**Received:** April 20, 2023; **Accepted:** August 19, 2023.
DOI. 10.30473/coam.2023.67660.1231
Summer-Autumn (2023) Vol. 8, No. 2, (1-15)
**Research Article**

Open Access

**Control and Optimization in Applied Mathematics - COAM**

# Graph Feature Selection for Anti-Cancer Plant Recommendation

**Mahmood Amintoosi**[1*], **Eisa Kohan-Baghkheirati**[2]

[1]Department of Computer Science, Hakim Sabzevari University, Sabzevar, Iran.
[2]Department of Biology, Hakim Sabzevari University, Sabzevar, Iran.

**Correspondence:**
Mahmood Amintoosi
**E-mail:**
m.amintoosi@hsu.ac.ir

**How to Cite**
Amintoosi, M., Kohan-Baghkheirati, E. (2023). “Graph feature selection for anti-cancer plant recommendation”, Control and Optimization in Applied Mathematics, 8(2): 1-15.

**Abstract.** Every year, extensive experimental analysis is conducted to evaluate the anti-cancer properties of plants. Developing a well-ranked list of potential anti-cancer plants based on verified anti-cancer metabolites can significantly reduce the time and cost required for plant evaluation. This paper proposes a method for generating such a ranked list by analyzing biological graphs of plant-metabolite interactions. In this approach, graph nodes are ranked based on specific graph features. However, a challenge arises in selecting the most informative graph features that ensure the resulting ranked plant list is more relevant, prioritizing plants with greater anti-cancer properties at the top. To address this challenge, we propose the use of the Average Precision metric commonly used in information retrieval and recommender systems, to compare different ranked lists. By constructing a network that captures the similarities between plants based on their shared metabolites, and ranking plants using different combinations of graph features, we can identify the subset of features that yields a ranked list with a higher Average Precision score. This subset of features can then be considered the most suitable for recommending anti-cancer plants. The proposed method can be used to select the best graph features for screening unverified plant lists for anti-cancer properties, increasing the likelihood of identifying plants with higher scores in the list that possess anti-cancer properties.

**Keywords.** Anti-cancer plant recommendation, Graph feature, Recommender systems, Medicinal plants, Herbal medicine, Breast cancer, Stomach cancer, Gastric cancer, Gastric neoplasms.

**MSC.** 05C90; 92C80.

## 1 Introduction

Graph theory finds applications in various fields such as mathematics, computer sciences, engineering, and biology. Network analysis is a common technique used to study biological graphs, which aids in tasks like drug target identification, protein or gene function prediction, and more [19]. Node Ranking is a specific type of network analysis that ranks the vertices of a graph based on certain graph and node features. Examples of these features include Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, and others. Previous studies have employed different combinations of these features for network analysis [2, 9, 15, 16]. However, there is currently no systematic approach to selecting the best subset of features for a given problem, which is the main focus of this paper. We apply our method to an anti-cancer plant recommendation system.

Cancer presents a significant global health challenge, with 19.3 million new cases and 10 million related deaths reported in 2020 [25]. Numerous studies, such as those by [10, 22], are dedicated to cancer detection. Medicinal plants hold promise as potential sources for cancer prevention and treatment. Several studies have demonstrated the therapeutic effects of plants on various types of cancer [12].

The anti-cancer effects of plants largely depend on their secondary metabolites, which are compounds not essential for plant growth and development but play a role in plant responses to environmental interactions [11, 18]. These metabolites can prevent or treat cancer through various mechanisms, such as enzyme inhibition, regulation of signaling-metabolic pathways, and inducing anti-oxidant and anti-inflammatory activities [6, 11, 12]. Hence, it is crucial to develop an anti-cancer plant recommendation method that considers the composition of these metabolites and ranks plants based on their anti-cancer potential. In ranking approaches, a graph is constructed based on the plants and metabolites, and the nodes of the graph are ordered based on specific graph features. The top-ranked plants are potential candidates for further investigation of their anti-cancer properties. An important question regarding this approach is *"What is the best subset of graph features for a specific task?"*. In this paper, we propose a method for selecting the most suitable graph features for anti-cancer plant recommendation. We employ an exhaustive search method to identify the best graph features in the context of anti-cancer plant recommendation. The objective function utilized for selecting the best features is a metric commonly employed in information retrieval. This metric involves the comparison of two distinct ranked lists.

