Efficient Discovery of Similar Aspects by Gene Community Search for Similarity Explanation

Submitted for blind review

Abstract. Gene similar aspects provide reliable explanation in understanding the biological roles and gene functions. As the volume of biomedical data expands, most of the current methods for similar explanation among genes are no longer applicable. Limited by search efficiency and restrictions on the query genes, these methods cannot be flexible and efficient for the similarity analysis. We hereby propose a flexible method *VENUS* to analyze gene similar aspect among multiple genes on heterogeneous information networks. *VENUS* infers the semantic and structural similarity of the query genes by gene community search. In this way, our approach narrows the search space when searching information network within an acceptable time cost. Besides, *VENUS* is not limited by inherent domain knowledge and is adaptive to large-scale networks. Through experiments on five different public data sources, it demonstrates that *VENUS* is effective and efficient.

Keywords: Gene similar aspect · Gene information network · Community search

1 Introduction

Gene similar aspect analysis plays a vital role in figuring out biochemical reactions in vivo at the molecular level. With the further understanding of gene function and gene similar aspect, more far-reaching studies have emerged in the research fields of genetic diseases, cancer detection, treatment, and gene medicine. According to the FDA, twenty-two generalized cell and gene therapy drugs have been approved for marketing worldwide as of 2020 [11]. Benefiting from the accelerated advancement of the gene similar aspect, this number will continue to increase in the near future. For instance, Comirnaty, the first practical mRNA- based COVID-19 vaccine developed by BioNTech and Pfizer, became the first vaccine against COVID-19 approved by the world Health Organization (aka. WHO) on December 31, 2020 [5]. Its success is inseparable from detailed comprehension of viral gene similar aspect by medical scientists.

Current approaches to explan the gene similarity can usually be divided into three categories: gene sequence-based, gene ontology-based, and gene function-based. The first two rely entirely on the in-depth study of the nature of genes by scholars. However, gene function-based explanations are more focused on discovering gene correlations that are imperceptible from the existing knowledge. The function of genes can be demonstrated through interactions with other biological entities, such as mRNA, proteins and diseases. Therefore, the gene similarity

aspect can be used to analyze the gene similarity by biological pathways, which is also called the meta path, on the gene co-expression networks.

Presently, there are still some topical issues in recent gene similarity-based reasearch:

- How to represent gene similar aspect according to the gene co-expression network?
- How to measure the similarity between genes in accordance with the gene similar aspect?

With the advancement development of biotechnology, it becomes more convenient to obtain interactive information about biological entities. Gene similar aspect analysis based on information network has attracted more and more attention of researchers. Zhang *et al.* [19] first normalizes the concept of gene similar aspect and provide an explainable similarity measure based on the normalized definition. Nevertheless, there still exist some limitations to analyze the gene similar aspect by information network: (1) *Analysis limited by domain knowledge.* The similarity measures in these methods need to be pre-defined and usually need a threshold to limit the search space in the network. In addition, it is not quite reasonable to use the same threshold for different query genes. (2) *Inefficiency on large data sets.* As the amount of gene-related data increasing, the search space of traditional global search-based methods can be large, and the algorithm becomes less efficient. (3) *Lack of flexibility.* Methods of gene similar aspect detection have restrictions on query genes. These methods can only be applied to a pair of genes related to the same disease. They do not consider the synergy that exists between a group of genes, nor the between genes.

As more biological pathways (e.g., gene-miRNA, gene-protein) are discovered, the scale of the network becomes larger. Therefore, a medium is needed to put a large amount of heterogeneous data in the same standard metric. Inspired by the concept of community in social networks, in which community members are associated closely, we propose a novel concept of gene community. Besides, community search method can avoid the above-mentioned limitations, and implements an efficient search of genes without prior domain knowledge. To explain gene similarity from similar aspects by gene community, we propose a method *VENUS* (short for discovery of similar aspect by gene community search) to analyze gene similarity based on community search. The main contributions of this work are as follows:

- We propose a novel concept of gene community and use a community search approach to explain the similar aspect mining between query genes and search for similar genes.
- We design a framework, *VENUS*, to acquire the semantic relations and structure information of genes based on the gene community, and used an iterative algorithm to analyze similar gene aspect.
- We evaluate *VENUS* on a random gene set to demonstrate its effectiveness in searching the most similar aspect among multiple genes.

