI and gene expression networks. Despite the performance enhancements achieved through multi-omics integration, the PageRank formulation was not tailored to specific data types or disease contexts, and pathway-specific or annotation data were not incorporated, thereby constraining the model's ability to accurately identify disease-associated genes. Building on this research trajectory, [8] proposed a multi-layer molecular interaction network derived from heterogeneous omics data such as RNA-seq, miRNA-seq, and other gene-level features. While this approach leveraged network heterogeneity, its efficacy was heavily dependent on the quality and completeness of the input data, with incomplete or inconsistent datasets significantly impairing predictive accuracy. More recently, [9] introduced the Constrained PageRank (CPR), which integrates multiple omics layers, including RNA-seq, DNA methylation, miRNA, and somatic mutations, into a coherent biological network to enhance disease gene prediction. Despite demonstrating substantial improvements over previous models, CPR also contended with issues of missing or inconsistent input data. Additionally, the lack of pathway-specific or gene annotation integration limited its applicability to complex diseases such as cancer, where such information is crucial for precise gene ranking.

In this study, we introduce BioRank, a tool designed to facilitate the identification and prioritization of therapeutic target genes, reinforce the scientific basis for potential candidate genes, and propose novel candidate genes prioritized for subsequent biological experiments. BioRank represents an advancement over PageRank, demonstrating efficiency in biological data analysis due to its comprehensive strategy of integrating essential biological features of genes and proteins from diverse heterogeneous sources. Specifically, gene annotation information from Gene Ontology [10], Reactome [11], and KEGG Pathway [12,13] is utilized to assess the functional relevance of genes to cancer. Moreover, differential gene expression analysis is employed to determine expression changes between tumor samples and controls. A personalized vector synthesized from multiple biological information sources, is employed to more accurately capture the biological role and significance of each gene within the network. Furthermore, a convex combination strategy is implemented to optimize the contributions of various data sources. These enhancements allow BioRank to surpass previous gene prioritization methods by leveraging both the network structure and the detailed biological attributes of each gene. Our findings contribute to the advancement of precision medicine [14,15] and facilitate the reduction of sample sizes needed for assay validation in clinical settings [16] (Figs. 1 and 2).

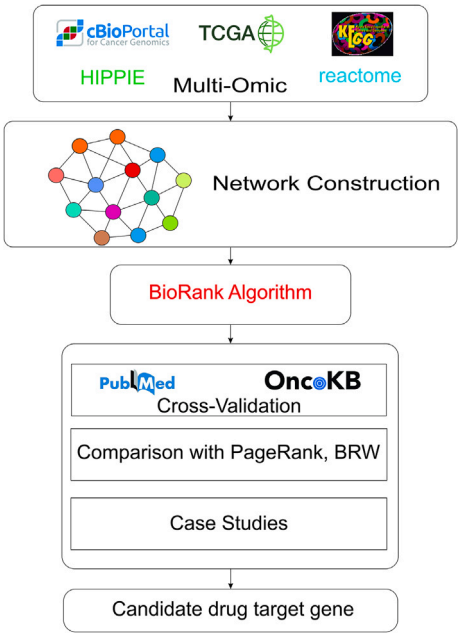


Fig. 1. Overview of the BioRank framework for therapeutic gene prioritization. The inputs consist of TCGA RNA-seq tumor data, the HIPPIE protein-protein interaction (PPI) network, and functional annotations derived from Gene Ontology (GO), KEGG, and Reactome. Node evidence is synthesized from (i) enrichment of annotations and (ii) differential expression between tumor and control conditions, as well as network proximity, culminating in a personalized vector ($q = \alpha\theta + (1 - \alpha)\phi$). Edges are assigned weights based on functional similarity and co-expression ($W = \beta W_1 + (1 - \beta)W_2$). A personalized PageRank algorithm, incorporating a damping factor d, is employed to propagate scores within the weighted network to generate a prioritized list of candidate genes. The top predictions are evaluated against OncoKB and corroborated by PubMed, with performance assessed using Recall@ and nDCG@ metrics across seven TCGA cancer types: BRCA, COAD, LUAD, THCA, BLCA, PRAD, and STAD.

2. Data and method

2.1. Data

In the development and training of the BioRank model, four primary types of input data were employed:

The protein–protein interaction (PPI) network was acquired from HIPPIE version 2.2 database [17], a dedicated repository offering detailed information on human protein interactions, accompanied by confidence scores derived from diverse evidence sources. Only those interactions with confidence scores exceeding 0.7 were retained, thereby ensuring the precision and dependability of the network. The refined PPI network constituted the foundational graph for signal propagation within the enhanced PageRank algorithm.