The remainder of this paper is organized as follows. We begin by reviewing some necessary background information, followed by a description of the proposed method as an algorithm. Subsequently, we present the experimental results obtained from several cancer datasets.

## 2 Preliminaries

This section provides the necessary background information for understanding the proposed method, such as evaluation metrics in recommender systems, Average Precision, and Average Precision at $k$.

### 2.1 Evaluation metrics in classification and recommender systems

A confusion matrix is a way to represent the results of a classifier, which is a two-dimensional table with a row and column for each class. Each element of the matrix shows the number of test examples that are classified or predicted by a learned model. The diagonal elements of the matrix represent correctly classified instances and the off-diagonal elements indicate incorrectly classified instances [21]. A recommender system can be seen as a binary classification problem, where the task is to decide whether each item should be recommended or not; in other words, whether it is relevant or not. When the model predicts a relevant item as relevant, it is considered as a True Positive. The possible situations are as follows:

- True Positive ($TP$): It means a relevant item is predicted as relevant
- True Negative ($TN$): It means a non-relevant item is predicted as not relevant
- False Positive ($FP$): It means a non-relevant item is predicted as relevant
- False Negative ($FN$): It means a relevant item is predicted as not relevant.

Based on these terms, some classification criteria such as Accuracy and Precision are defined. Accuracy is the fraction of correct predictions and is defined by

$$\frac{(TP+TN)}{(TP+TN+FP+FN)},$$

and precision is the fraction of retrieved items that are relevant, which is defined by

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = \frac{TP}{TP+FP} = P(\text{relevant}|\text{retrieved}). \tag{1}$$

In other words, precision can provide answers to the question: How many of the items that the model classified as positive were actually positive? The precision formula (1) is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP). True positives are the cases where the model correctly predicted a positive instance, while false positives are the cases where the model incorrectly predicted an instance as positive when it was actually negative. A high precision score indicates that the model is making accurate predictions for the positive class, while a low precision score indicates that the model is making too many false positive errors.

### 2.2 Objective function for measuring the ranked list of plants

Precision is suitable for evaluating the performance of binary retrieval systems, but it cannot assess rankings. Most modern information retrieval (IR) systems generate ranked results. In this paper, which is focuses on recommending a list of anti-cancer plants, the situation is similar to a search engine. An ideal search engine ranks all relevant documents (items) before non-relevant ones. Evaluation metrics should consider the ranks of relevant documents, and in this case, we need a metric that incorporates the order of the produced plant list. By default, precision takes into account all the retrieved items. However, it can also be evaluated at a specific number of retrieved items, commonly known as the cut-off rank. In this evaluation, the model is assessed by considering only its top-most queries. This measure is called precision at $k$ or $P@k$. In some documents, it is denoted as $P@r$, corresponding to precision at rank $r$. Based on this criterion, two other metrics were introduced to measure the quality of a recommender system: Average precision and average precision at $k$; which will be described shortly.

#### 2.2.1 Average precision

The average precision [26] is the mean of the precision scores after each relevant document is retrieved:

$$\text{Average Precision} = \frac{\sum_r P@r}{R}. \tag{2}$$

Here, $r$ represents the rank of each relevant document, $R$ is the total number of relevant documents, and $P@r$ is the precision of the top-$r$ retrieved items. Average Precision ($AP$) is a metric that evaluates whether all of the ground-truth relevant items selected by the model are ranked higher or not.

#### 2.2.2 Average precision at $k$ ($AP@k$)

Average Precision at $k$ [7] is a variant of Average Precision ($AP$) that considers only the top $k$ ranked documents. While $AP$ is already biased towards the top of the ranking, it also incorporates a recall component by normalizing according to $R$, the number of relevant documents for a query. $AP@k$ takes into account both the number of relevant documents in the top $k$ and their positions, and unlike Precision at $r$ ($P@r$), which disregards position. It is defined as (3):

$$AP@k = \frac{\sum_{r=1}^{k} P@r \times \text{rel}(r)}{R}, \tag{3}$$

where $\text{rel}(r) = 1$ if the $r$-th retrieved document is relevant and $\text{rel}(r) = 0$ otherwise. In Average Precision, it is customary to normalize by the number of relevant documents, i.e., $R$.