2 Related Work

2.1 Gene Similarity Search

The similarity search based on gene sequence These methods mainly measures the similarity between genes by comparing the similarity degree of gene sequence. According to the different methods of comparison, it is mainly divided into the Alignment-based and Alignment-free methods. The representative work includes BLAST [1], Identity [8] and etc.

Generally speaking, alignment-based methods can only use a single data source and are more time-consuming. In contrast, alignment-free methods have a significantly faster calculation speed than alignment methods, and can obtain more accurate experimental results in a shorter time. But compared to other methods, the data used by sequence-based methods is very simple and limited in information.

The similarity search based on gene function : Some methods use Gene Ontology (GO) to annotate information. At present, the common methods to determine the similarity of GO terms are based on path distance, information content, vector space, and fusion [12]. Besides, there are often complex biochemical reactions between different biological entities. With the concept of gene co-expression network, protein-protein interaction network and other concepts, we can use the associations between genes and their related biological entities to analyze gene similarity. Zhang *et al.* [19] put forward the *SCENARIO* which normalizes the concept of gene similar aspect. By comparing the scores of two given genes on different gene aspects.

These methods are dependent on the quality and availability of the data and have limited plausibility for explanation of the results. Besides, they did not take into account the similar aspects of a set of genes.

2.2 Community Search

Detection of community structure can reveal the structural features of complex networks, i.e. the division of a network into groups (clusters or modules) of nodes having dense intra-connections, and sparse interconnections [20]. Generally, community detection algorithms aim to identify all communities for a graph. With the rapid development of information technologies, various big graphs are prevalent in many real applications, however, it often takes a long time to detect all the communities. Therefore, query-based online community search tasks on large graphs have gained the attention of researchers. They aim to provide efficient solutions for searching high-quality communities from large networks in real-time. Cui *et al.* [6] firstly propose a local search strategy, which searches in the neighborhood of a vertex to find the best community for the vertex.

Note that though query-based is executed very efficiently, it is still necessary to design pruning strategies based on different network structure and community standard to obtain excellent performance.

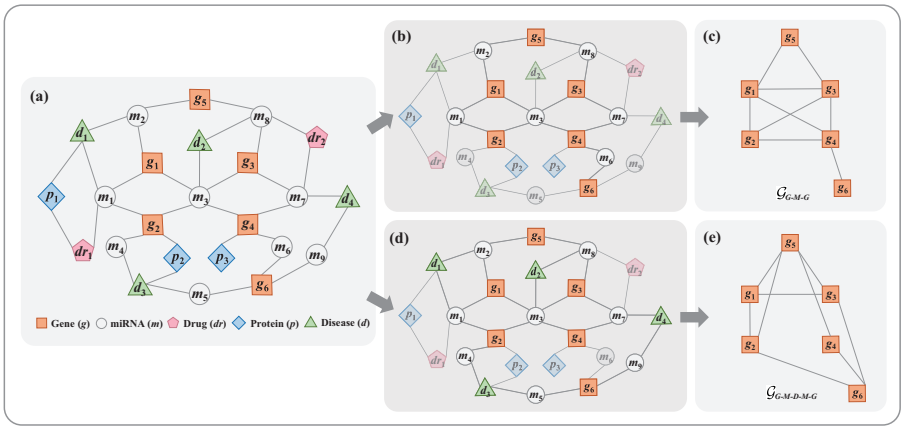

Fig. 1. An example of gene communities in GIN

3 Problem Definition

In order to discover the gene similar aspects among genes, we give some formal definitions that are crucial for the similarity computation.

