**Prioritizing Cancer Therapeutic Genes Using BioRank: A Biologically-Informed PageRank Framework**

Duc-Tinh Pham1\*, Huu-Tam Nguyen2, Van-Hai Pham2

1School of Information and Communications Technology, Hanoi University of Industry, 298 Cau Dien Street, Bac Tu Liem District, Hanoi, Vietnam.

2School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi, Vietnam.

\*Corresponding: Duc-Tinh Pham (tinhpd@haui.edu.vn)

To assess the performance of the gene prioritization algorithms, we employ two widely used metrics in information retrieval and network-based gene ranking: Recall@*K* and normalized Discounted Cumulative Gain (nDCG@*K*).

Recall@*K* quantifies the proportion of relevant (i.e., known disease-associated) genes that are successfully retrieved within the top *K* predictions generated by the algorithm. Given a ranked list of candidate genes *R={g1,g2,…,gK}* and a ground truth set of disease-associated genes GGG, Recall@*K* is defined as:

This metric evaluates the coverage of true disease genes among the top *K* ranked results, regardless of their specific positions in the list. Recall@*K* is particularly relevant when downstream applications, such as wet-lab validation or drug target selection, are limited to a small number of top candidate genes.

Normalized Discounted Cumulative Gain (nDCG@*K*) evaluates both the presence and the ranking positions of relevant genes. It reflects the intuition that identifying relevant genes at higher ranks is more valuable than identifying them further down the list. Formally, Discounted Cumulative Gain at rank *K* (DCG@*K*) is computed as:

where indicates whether the *i-th* gene in the ranked list *R* belongs to the ground truth set *G*. The ideal DCG (IDCG@K) corresponds to the DCG of a perfectly ranked list, where all relevant genes are placed at the top. nDCG@K is then calculated as:

nDCG@K yields values in the range [0, 1], where a score of 1 denotes a perfect ranking. This metric is particularly well-suited for prioritization tasks in which it is essential to rank relevant genes as highly as possible, even if not all relevant genes are retrieved within the top *K* positions.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table S2.** Performance Comparison of PageRank, BRW, and BioRank on the top 100 genes across Cancer Datasets | | | | | | | | | | | |
| **Biomolecular network** | **Properties** | | **PageRank** | | | **BRW** | | | **BioRank** | | |
| **Nodes** | **Edges** | **Match** | **Recall@100** | **nDCG@100** | **Match** | **Recall@100** | **nDCG@100** | **Match** | **Recall@100** | **nDCG@100** |
| **BRCA** | 12148 | 219166 | 32 | 0.0273 | 0.3412 | 56 | 0.0477 | 0.5131 | **66** | 0.0562 | **0.6999** |
| **COAD** | 12148 | 799078 | 27 | 0.0230 | 0.2970 | 36 | 0.0307 | 0.3250 | **46** | 0.0392 | **0.5115** |
| **LUAD** | 12148 | 337686 | 27 | 0.0230 | 0.2999 | 32 | 0.0273 | 0.2975 | **58** | 0.0494 | **0.6242** |
| **THCA** | 12148 | 547306 | 32 | 0.0273 | 0.3413 | 35 | 0.0298 | 0.3310 | **55** | 0.0468 | **0.5467** |
| **BLCA** | 12148 | 237288 | 32 | 0.0273 | 0.3391 | 61 | 0.0520 | 0.6287 | **69** | 0.0588 | **0.7208** |
| **PRAD** | 12148 | 607492 | 34 | 0.0290 | 0.3576 | 59 | 0.0503 | 0.5554 | **57** | 0.0486 | **0.6032** |
| **STAD** | 12148 | 271464 | 32 | 0.0273 | 0.3359 | 70 | 0.0596 | 0.7164 | **65** | 0.0554 | **0.6936** |

Table S2 presents a comprehensive comparison of three gene prioritization methods PageRank, BRW, and BioRank based on three performance metrics: Match (number of validated therapeutic targets), Recall@100, and nDCG@100 across seven cancer datasets. This table serves as a critical component in demonstrating the effectiveness and robustness of the proposed BioRank method. The table includes all necessary components for comparison: number of nodes and edges in the PPI network, and three performance indicators. This level of detail provides transparency regarding network scale and performance dynamics. BioRank achieves the highest scores across all datasets in both Recall@100 and nDCG@100, confirming its superior capacity to retrieve more validated genes and rank them more accurately compared to the baselines. For example, in the LUAD dataset, BioRank increases the match count by over 2x compared to PageRank (58 vs. 27), and boosts nDCG@100 from 0.2999 (PageRank) to 0.6242. The improvement in nDCG@100 is particularly meaningful, as it reflects the ability of BioRank to push highly relevant genes toward the top of the ranked list critical for downstream experimental validation.