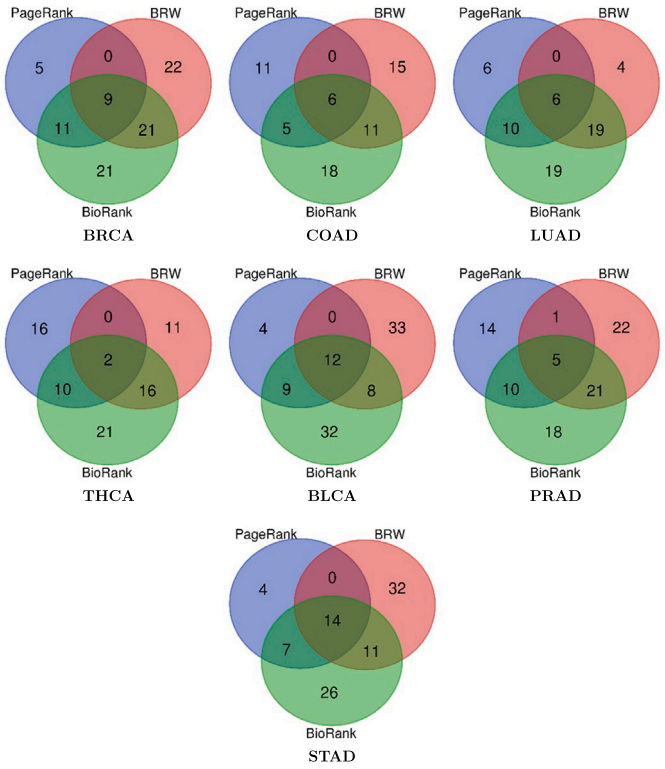


Fig. 2. Shared and unique target genes predicted top 100 by BioRank, BRW, and PageRank across cancer datasets. A comparative visualization of the therapeutic gene predictions generated top 100 by BioRank, BRW, and PageRank across seven cancer datasets. This figure effectively demonstrates both the concordance and divergence among the three algorithms. It is noteworthy that BioRank consistently identifies the majority of known targets detected by existing methods, while also contributing a significant number of unique predictions, many of which are corroborated by PubMed evidence. This dual characteristic highlights BioRank's capacity to balance sensitivity (recovering established cancer genes) and novelty (identifying underexplored candidates).

Gene Ontology Annotations (Ontology Graph). Each gene can be associated with a variety of biological descriptors, including functions, biological processes, and signaling pathways. Annotation data were assimilated from three principal sources: Reactome [11], Gene Ontology (GO) [10], and KEGG pathways [12,13]. The annotations underwent statistical filtration using Fisher's Exact Test and False Discovery Rate (FDR) correction [18,19] to exclude associations deemed unreliable. This procedure guarantees that only statistically significant associations are preserved for the construction of the ontology graph.

Seed Genes. For each cancer type, a seed gene set was developed consisting of genes verified to be linked with the respective disease. These genes were sourced from cBioPortal for Cancer Genomics [25] and subsequently refined to include solely cancer driver genes exhibiting a mutation frequency exceeding 1 %. This seed set served as the initial input for the personalized vector in the enhanced PageRank model.

Gene expression profiles were retrieved from the TCGA database [20], encompassing both tumor and control samples. We conducted an analysis of RNA-seq data to generate cancer-specific tumor-control gene expression matrices for each cancer type. These matrices facilitated the identification of differentially expressed genes and co-expression networks throughout the preprocessing phase.

In order to assess the performance of the model, a validation set consisting of disease-associated genes sourced from the OncoKB database [26,27] was utilized. The OncoKB database is a curated repository of cancer-related genes, underpinned by clinical evidence and associations with FDA-approved therapeutics. This gene set functioned as the ground truth for evaluating the accuracy and effectiveness of the proposed model.

2.2. Node weight computation based on biological information

In the present investigation, we employed the methodology delineated in [21] to determine both node and edge weights by integrating protein–protein interaction (PPI) networks with a wide array of biological data sources. Our algorithm deviates from the traditional reliance on node degree, which is the count of interactions per gene, by also incorporating gene annotation data and gene expression levels sourced from repositories such as Gene Ontology (GO) [10] and KEGG pathways [12,13]. In contrast to the conventional PageRank algorithm, where all nodes are initialized with uniform importance scores [22], which inadequately addresses the biological heterogeneity inherent in real-world molecular networks, BioRank initializes node scores using a composite metric derived from both gene annotations and gene expression data. This approach more accurately mirrors the functional and transcriptional significance of each gene.

A) annotation-based node weight computation. Assume we have l distinct biological sources, each providing gene annotations for the set of genes under study. Let S denote the seed gene set. To eliminate unreliable annotations, we apply statistical enrichment analysis using the Fisher Exact Test [18]. Subsequent correction for multiple comparisons is performed using the False Discovery Rate (FDR) approach [19], with level $P\_value < 10^{-5}$.

For every source $j \in \{1, \ldots, l\}$, define F^j as the subset of annotations exhibiting significant enrichment in at least one gene from the seed set S. The aggregate set of reliable annotations F encompassing all sources is delineated as:

$$F = \bigcup_{j=1}^{l} F^j \tag{1}$$

For each gene i (regardless of whether $i \in S$), let $\mathcal{A}(i)$ be the set of annotations assigned to gene i from all sources. The annotation-based biological score θ_i of gene i is then computed as follows:

$$\theta_i = \begin{cases} \ell, & \text{if } i \in S \\ \sum_{j=1}^{l} \frac{|\mathcal{A}(i) \cap F^j|}{|F^j|}, & \text{otherwise} \end{cases} \tag{2}$$

where, ℓ is a large constant introduced to ensure that known disease genes are assigned the highest possible priority during initialization. For genes not included in S, the value of θ_i is calculated based on the degree of overlap between the gene's annotations and the set of reliable annotations, normalized by the size of each source.