To illustrate these metrics, consider the following example.

**Example 1.** Suppose we have two different IR systems, Model 1 and Model 2, and their rankings of documents for a given input query $q$. The collection contains six documents, where odd documents ($d_1$, $d_3$, $d_5$) are relevant to the query $q$, and even documents ($d_2$, $d_4$ and $d_6$) are not relevant to $q$. If the ranked output list of Model 1 is [$d_1$, $d_3$, $d_5$, $d_2$, $d_6$], and the output of Model 2 is [$d_1$, $d_2$, $d_3$, $d_4$, $d_5$], both methods retrieve the relevant documents $d_1$, $d_3$ and $d_5$. However, since these items appear at the top of recommended items of Model 1, a good measure would assign a higher score to Model 1 compared to Model 2. Table 1 illustrates this situation, where the last row shows that $AP@5$ for Model 1 is equal to 1, which is higher than the score of 0.756 for Model 2, as expected. It is evident that removing each relevant document from Model 2 decreases its score.

**Table 1:** Computation of $AP@k$ for Example 1, comparing two IR models. The relevant documents are highlighted in bold letters

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| rank ($r$) | Retrieved Items | $P@r$ × rel($r$) | Retrieved Items | $P@r$ × rel($r$) |
| 1 | **d₁** | 1/1×1 | **d₁** | 1/1×1 |
| 2 | **d₃** | 2/2×1 | $d_2$ | 1/2×0 |
| 3 | **d₅** | 3/3×1 | **d₃** | 2/3×1 |
| 4 | $d_2$ | 3/4×0 | $d_4$ | 2/4×0 |
| 5 | $d_6$ | 3/5×0 | **d₅** | 3/5×1 |
| $AP@5$ | | 1/3(1+1+1+0+0)<br>=1 | | 1/3(1+0+2/3+ 0+3/5)<br>= 0.756 |

## 3 The Proposed Algorithm

The proposed algorithm aims to identify the best graph features for ranking plants based on their anti-cancer potential. In this algorithm, we consider a graph where the nodes represent plants and the edges represent shared metabolites between plants. Based on the assumption that plants containing a greater number of common anti-cancer metabolites are more likely to possess higher anti-cancer potential, the nodes within the graph are ranked using various graph features, including Degree, Closeness, Betweenness, and Eigenvector. These features serve to quantify the significance of the nodes within the graph. Different combinations of graph features yield distinct rankings of plants. The proposed algorithm employs a global search to determine the best subset of features.

The algorithm is illustrated in Algorithm 1. Let $A$ be a given list of verified anti-cancer metabolites (along with the corresponding plants), and $B$ be a list of verified anti-cancer plants (and their known metabolites). The plants in list $A$ may or may not possess verified anti-cancer properties. The objective is to prioritize the ranking of these plants based on their higher likelihood of possessing anti-cancer properties. For each subset of graph features, a ranking of plants is generated. The ranking that closely aligns with list $B$ (serving as the ground truth list) is expected to indicate higher anti-cancer potential. The quality of the ranking is evaluated using $AP@k$. The subset of graph features that produces the best ranking is reported as the best graph feature for the anti-cancer plant recommendation system. Plants at the top of the best ranking are anticipated to possess higher anti-cancer potential compared to those at the bottom of the ranking. Thus, they are considered better candidates for further experimental research.

**Algorithm 1** Finding the best graph features for anti-cancer plant recommendation.

```
Input: $A$, A list of anti-cancer metabolites and the corresponding plants,
    $B$, A list of anti-cancer plants and their metabolites,
    $S$, The set of graph features,
Output: The best graph feature for plant recommendations.
Create a graph using the plants from list $A$, as nodes (vertices).
Find the largest sub-graph in the above graph and name it $G$.
Assign edge weights to $G$ based on the number of common metabolites of two vertices (plants).
For each metabolite of each edge, if the metabolite also exists in list $B$, increase the weight of corresponding edge.
Compute all $2^{|S|} - 1$ non-empty subsets of $S$.
for each subset, do the following do
    Compute the features of $G$.
    Rank the nodes (plants) of $G$ based on the sum of their features
    Use the recommender metric $AP@k$ to evaluate the quality of the ranked list, considering the plants from list $B$ as the Truth list.
    Compute the aforementioned measure for various values of $k$, and save the average.
end for
Return the subset with a maximum average of $AP@k$.
```

## 4 Results

As previously stated, we require two distinct lists: A list of anti-cancer plants, and a list of anti-cancer metabolites. In this section, we provide a more detailed description of these lists as our datasets and present the results of the proposed algorithm on these datasets.