Definition 1 (Gene Information Network). *A gene information network (GIN) is defined as an undirected graph* $\mathcal{G} = (V, E)$ *with an object mapping function* $\phi : V \to \mathcal{A}$ *and a link mapping function* $\psi : E \to \mathcal{R}$ *subject to* $|\mathcal{A}| > 1$ *and* $|\mathcal{R}| > 1$*, where* $\mathcal{A}$ *refers to the set of gene-related biological object types and* $\mathcal{R}$ *denotes the set of relations between objects. Each object* $v \in V$ *belongs to an object type* $\phi(v) \in \mathcal{A}$*, and each link* $e \in E$ *belongs to a relation* $\psi(e) \in \mathcal{R}$*.*

To study the semantic relationships between specific type of nodes, *gene meta path* and *gene similar aspect* are proposed to better understand meta-level description of a given GIN. In a GIN, any pair of genes can be connected via a gene meta path instance.

Definition 2 (Gene Meta Path). *A gene meta path* P *is a path defined on the graph of GIN where the two end nodes of* P *are two genes, and is denoted in the form of* $Gene \xrightarrow{R_1} A_1 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_l \xrightarrow{R_l^{-1}} \dots \xrightarrow{R_2^{-1}} A_1 \xrightarrow{R_1^{-1}} Gene$*, which defines a composite relation* $R = R_1 \circ R_2 \circ \cdots \circ R_2^{-1} \circ R_1^{-1}$ *between two gene objects, where* R^{-1} *denotes the inverse relation and* $\circ$ *denotes the composition operator on relations. And* $Ins(P_{g \to g'})$ *denotes the set of paths which go from* g *to* g' *following* P*, where* g *and* g' *belong to gene set.*

Example 1. Take Figure 1(a) as an example, there are two meta paths between g_1 and g_2: "*G-M-G*" (short for "*Gene-MiRna-Gene*") and "*G-M-D-M-G*" (short for "*Gene-MiRna-Disease-MiRna-Gene*"), cause there are two instances: "$g_1 - m_3 - g_2$" and "$g_1 - m_2 - d_1 - m_1 - g_2$" .

Definition 3 (Gene Similar Aspect). *Given two gene nodes g_1 and g_2, a gene similar aspect is a nonempty subset of all gene meta paths between g_1 and g_2. Intuitively, a gene similar aspect is a meta path or a combination of multiple meta paths, which can express gene similarity more comprehensively.*

As mentioned by Sun *et al.* [15], the similarity between any pair of gene nodes (g_1, g_2) which are connected by meta path P can be measured as:

$$Sim(g_1, g_2, P) = \frac{2\times \mid Ins(P_{g_1 \to g_2}) \mid}{\mid Ins(P_{g_1 \to g_1}) \mid + \mid Ins(P_{g_2 \to g_2}) \mid} \tag{1}$$

Though Eq.(1) is effective in computing the similarity based on a single meta path, there may be multiple meta paths among nodes. Thus, we extend it to a more general scenario where multiple genes are considered. Therefore, we consider the case where multiple meta paths form a gene similar aspect.

To compare the similarity of query genes based on meta path, we propose the concept of gene community. Each community has an aggregation degree to describe the level of association of nodes within the community, defined as follows.

Definition 4 (Gene Community). *A gene community $\mathcal{G}_P = (V_P, E_P)$ is a graph which is generated based on the meta path P of interest and query nodes S from a GIN. Specifically, The generated gene community should satisfy the following rules:*

1) $S \subseteq \mathcal{V}_P$.
2) any two gene nodes are neighbors of each other only if there is an instance of meta path P between them.