B) expression-based node weight computation. Let $L \in \mathbb{R}^{n \times m}$ denote the gene expression matrix. Here, n denotes the total number of genes, while m represents the number of patient samples. The element l_{ij} represents the expression level of gene i in sample j.

To normalize the data, the Z-score for each element is computed as follows:

$$z_{ij} = \frac{l_{ij} - \mu_i}{\sigma_i} \tag{3}$$

where μ_i is the mean and σ_i is the standard deviation of the expression level of gene i across all samples.

Construct a binary matrix Z based on the Z-matrix, where:

$$\bar{z}_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 2.5 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

A threshold of $z_{ij} > 2.5$ was selected to ensure that only genes with expression levels substantially higher than the mean are considered as

differentially expressed. This threshold is commonly used in transcriptomic analyses to reduce noise and increase the specificity for identifying target genes.

A gene g is considered differentially expressed if the number of patients in which gene g is identified as differentially expressed exceeds the mean value, calculated as follows:

$$\frac{1}{m}\sum_{j=1}^{m} z_{gj} > \frac{1}{n\times m}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \tag{5}$$

Gene weight computation: Use the PPI network to determine a topological vector ϕ and normalize ϕ_i for the differentially expressed genes:

$$\phi_i = \frac{|\mathcal{N}(i)\cap S|}{|\mathcal{N}(i)|} + \frac{|\mathcal{N}_2(i)\cap S|}{|\mathcal{N}_2(i)|} \tag{6}$$

where $\mathcal{N}(i)$ denotes the set of genes that are immediate neighbors (first-order neighbors) of gene i, and $\mathcal{N}_2(i)$ represents the set of genes located at a distance of two (second-order neighbors).

After obtaining the two weight vectors, we integrate them using a convex combination to construct the personalization vector q as input for the BioRank algorithm:

$$q = \alpha\theta + (1-\alpha)\phi \tag{7}$$

where $\alpha \in [0;1]$ is a parameter that adjusts the balance of contribution of each data source.

2.3. Edge weight computation based on biological information

Edge weights reflect the interaction strength between genes in the PPI network.

A) using gene annotations. Construct a weighted transition matrix $\mathbf{W}$, where W_{ij} depends on the extent to which genes i and j share common annotations related to biological processes specific to the disease:

$$DSI(i,j) = |\mathcal{A}(i)\cap\mathcal{A}(j)\cap F| \tag{8}$$

where $\mathcal{A}(i)$ is the set of annotations associated with the gene i; F is the set of statistically significant disease-related annotations.

Update the matrix $\mathbf{W}$ using the DSI function:

$$W_{ij}^{1} = \begin{cases} c + DSI(i,j), & \text{if } (i,j)\in E \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where c is a positive constant used to ensure a minimum weight when no biological information is available at the ends of a PPI edge.

B) using co-expression information.

$$W_{ij}^{2} = \frac{P_{ij}}{\sum_{k\in\mathcal{V}(i)} P_{ik}} \tag{10}$$

where P_{ij} is the Pearson correlation coefficient [23] between genes i and j.

C) integration of edge weights from two sources.

$$\mathbf{W} = \beta\cdot\mathbf{W}^{1} + (1-\beta)\cdot\mathbf{W}^{2} \tag{11}$$

where $\beta \in [0;1]$ is a parameter that adjusts the balance of contributions from each data source.

2.4. Proposed enhanced PageRank algorithm for disease gene prioritization

The PageRank algorithm [22] can be used to evaluate the importance of each node based on the link structure of the network. The general formula for computing the PageRank score of any node v in the network is as follows:

$$PR(v) = \frac{1-d}{N} + d\sum_{u\in B(v)}\frac{PR(u)}{L(u)} \tag{12}$$

where N is the total number of nodes, d is the damping factor (typically set to 0.85), $B(v)$ is the set of nodes that link to node v, $L(u)$ is the number of outbound links from node u, and $PR(u)$ is the PageRank score of node u.

In this research, we present an innovative methodology aimed at accurately predicting potential therapeutic target genes in cancer through the introduction of an enhanced version of the PageRank algorithm. Our approach incorporates crucial bioinformatics features related to genes and proteins, drawing from annotations provided by Gene Ontology [10], Reactome [11], and KEGG Pathway information [12,13, 24], to evaluate the relevance of genes to cancer. Subsequently, we conduct gene expression analysis to identify genes that are differentially expressed between disease samples and control samples. A protein–protein interaction (PPI) network [25] is employed to establish the network topology and allocate initial weights. Ultimately, the PageRank formula is augmented by incorporating a personalized vector, which accounts for the specific biological roles of individual genes.