The data and the code used in this paper are available at http://github.com/mamintoosi/FS-in-Bio-Graphs

### 4.1 Dataset analysis

We utilized several datasets to demonstrate the effectiveness of the proposed method. We obtained the anti-cancer metabolites from PubChem and KNApSAcK databases, which we will refer to as AC (Anti-Cancer) in the following. First, we briefly describe this dataset, and then integrate it with the known anti-cancer plants for Stomach and Breast Cancers. For each dataset, we provide some statistics about the plants and metabolites involved.

#### 4.1.1 Anti-cancer (AC) metabolites dataset

Here, we obtained the list of anti-cancer plants from PubMed and Google Scholar databases using keywords such as plant extract, plant metabolites, plant secondary metabolites, stomach cancer, and cancer-related cell lines. We then collected and stored separately the articles that confirmed the anti-cancer effects of plants. All plant anti-cancer metabolites were also extracted from plant metabolite databases such as PubChem [13] and KNApSAcK [23]. Additionally, the KNApSAcK Family database was used to obtain the plant-metabolite associations. We will use the term AC, refer to the Anti-Cancer metabolites that are collected from PubChem and KNApSAcK databases. Table 2 shows some statistics of this dataset. We have 667 unique metabolites and 3251 unique plants containing some of the anti-cancer metabolites. Each metabolite occurs in 1 to 244 plants, with an average of 8. C00002526 ($C_{15}H_{10}O_5$, Genistein, or 4’,5,7-Trihydroxyisoflavone) is the most frequent anti-cancer metabolite occurring in 244 plants [14, 17, 24], and *Annona muricata* is the most frequent plants. We hypothesized that those plants having more common metabolites to AC, should have more anti-cancer properties, as literature confirmed for *Annona muricata* [4, 5, 8].

**Table 2:** Description of the dataset on AC metabolites

| | |
|---|---|
| Number of unique metabolites: | 667 |
| Min number of replicates of a metabolite in plants: | 1 |
| Max number of replicates of a metabolite in plants: | 244 |
| Avg number of replicates of a metabolite in plants: | 8 |
| Number of unique plants: | 3251 |
| Min Number of plants having each metabolite: | 1 |
| Max Number of plants having each metabolite: | 14 |
| Avg Number of plants having each metabolite: | 2 |

Not all plants in AC have verified anti-cancer properties. To select one plant that has not been verified yet for further investigation, we can rank the plants in AC based on their anti-cancer potential and choose the first plant that has not been verified yet. Ranking the plants can be done by building a graph over these plants, and sorting them according to graph features. Selecting the best features is done by the proposed algorithm (Algorithm 1). The best features may vary for different cancers, here, we used two anti-cancer plant datasets: Stomach and Breast cancer. Each of these datasets is used separately as a list $B$ in Algorithm 1.

### 4.1.2 Stomach dataset

The stomach dataset comprises a list of plants with verified anti-cancer properties, based on experimental studies and published data from scientific articles. According to the dataset, C00000615 (caffeic acid, $C_9H_8O_4$) and C00000805 (alpha-pinene, $C_{10}H_{16}$) are the most common metabolites found in 33 stomach anti-cancer plants. Our hypothesis is that the metabolites that occur more frequently in stomach anti-cancer plants have higher anti-cancer potential; This hypothesis is supported by a literature review [3]. Table 3 presents some statistics of this dataset. There are 272 shared metabolites between AC and stomach anti-cancer plants. As previously mentioned, the anti-cancer potential of plants depends on their metabolites. Therefore, the presence of shared anti-cancer metabolites is expected.