The modified PageRank score for a node v is calculated as follows:

$$PR(v) = (1-d)\cdot q(v) + d\sum_{u\in B(v)}\frac{PR(u)\cdot w(u,v)}{\sum_{k\in V(u)} w(u,k)} \tag{13}$$

where:

- $w(u,v)$: edge weight between nodes u and v derived from co-expression or biological similarity, computed by:

  $\mathbf{W} = \beta\cdot\mathbf{W1} + (1-\beta)\cdot\mathbf{W2}$

- $\sum_{k\in V(u)} w(u,k)$: total weight of outgoing edges from node u.
- $q(v) = \alpha\cdot\theta + (1-\alpha)\cdot\phi$: personalized vector of vertex v (vertex weight of v).

Algorithm 1: Improved_PageRank for BioRank model.

Input: $G(V,E)$: protein–protein interaction network
q: personalized initialization vector
d: damping factor (typically 0.85)
N: number of nodes
ε: convergence threshold
$w(u,v)$: edge weight from node u to node v
Output: PR: final PageRank score vector

```
foreach v ∈ V do
    PR[v] ← q[v]
repeat
    PR_new ← { }
    foreach v ∈ V do
        s ← 0
        foreach u ∈ V(v) do
            W ← Σ_{k∈V(u)} w(u,k)
            s ← s + PR[u]·w(u,v) / W
        PR_new[v] ← (1−d)·q(v) + d·s
    δ ← max_{v∈V} |PR_new[v] − PR[v]|
    PR ← PR_new
until δ < ε
return PR
```

3. Results and discussion

3.1. Predicted results of cancer therapeutic target genes

In this study, we utilized the OncoKB database to benchmark our experimental results. OncoKB (Oncology Knowledge Base) is a well-established, cancer-specific database maintained by the Memorial Sloan Kettering Cancer Center (MSKCC), which provides comprehensive and curated information on cancer-related gene variants [26,27]. Experiments were conducted on seven datasets corresponding to seven prevalent cancer types: breast carcinoma (BRCA), colorectal cancer (COAD), lung adenocarcinoma (LUAD), thyroid cancer (THCA), bladder cancer (BLCA), prostate cancer (PRAD), and stomach cancer (STAD). We evaluated the prediction performance of our algorithm based on the top 15 ranked genes using three metrics: (1) The number of matched therapeutic target genes found in the OncoKB database; (2) Recall@ quantifies the proportion of relevant (i.e., known disease-associated) genes that are successfully retrieved within the top K predictions generated by the algorithm; (3) nDCG@ (normalized discounted cumulative gain), which measures ranking accuracy with respect to the positions of listed genes in the prioritized results. The complete list of the top 100 ranked genes is available in the accompanying supplementary information file.

Table 1 provides a summary of the top 15 genes prioritized by BioRank across seven cancer types. It details, for each gene, its classification by OncoKB as either an oncogene or a tumor-suppressor gene, supported by relevant PubMed identifications (PMIDs listed in the table). Across various cancers, the majority of these top-ranked genes are established drivers, including tumor-suppressor genes such as TP53 (TSG), BRCA1 (TSG), and EP300 (TSG), alongside oncogenes like EGFR, ESR1, MYC, PIK3CA, and AKT1.

CTNNB1, ERBB2, MAPK1, and SRC, thus indicating that BioRank emphasizes biologically credible targets at the top of its list. This is corroborated by high OncoKB validation rates among the top 15 (e.g., BRCA 93.3 %, BLCA 86.6 %, STAD 86.6 %, LUAD 80 %). In addition, entries within the top-ranked list that currently lack OncoKB tags are nonetheless recognized in literature as “potential candidate genes” (e.g., GRB2, SUMO2, RELA, TRIM28, FN1, ALB, PNP, CCR3, CDH7, SOX1, CCL18), suggesting they are plausible emerging targets. Furthermore, BioRank introduces a limited set of genes for “further studies” (PCDHA4, GPR161, CCL17) which are not yet curated but are prioritized through biology-based graph propagation, thereby presenting concrete hypotheses for experimental investigation. The rightmost column presents the run time per cancer-specific network (ranging from 334 to 4995 s), demonstrating scalability across networks with comparable numbers of nodes but varying edge structures. Collectively, these findings elucidate that BioRank effectively recovers established oncogenes and tumor-suppressors and identifies credible novel candidates for further validation.

3.2. Compared to other methods

To assess the predictive efficacy of BioRank, we conducted a comparative analysis of its outputs with those of two alternative methodologies: the original PageRank algorithm and the BRW algorithm [21], utilizing seven datasets related to prevalent cancer types. It is pertinent to mention that in study [21], the BRW algorithm underwent empirical comparison against four additional methods: RWR [28], DIAMOnD [29], DADA [30], and RWR-M [31]. Consequently, these four algorithms are excluded from our evaluation, as their comparative performance with BRW has been extensively documented. Experiments were executed using both the top 15 and top 100 ranked genes, scrutinizing various facets including the quantity of predicted genes, ranking quality, coverage, and the extent of concordance and reliability among the predicted outcomes.

It is important to acknowledge that all seven biological network datasets originate from the HIPPIE version 2.2 database, hence they encompass an identical number of genes (refer to Section 4.4, DE Gene and Co-expression Network, in the UserManual-BioRank for further details). Nonetheless, the number of edges varies as a consequence of the distinctive properties of the seed set correlated with each disease.