**Table 3:** Stomach anti-cancer plants dataset description

| | |
|---|---|
| Number of unique metabolites: | 7443 |
| Min number of replicates of a metabolite in plants: | 1 |
| Max number of replicates of a metabolite in plants: | 33 |
| Avg number of replicates of a metabolite in plants: | 2 |
| Number of unique plants: | 367 |
| Min number of plants having each metabolite: | 1 |
| Max number of plants having each metabolite: | 241 |
| Avg number of plants having each metabolite: | 33 |

### 4.1.3 Breast dataset

The breast anti-cancer plants dataset is another dataset that we will analyze in the following (Table 4). It comprises a list of 403 plants with verified anti-breast cancer properties, based on experimental studies and published data. According to our previous hypothesis, the top-ranked metabolite is C00003672 (beta-sitosterol, $C_{29}H_{50}O$), which according to our previous hypothesis, has confirmed anti-cancer potential [20, 1].

**Table 4:** Breast anti-cancer plants dataset description

| | |
|---|---|
| Number of unique metabolites: | 7416 |
| Min number of replicates of a metabolite in plants: | 1 |
| Max number of replicates of a metabolite in plants: | 61 |
| Avg number of replicates of a metabolite in plants: | 2 |
| Number of unique plants: | 403 |
| Min number of plants having each metabolite: | 1 |
| Max number of plants having each metabolite: | 627 |
| Avg number of plants having each metabolite: | 31 |

### 4.2 Plants' graph construction

As explained before, recommending a good anti-cancer plant candidate with a high probability of having anti-cancer properties requires determining the best features of the graph for node ranking. Not all of the 3251 plants that contain AC metabolites have verified anti-cancer properties. Therefore, we use the other two anti-cancer plant datasets to rank the AC plants based on the features of the graph constructed using AC dataset. Here, we build a graph over AC plants, where metabolites represent the edges between plants. Then, the graph nodes (plants) are sorted according to some graph features. The resulting graph has 3251 nodes, which is equal to the number of plants in AC dataset. Since many graph features are defined on connected graphs, we select the largest connected sub-graph of this graph for further analysis. The largest sub-graph of AC plants has 3050 nodes.

### 4.3 Graph features

Among the various graph features, we have selected the following features: *degree, degree cent, betweenness, closeness, eccentricity, and eigenvector*. Our goal is to select the best features for plant recommendation. We use the combinatorial approach described in Algorithm 1: exhaustive search for choosing the best subset of features. Since we have six features, the total number of all non-empty subsets is $2^6 - 1 = 63$ (Table 5). With a global search, we will choose the best subset. As mentioned in the previous section, in this paper, we use $AP@k$ as an objective function to measure the performance of each subset. After creating the sub-graph with each subset, the graph nodes (plants) are ranked based on the features in the subset. We will demonstrate the proposed method on Stomach and Breast anti-cancer plant datasets as case studies.

**Table 5:** Features subsets for 6 graph features used in node ranking

| | |
|---|---|
| 1 | {degree} |
| 2 | {degree_cent} |
| 3 | {betweenness} |
| 4 | {closeness} |
| 5 | {eccentricity} |
| 6 | {eigenvector} |
| 7 | {degree, degree-cent} |
| 8 | {degree,betweenness} |
| 9 | {degree, closeness} |
| 10 | {degree, eccentricity} |
| 11 | {degree,eigenvector} |
| 12 | {degree_cent, betweenness} |
| ⋮ | ⋮ |
| 61 | {degree, betweenness, closeness, eccentricity, eigenvector} |
| 62 | {degree_cent, betweenness, closeness, eccentricity, eigenvector} |
| 63 | {degree, degree_cent, betweenness, closeness, eccentricity, eigenvector} |

### 4.4 Results on stomach anti-cancer plants dataset

The largest connected sub-graph of AC consists of 3050 plants. For each of the $2^6 - 1$ subsets, and for each $k \in [1, 2, \dots, 9]$, the average precision at $k$ ($AP@k$) is computed. The subset with the best average $AP@k$ can be considered as the best subset of graph features. Figure 1 shows the result of running Algorithm 1 on AC and Stomach datasets, represented as list $A$ and list $B$ in the Algorithm. The subset with three features: *degree, eccentricity, and eigenvector* demonstrates the best average scores among all other subsets, including the complete set of six features. In this dataset, *betweenness* exhibits the lowest score. The best and the worst subsets are highlighted respectively by green and red letters in Figure 1.