Table 2 provides an in-depth comparative analysis of three methodologies for gene prioritization, focusing on the top 15 genes: the conventional PageRank, the Biological Random Walk (BRW), and the newly introduced BioRank, evaluated across seven prominent cancer datasets. The assessment emphasizes three crucial metrics: the number of therapeutic target genes aligned with OncoKB (Match), the recall in the top 15 predictions (Recall@15), and the normalized Discounted Cumulative Gain (nDCG@15), appraising both relevance and ranking position. The results unequivocally demonstrate that BioRank consistently outperforms both PageRank and BRW in all three metrics across all datasets.

4. Discussion

The predictive outcomes indicate that BioRank constitutes a promising tool engineered to aid in the identification and prioritization of therapeutic target genes, including TP53, ESR1, EGFR, AKT1, and MYC. It solidifies the scientific basis for potential candidate genes such as GRB2, SUMO2, RELA, TRIM28, FN1, ALB, PNP, CCR3, CDH7, SOX1, and CCL18. Additionally, it introduces novel candidate genes like PCDHA4, GPR161, and CCL17, thereby emphasizing them as priorities for ensuing biological experiments. A critical aspect contributing significantly to the enhanced predictive performance is the formulation of a tailored weighted vector that incorporates biological information derived from seed genes. Rather than utilizing a uniform vector, the incorporation of biologically relevant information facilitates signal propagation in a manner that accurately mirrors biological mechanisms, thereby enhancing the identification of functionally relevant genes. Moreover, edge weighting, representing both the reliability and biological significance of protein–protein interactions, is indispensable. By steering signal propagation more selectively and mitigating the dilution of signal through low-confidence edges, the method further hones its ability to prioritize biologically significant genes. An exhaustive analysis of the results implies that Recall@ and nDCG@ provide considerable advantages for gene ranking objectives. A high Recall@ and nDCG@ value signifies that validated target genes are prioritized toward the top of the predictive hierarchy, an essential factor in assisting researchers with the prioritization of candidate genes for subsequent biological experimentation.

For instance, within the context of the breast cancer dataset, TP53 functions as a critical tumor suppressor gene that is frequently inactivated in breast cancer. The TP53 Y220C variant constitutes a missense mutation within the DNA-binding domain, leading to destabilization of p53 by disrupting five electrostatic interactions [32], and fails to restore the transcriptional activity of wild-type TP53 in reporter assays [33]. ESR1 (estrogen receptor alpha) acts as a transcription factor that is commonly mutated in hormone-resistant metastatic breast carcinomas. ESR1 encodes ERα; upon estrogen binding, it facilitates the release of HSP90, instigates the dimerization of ERα/ERβ, and facilitates their translocation to the nucleus. Through interaction with ERE/AP-1/SP1 in conjunction with co-regulators, it governs cellular processes such as proliferation, migration, and differentiation [34–37]. AKT1, an intracellular kinase, is frequently subject to mutation in various cancer types, including breast and endometrial carcinomas. The AKT1 E17K mutation, located within the pH-domain, is an activating mutation that enhances PI3K/AKT signaling and promotes oncogenic phenotypes [38–42]. This mutation is additionally associated with Proteus syndrome and breast cancer [43,44].

Within the colorectal cancer (COAD) dataset, the epidermal growth factor receptor (EGFR) gene harbors a specific missense mutation, G465E, which is localized in the extracellular domain of the protein. This mutation has been identified in instances of colorectal cancer [45].

Table 1
Top 15 Prioritized Genes Across Seven Cancer Types by BioRank with Validation from OncoKB and PubMed.