### 4.5 Results on the breast anti-cancer plants dataset

Figure 2 illustrates the result obtained by running Algorithm 1 on the AC and breast anti-cancer plants datasets, using list A and list B as inputs, respectively. The subset with four features: *degree, degree centrality, betweenness, and eccentricity*, achieves the best average scores among all other subsets. In contrast to the previous dataset, the score of the entire set of six features is

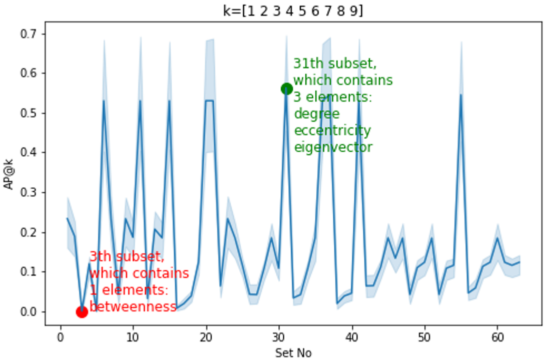

**Figure 1:** Stomach anti-cancer plants dataset: $AP@k$ results for various values of $k$ measuring the performance of the ranked plant list based on the subset of graph features. The subset $\{degree, eccentricity, eigenvector\}$ represents the best subset of features (highlighted in green) and $\{betweenness\}$ represents the worst set (highlighted in red).

high, closely approaching the performance of the best subset. Notably, *eccentricity* yields the worst result.

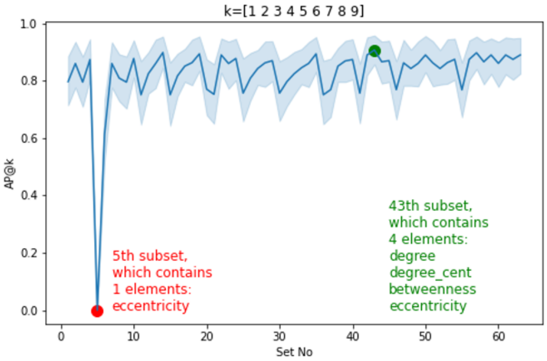

**Figure 2:** Breast anti-cancer plants dataset: $AP@k$ results for various values of $k$, measuring the performance of the ranked plant list based on the subset of graph features. The best subset of features $\{degree, degreecentrality, betweenness, eccentricity\}$ is highlighted in green letters, while the worst subset $\{eccentricity\}$ is highlighted in red letters.

Table 6 displays the top 12 recommended plants using the best features for the mentioned datasets. Similar to Example 1, the items present in the Truth list are highlighted in bold letters. Since the order of "hits" and "misses" significantly impacts the $AP$ score, these highly ranked lists are expected to be more suitable than others. Moreover, Table 6 includes the ranked lists of the worst features for each dataset, allowing for comparison. For instance, in the breast cancer dataset's best section (column Breast), 4 out of 5 top recommended plants are relevant (present in the list of plants with verified breast anti-cancer properties). However, in the worst section, none of the first 10 items are relevant. As observed in Table 6 and as expected, the plants recommended by the best features exhibit a significant overlap with the list of anti-cancer plants. The plants highlighted in bold letters have confirmed anti-cancer properties, while the other plants are good candidates for investigating their anti-cancer properties.

**Table 6:** The first 12 recommended plants using the best or the worst features for the mentioned datasets. The plants highlighted in bold letters have proven anti-cancer properties. The best and the worst feature sets are shown in Figures 1 and 2, respectively.

| Stomach Dataset | | Breast Dataset | |
|---|---|---|---|
| Plants Recommended by the: | | Plants Recommended by the: | |
| Best Features | Worst Features | Best Features | Worst Features |
| **Trifolium pratense** | Phellodendron amurense | **Arabidopsis thaliana** | Opuntia dellenii |
| Arabidopsis thaliana | Arabidopsis thaliana | **Phellodendron amurense** | Tsuga heterophylla |
| Prunus cerasus | Glehnia littoralis | **Trifolium pratense** | Phellodendron japonicum |
| **Punica granatum** | Ziziphus jujuba | Glehnia littoralis | Ferula foetida |
| Glycine max | Galanthus caucasicus | **Punica granatum** | Ferula assafoetida |
| Lespedeza bicolor | Nelumbo nucifera | Salvia officinalis | Ferula assa-foetida |
| Viscum coloratum | Morus alba | Lespedeza bicolor | Angelica gigas |
| Crataegus pinnatifida | Juniperus thurifera | Viscum coloratum | Falcaria vulgaris |
| Medicago sativa | Juniperus phoenicea | Diospyros kaki | Aeschynanthus bracteatus |
| Glehnia littoralis | **Punica granatum** | **Crataegus pinnatifida** | Ligusticum jeholense |
| Cytisus scoparius | **Salvia officinalis** | Phyllanthus emblica | **Phellodendron amurense** |
| Glycyrrhiza glabra | Narcissus tazetta | **Ziziphus jujuba** | Phebalium clavatum |