Biomolecular Networks	Gene name	Evidence from the OncoKB		Evidence from the PubMed	Execution Time (s)	Biomolecular Networks	Gene name	Evidence from the OncoKB		Evidence from the PubMed	Execution Time (s)
		Is Onco Gene	Is Tumor Suppressor Gene					Is Onco Gene	Is Tumor Suppressor Gene		
BRCA	TP53		Yes	12619115	333.86	THCA	MYC	Yes		30226440	2063.53
	ESR1	Yes		31318440			CCL17	For further studies			
	EGFR	Yes		16261406			CTNNB1	Yes		33846546	
	GRB2	Potential candidate gene		29550383			PIK3CA	Yes		18000091	
	PIK3R1		Yes	38153569			ETV4	Yes		34283663	
	EP300		Yes	28341962			FN1	Potential candidate gene		39268167	
	AKT1	Yes		35892586			EP300		Yes	10700188	
	BRCA1		Yes	12767038		BLCA	TP53		Yes	36568387	410.32
	MYC	Yes		21779462			EGFR	Yes		34991599	
	PIK3CA	Yes		36279023			CTNNB1	Yes		37740194	
	ERBB2	Yes		31037288			GRB2	Potential candidate gene		10995035	
	HDAC1	Yes		28779562			PIK3R1		Yes	34668023	
	RAF1	Yes		7834453			EP300		Yes	36647005	
	MAPK1	Yes		33213267			HDAC1	Yes		31861435	
	CTNNB1	Yes		7005411			MYC	Yes		37105989	
COAD	CTNNB1	Yes		33115416	4995.42		ESR1	Yes		30511377	
	EGFR	Yes		33825902			SRC	Yes		39794543	
	TP53		Yes	33924934			AKT1	Yes		35317488	
	EP300		Yes	12385008			BRCA1		Yes	35395863	
	MYC	Yes		35972682			HSP90AA1	Yes		37000291	
	SUMO2	Potential candidate gene		37338025			RELA	Potential candidate gene		28586003	
	RELA	Potential candidate gene		34867383			RAF1	Yes		34554931	
	HDAC1	Yes		39260334		PRAD	CTNNB1	Yes		36750551	3692.15
	TRIM28	Potential candidate gene		29631612			TP53		Yes	37163614	
	WIF1		Yes	34627187			CDH7	Potential candidate gene		37444571	
	ESR1	Yes		32266127			EP300		Yes	33705753	
	FN1	Potential candidate gene		29274284			EGFR	Yes		32678075	
	AKT1	Yes		38891994			POU3F2	Yes		27784708	
	PIK3R1		Yes	31203132			PCDHA4	For further studies			
	GRB2	Potential candidate gene		18192688			GRB2	Potential candidate gene		33707553	
LUAD	EGFR	Yes		32053675	1159.37		HDAC1	Yes		32546700	
	CTNNB1	Yes		32442860			MYC	Yes		35562350	
	MYC	Yes		32003251			PIK3R1		Yes	35670774	
	GRB2	Potential candidate gene		27449805			ESR1	Yes		23805288	
	AKT1	Yes		36350496			AKT1	Yes		32451180	
	PIK3R1		Yes	34858053			SOX1	Potential candidate gene		20929579	
	TP53		Yes	38164123			CCL18	Potential candidate gene		25197632	
	RAF1	Yes		17315157		STAD	TP53		Yes	32007736	1480.75
	FN1	Potential candidate gene		27207836			EGFR	Yes		20430735	
	SUMO2	Potential candidate gene		37948404			CTNNB1	Yes		37054973	
	MAPK1	Yes		30972766			GRB2	Potential candidate gene		19337752	
	ALB	Potential candidate gene		38973954			HDAC1	Yes		35686089	
	EP300		Yes	38048728			EP300		Yes	21390126	
	ERBB2	Yes		38154514			PIK3R1		Yes	38948250	
	SRC	Yes		12826049			MYC	Yes		38169774	
THCA	TP53		Yes	39940804	2063.53		ESR1	Yes		33438526	
	PNP	Potential candidate gene		12629124			BRCA1		Yes	35077220	
	PCDHA4	For further studies					AKT1	Yes		39724412	
	EGFR	Yes		15623643			SRC	Yes		20406949	
	GPR161	For further studies					RELA	Potential candidate gene		39821576	
	GRB2	Potential candidate gene		19027225			CREBBP		Yes	23839013	
	CCR3	Potential candidate gene		19731977			JUN	Yes		27882939	
	PIK3R1		Yes	21487925							

Table 2
Performance Comparison of PageRank, BRW, and BioRank on the top 15 genes across Cancer Datasets.

Biomolecular Networks	Properties		PageRank			BRW			BioRank		
	Nodes	Edges	Match	Recall@15	nDCG@15	Match	Recall@15	nDCG@15	Match	Recall@15	nDCG@15
BRCA	12,148	219,166	6	0.0051	0.4020	6	0.0060	0.3255	**14**	0.0119	**0.9265**
COAD	12,148	799,078	6	0.0051	0.3964	2	0.0034	0.2070	**10**	0.0085	**0.7422**
LUAD	12,148	337,686	7	0.0060	0.4477	2	0.0062	0.2106	**11**	0.0102	**0.8258**
THCA	12,148	547,306	7	0.0050	0.4404	3	0.0034	0.2113	**8**	0.0068	**0.5304**
BLCA	12,148	237,288	6	0.0051	0.4020	9	0.0074	0.7168	**13**	0.0111	**0.8829**
PRAD	12,148	607,492	7	0.0060	0.4194	6	0.0054	0.3755	**10**	0.0101	**0.7177**
STAD	12,148	271,464	6	0.0051	0.4005	10	0.0102	0.8604	**13**	0.0116	**0.8817**

Within the LUAD dataset, the epidermal growth factor receptor (EGFR), a receptor tyrosine kinase, exhibits alterations through amplification and/or mutation in lung and brain malignancies, among others. In-frame deletions of exon 19 of the EGFR gene lead to constitutive activation of EGFR tyrosine kinase activity and render the receptor sensitive to tyrosine kinase inhibitors (TKIs), such as gefitinib, erlotinib, and afatinib, in lung adenocarcinoma [46–48]. The aforementioned drugs, afatinib, erlotinib, and gefitinib, have received FDA approval for the treatment of patients with non-small cell lung cancer that harbors EGFR exon 19 deletions.