## 5 Conclusion

Graphs, as the fundamental tool for network representation, have gained significant attention in various fields, including biology. Graph features, such as node degree and centrality, play a crucial role in ranking nodes. This paper focuses on the selection of the best subset of graph features using a global search approach. One of the challenges in selecting the optimal features for bio-graphs is finding an effective objective function to compare recommended lists associated with different feature subsets. In this study, we employed Average Precision at $k$ ($AP@k$) as a metric for comparing ranked lists. The paper provides comprehensive explanations of the proposed algorithm and demonstrates its application to datasets involving anti-cancer plants and metabolites. By constructing a graph based on two lists of known anti-cancer plants and metabolites, we evaluated the ranked lists using the $AP@k$ metric. The proposed ranked list includes plants that have not been previously investigated for their anti-cancer properties. These plants possess potential for further experimentation, which can save time and effort by avoiding the evaluation of unrelated plants.

## Declarations

### Availability of supporting data

All data generated or analyzed during this study are included in this published paper.

### Funding

This study received no funds, grants, or other financial support.

### Competing interests

The authors declare no competing interests are relevant to the content of this paper.

### Authors' contributions

The main manuscript text is collectively written by all authors.

### Acknowledgements

We would like to appreciate Ms. S. Asadi and Ms. Z. Samadi for preparing dataset of anti-cancer plants.

### References

[1] Abeesh, P., Guruvayoorappan, C. (2022). "Preparation and characterization of beta sitosterol encapsulated nanoliposomal formulation for improved delivery to cancer cells and evaluation of its anti-tumor activities against Daltons Lymphoma Ascites tumor models", Journal of Drug Delivery Science and Technology, 70, 102832.

[2] Ahmed, M.M., Tazyeen, S., Ali, R., Alam, A., Imam, N., Malik, M.Z., Ali, S., Ishrat, R. (2022). "Network centrality approaches used to uncover and classify most influential nodes with their related miRNAs in cardiovascular diseases", Gene Reports, 27, 101555.

[3] Allenspach, M., Steuer, C. (2021). "$\alpha$-Pinene: A never-ending story", Phytochemistry, 190, 112857.

[4] Amadea, I.C., Atrasina, D. (2021). "The effects of simvastatin and soursop (annona muricata) leaf extract on colorectal cancer", Indonesian Journal of Life Sciences| ISSN, 3(1), 1-172656-0682.

[5] Behl, S., Inbanathan, A., Sundaram, M.K., Hussain, A. (2022). "Plants of the genus Annona: Source of potential anti-cancer therapeutics", In Functional Foods and Nutraceuticals in Metabolic and Non-Communicable Diseases, 741-753, Elsevier.

[6] Bouyahya, A., El Allam, A., Zeouk, I., Taha, D., Zengin, G., Goh, B.H., Catauro, M., Montesano, D., El Omari, N. (2022). "Pharmacological effects of Grifolin: Focusing on anticancer mechanisms", Molecules, 27(1), 284.

[7] Craswell, N., Robertson, S. (2009). “Average Precision at $n$”, In Liu, L. and ÖZsu, M.T., editors, Encyclopedia of Database Systems, Springer US, Boston, MA, 193-194.

[8] El-Beltagy, A. E.-F.B., Elsyyad, H.I., Abdelaziz, K.K., Madany, A.S., Elghazaly, M.M. (2021). “Therapeutic role of Annona muricata fruit and bee Venom against MNU-induced breast cancer in pregnant rats and its complications on the ovaries”, Breast Cancer: Targets and Therapy, 13, 431.

[9] Gopalakrishnan, S., Sridharan, S., Nayak, S.R., Nayak, J., Venkataraman, S. (2022). “Central hubs prediction for bio networks by directed hypergraph-GA with validation to COVID-19 PPI”, Pattern Recognition Letters, 153, 246-253.