5. Conclusion

In this study, we introduce BioRank, an advanced gene prioritization method that enhances the classical PageRank algorithm through the incorporation of multimodal biological data. Unlike conventional methodologies that rely solely on network topology, BioRank integrates gene-level annotations and expression profiles using a personalized vector, thereby enabling a biologically more meaningful initialization of nodes. The algorithm further refines the accuracy of rankings by weighting edges based on functional similarity and gene co-expression. The design of BioRank ensures that genes with biological relevance receive greater initial priority, and that signal propagation is guided by reliable interaction strengths. This approach enhances the identification of genes potentially involved in cancer pathogenesis. Validation across seven cancer-related datasets demonstrates BioRank's superior performance compared to existing methods, as evidenced by higher Recall@, nDCG@ scores and a greater number of matched therapeutic targets listed in OncoKB and PubMed. Notably, BioRank effectively prioritized well-established cancer genes such as TP53, ESR1, EGFR, AKT1, and MYC at top ranks, while also identifying less-explored but potentially promising candidates such as GRB2, SUMO2, RELA, TRIM28, FN1, ALB, PNP, CCR3, CDH7, SOX1, and CCL18. Furthermore, it suggested novel genes not previously reported, such as PCDHA4, GPR161, and CCL17, thereby highlighting them as priorities for future biological investigations. These findings underscore the practical utility of integrating network structure with functional genomics to improve the accuracy and interpretability of cancer gene prioritization. The algorithm and its supporting resources are made publicly available to facilitate reproducibility and further research.

Planned future work (beyond the current scope) involves evaluating BioRank with harmonized

TCGA–GTEx resources (and explicit batch-correction pipelines) as a sensitivity analysis. Additionally, further experiments could assess the impact of the α and β values on predictive performance, as well as the specific contribution of each data source to the model's predictive capacity. We contend that these additions will strengthen the manuscript while preserving the integrity and comparability of the results presented here.

CRediT authorship contribution statement

Duc-Tinh Pham: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Software, Conceptualization. **Huu-Tam Nguyen:** Writing – original draft, Visualization, Software, Resources, Data curation. **Van-Hai Pham:** Writing – review & editing, Supervision, Project administration. **Van-Thanh Le:** Visualization, Software, Resources, Funding acquisition.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the authors used ChatGPT in order to improve language of the manuscript for readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by Hanoi University of Industry under Project No. 17–2025-RD/HD-DHCN.

References

[1] Tran T-D, Pham D-T. Identification of anticancer drug target genes using an outside competitive dynamics model on cancer signaling networks. Sci Rep 2021;11(1):14095.
[2] Pham DT, Tran TD. Drivergene.net: a Cytoscape app for the identification of driver nodes of large-scale complex networks and case studies in discovery of drug target genes. Comput Biol Med 2024;179:108888.
[3] Ma K, et al. PPRTGI: a personalized PageRank graph neural network for TF-target gene interaction detection. IEEE ACM Trans Comput Biol Bioinform 2024;21(3):480–91.
[4] Bray F, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
[5] Nguyen T-T, et al. Exploring the molecular terrain: a survey of analytical methods for biological network analysis. Symmetry 2024;16(4).
[6] Hui TX, et al. A review of random walk-based method for the identification of disease genes and disease modules. IEEE Access 2023;11:116366–83.
[7] Cowen L, et al. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet 2017;18(9):551–62.
[8] Nguyen T, et al. WINNER: a network Biology tool for biomolecular characterization and prioritization. Frontiers Big Data 2022;5:1016606.
[9] Shang H, Liu ZP. Network-based prioritization of cancer genes by integrative ranks from multi-omics data. Comput Biol Med 2020;119:103692.
[10] The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47(D1):330–8.
[11] Jassal B, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2020;48(D1):498–503.
[12] Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci 2019;28(11):1947–51.
[13] Kanehisa M, et al. New approach for understanding genome variations in KEGG. Nucleic Acids Res 2019;47(D1):590–5.
[14] Asrina A. Precision medicine approaches in oncology: current trends and future directions. Adv Healthc Res 2024.
[15] Papadopoulou P, Lytras M. Advancing precision medicine in medical education: integrated, precise and data-driven smart solutions. Appl Res 2023.
[16] Castaneda C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinformatics 2015;5:4.
[17] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res 2017;45(D1):408–17.
[18] Jafari M, Ansari-Pour N. Why, when and how to adjust your p values? Cell J 2019;20(4):604–7.
[19] Gossmann A, et al. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. IEEE Trans Med Imaging 2018;37(8):1761–74.
[20] Jovic D, et al. Single-cell RNA sequencing technologies and applications: a brief overview. Clin Transl Med 2022;12(3):e694.
[21] Gentili M, et al. Biological random walks: multi-omics integration for disease gene prioritization. Bioinformatics 2022;38(17):4145–52.
[22] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. 1999, Stanford InfoLab.
[23] Benesty J, et al., Pearson correlation coefficient. In: Noise reduction in speech processing. Springer; 2009. pp. 1–4.
[24] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[25] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12(1):56–68.
[26] Chakravarty D, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017:2017.
[27] Suehnholz SP, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discovery 2024;14(1):49–65.
[28] Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics 2010;26(8):1057–63.
[29] Ghiassian SD, Menche J, Barabási AL. Diamond: a disease module detection algorithm derived from connectivity patterns in the human interactome. PLoS Comput Biol 2015;11(4):e1004120.
[30] Erten S, et al. DADA: Degree-aware algorithms for network-based disease gene prioritization. BioData Min 2011;4:19.
[31] Valdeolivas A, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics 2019;35(3):497–505.
[32] Rauf SM, Endou A, Takaba H, Miyamoto A. Effect of Y220C mutation on p53 and its rescue mechanism: a computer chemistry approach. Protein J 2013;32(1):68–74. https://doi.org/10.1007/s10930-012-9458-x. PMID: 23315175.

[33] Baroni TE, Wang T, Qian H, Dearth LR, Truong LN, Zeng J, et al. A global suppressor motif for p53 cancer mutants. Proc Natl Acad Sci USA 2004;101(14):4930–5. https://doi.org/10.1073/pnas.0401162101. Epub 2004 Mar 22. PMID: 15037740; PMCID: PMC387351.

[34] Couse JF, Korach KS. Estrogen receptor null mice: what have we learned and where will they lead us? Endocr Rev 1999;20(3):358–417. https://doi.org/10.1210/edrv.20.3.0370. Erratum in: *Endocr. Rev.* 1999, **20**(4): 459. PMID: 10368776.

[35] Kushner PJ, Agard DA, Greene GL, Scanlan TS, Shiau AK, Uht RM, et al. Estrogen receptor pathways to AP-1. J Steroid Biochem Mol Biol 2000;74(5):311–7. https://doi.org/10.1016/s0960-0760(00)00108-4. PMID: 11162939.

[36] Saville B, Wormke M, Wang F, Nguyen T, Enmark E, Kuiper G, et al. Ligand-, cell-, and estrogen receptor subtype (alpha/beta)-dependent activation at GC-rich (Sp1) promoter elements. J Biol Chem 2000;275(8):5379–87. https://doi.org/10.1074/jbc.275.8.5379. PMID: 10681512.

[37] Thomas C, Gustafsson JÅ. The different roles of ER subtypes in cancer Biology and therapy. Nat Rev Cancer 2011;11(8):597–608. https://doi.org/10.1038/nrc3093. PMID: 21779010.

[38] Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, et al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 2007;448(7152):439–44. https://doi.org/10.1038/nature05933. Epub 2007 Jul 4. PMID: 17611497.

[39] Malanga D, Scrima M, De Marco C, Fabiani F, De Rosa N, De Gisi S, et al. Activating E17K mutation in the gene encoding the protein kinase AKT1 in a subset of squamous cell carcinoma of the lung. Cell Cycle 2008;7(5):665–9. https://doi.org/10.4161/cc.7.5.5485. Epub 2007 Dec 26. PMID: 18256540.

[40] Aoki M, Batista O, Bellacosa A, Tsichlis P, Vogt PK. The Akt kinase: molecular determinants of oncogenicity. Proc Natl Acad Sci USA 1998;95(25):14950–5. https://doi.org/10.1073/pnas.95.25.14950. PMID: 9843996; PMCID: PMC24556.

[41] Parikh C, Janakiraman V, Wu WI, Foo CK, Kljavin NM, Chaudhuri S, et al. Disruption of PH-kinase domain interactions leads to oncogenic activation of AKT in human cancers. Proc Natl Acad Sci USA 2012;109(47):19368–73. https://doi.org/10.1073/pnas.1204384109. Epub 2012 Nov 7. PMID: 23134728; PMCID: PMC3511101.

[42] Guo G, Qiu X, Wang S, Chen Y, Rothman PB, Wang Z, et al. Oncogenic E17K mutation in the pleckstrin homology domain of AKT1 promotes v-abl-mediated pre-B-cell transformation and survival of pim-deficient cells.. Oncogene 2010;29(26):3845–53. https://doi.org/10.1038/onc.2010.149. Epub [2010 May 3]. PMID: 20440266.

[43] Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med 2011;365(7):611–9. https://doi.org/10.1056/NEJMoa1104017. Epub 2011 Jul 27. PMID: 21793738; PMCID: PMC3170413.

[44] Hyman DM, Smyth LM, Donoghue MTA, Westin SN, Bedard PL, Dean EJ, et al. AKT inhibition in solid tumors with AKT1 mutations. J Clin Oncol 2017;35(20):2251–9. https://doi.org/10.1200/JCO.2017.73.0143. Epub 2017 May 10. Erratum in: J Clin Oncol. 2019 Feb 1;37(4):360. doi: 10.1200/JCO.18.02209. PMID: 28489509; PMCID: PMC5501365.

[45] Ye S, Hu X, Ni C, Jin W, Xu Y, Chang L, et al. KLF4 p.A472D mutation contributes to acquired resistance to cetuximab in colorectal cancer. Mol Cancer Ther 2020;19(3):956–65. https://doi.org/10.1158/1535-7163.MCT-18-1385. Epub 2020 Jan 10. PMID: 31924740.

[46] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 2004;350(21):2129–39. https://doi.org/10.1056/NEJMoa040938. Epub 2004 Apr 29. PMID: 15118073.

[47] Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304(5676):1497–500. https://doi.org/10.1126/science.1099314. Epub 2004 Apr 29. PMID: 15118125.

[48] Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib.. Proc Natl Acad Sci USA 2004;101(36):13306–11. https://doi.org/10.1073/pnas.0405220101. Epub 2004 Aug 25. PMID: 15329413; PMCID: PMC516528.