[10] Hasanzadeh, M., Movahedi, M., Rejali, M., Maleki, F., Moetamani-Ahmadi, M., Seifi, S., Hosseini, Z., Khazaei, M., Amerizadeh, F., Ferns, G.A., Rezayi, M., Avan, A. (2019). “The potential prognostic and therapeutic application of tissue and circulating microRNAs in cervical cancer”, J. Cell Physiol, 234(2), 1289-1294.

[11] Kaur, J., Mahey, S., Ahluwalia, P., Joshi, R., Kumar, R. (2022). “Role of plant secondary metabolites as anticancer and chemopreventive agents”, In Plant Secondary Metabolites, 97-119, Springer.

[12] Khan, M.I., Bouyahya, A., Hachlafi, N.E., Menyiy, N.E., Akram, M., Sultana, S., Zengin, G., Ponomareva, L., Shariati, M.A., Ojo, O.A. (2022). “Anticancer properties of medicinal plants and their bioactive compounds against breast cancer: a review on recent investigations”, Environmental Science and Pollution Research, 1–34.

[13] Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., and Shoemaker, B.A. (2016). “PubChem substance and compound databases”, Nucleic Acids Research, 44(D1), D1202-D1213.

[14] Ma, X., Yu, X., Min, J., Chen, X., Liu, R., Cui, X., Cheng, J., Xie, M., Diel, P., Hu, X. (2022a). “Genistein interferes with antitumor effects of cisplatin in an ovariectomized breast cancer xenograft tumor model”, Toxicology Letters, 355, 106-115.

[15] Ma, Y., Guo, J., Li, D., Cai, X. (2022b). “Identification of potential key genes and functional role of CENPF in osteosarcoma using bioinformatics and experimental analysis”, Experimental and Therapeutic Medicine, 23(1), 1-12.

[16] Naseri, A., Sharghi, M., Hasheminejad, S.M.H. (2021). “Enhancing gene regulatory networks inference through hub-based data integration”, Computational Biology and Chemistry, 95, 107589.

[17] Naujokat, C., McKee, D.L. (2021). “The “Big Five” phytochemicals targeting cancer stem cells: curcumin, EGCG, sulforaphane, resveratrol and genistein”, Current Medicinal Chemistry, 28(22), 4321-4342.

[18] Pagare, S., Bhatia, M., Tripathi, N., Pagare, S., Bansal, Y.K. (2015). “Secondary metabolites of plants and their role: Overview”, Current Trends in Biotechnology and Pharmacy, 9(3), 293-304.

[19] Pavlopoulos, G.A., Hooper, S.D., Sifrim, A., Schneider, R., Aerts, J. (2011). “Medusa: A tool for exploring and clustering biological networks”, BMC Research Notes, 4(1), 1-6.

[20] Rajavel, T., Mohankumar, R., Archunan, G., Ruckmani, K., Devi, K.P. (2017). "Beta sitosterol and Daucosterol (phytosterols identified in Grewia tiliaefolia) perturbs cell cycle and induces apoptotic cell death in A549 cells", Scientific Reports, 7(1), 1-15.

[21] Rawat, B., Dwivedi, S.K. (2019). "Selecting appropriate metrics for evaluation of recommender systems", International Journal of Information Technology and Computer Science.

[22] Rezayi, M., Farjami, Z., Hosseini, Z.S., Ebrahimi, N., Abouzari-Lotf, E. (2018). "MicroRNA-based biosensors for early detection of cancers", Curr Pharm Des, 24(39), 4675-4680.

[23] Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., Kanaya, S. (2006). "KNApSAcK: a comprehensive species-metabolite relationship database", In Plant Metabolomics, 165-181, Springer.

[24] Sohel, M., Biswas, P., Al Amin, M., Hossain, M.A., Sultana, H., Dey, D., Aktar, S., Setu, A., Khan, M.S., Paul, P. (2022). "Genistein, a potential phytochemical against breast cancer treatment-insight into the molecular mechanisms", Processes, 10(2), 415.

[25] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F. (2021). "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA: a cancer journal for clinicians, 71(3), 209-249.

[26] Zhang, E., Zhang, Y. (2009). "Average precision", In Liu, L. and ÖZsu, M.T., editors, Encyclopedia of Database Systems, Springer US, Boston, MA, 192-193.