*Example 2. Figure 1 illustrates the two gene community extracted from the gene information network shown in Figure 1(a) in accordance with two meta paths respectively. For the meta path "*G-M-G*", any edge appears in Figure 1(c) indicates the existence of a "*G-M-G*" instance on the GIN for the two endpoints of the edge. When the query gene set is $\{g_1, g_2, g_4\}$, it is obvious that they are more similar on the meta path "*G-M-G*" because they are related to each other on "*G-M-G*", but in Figure 1(e) only g_1 and g_2 are connected.*

Conceptually, a gene community is a set of vertices which are connected by a specific meta path cohesively.

By building a gene community, a heterogeneous GIN is reconstructed into a homogeneous network incorporating meta path information. Further, to assess the association level of gene communities under different meta paths, we propose the concept of core number.

Definition 5 (k-core community). *We define a gene community $\mathcal{G}_P$ to be a k-core community if it satisfies*

$$k = \min\{deg_{\mathcal{G}_P}(v) | v \in \mathcal{G}_P\} \tag{2}$$

where the degree of a vertex v of $\mathcal{G}_P$ is denoted by $deg_{\mathcal{G}_P}(v)$.

Example 3. As Figure 1(c) shows, the subgraph which only contains gene set $\{g_1, g_2, g_3, g_4, g_5\}$ is a 2-core community. However, in Figure 1(e), the subgraph which contains the same gene set is a 1-core community for the vertex g_4 has only one neighbor g_5.

In summary, gene communities incorporate both semantic and structural information among gene nodes from a GIN. The aggregation degree of a gene community can reflect the gene similarity under a specific gene meta path, and it can be evaluated by its core number.

4 Design of Method

In this section, we discuss the framework of *VENUS* in detail. It takes a query gene set S together with a GIN as input, and outputs a set of Top-K gene similar aspects of the given query gene set.

4.1 Candidate Meta Path Generation

In order to solve the problem of finding gene similar aspects among genes in a GIN, *VENUS* first generates candidate meta paths for the given query set. In the process of candidate meta paths generate, an adaptive and flexible method is used to search for meta path sets that can describe the query genes. As shown in Algorithm 1, the proposed method is not based on pre-defined domain knowledge (such as specific meta paths and limited length of the meta path) as opposed to the previous methods.

We only use a set of query genes as input for finding meta paths without pre-defined parameters. In Steps 1-3, we ensure that a reachable path exists between any two nodes in the query genes. In Steps 4-15 of the algorithm, the optional meta paths between each query gene to their one-hop genes are searched using breadth-first search (BFS), which is upper bounded by $\mathcal{O}(|V| \cdot |S|)$ on time complexity for there are no common edges between the paths and the size of the graph is fixed. Finally, we filter the meta paths that can jointly describe each query gene in Step 16. In this way, the semantic relationship between the current query gene and the neighbor genes are represented by the meta paths. After obtaining the set of candidate meta paths, all nonempty subsets of the set can be used as candidate gene similar aspects.

4.2 Gene Community Search

For each candidate gene meta path, we choose the gene community that contains the maximum number of cores generated by the query genes. Based on Definition 3, the following two observations can be obtained.

Observation 1: GIN is a 1-core gene community, because it is a contiguous graph and contains all query genes.

Algorithm 1 *Candidate Meta Paths Generate*

Input: Gene information network $\mathcal{G}$, a set of query gene nodes S
Output: The candidate meta paths set PS among the query nodes S

if S is not in same connected component of $\mathcal{G}$ **then**
  return $\emptyset$
end if
for g_i in S **do**
  $Vis \leftarrow \{g_i\}$, $NodeSet \leftarrow \{g_i\}$
  while not $NodeSet.empty()$ **do**
    $v \leftarrow NodeSet.pop()$
    $u \leftarrow v.neighbors()$
    for u_j in u **do**
      if the type of u_j is not gene and $u_j \notin Vis$ **then**
        $Vis \leftarrow u_j$, $NodeSet \leftarrow u_j$
      else if $P_{g_i \to u_j}$ is independent **then**
        $PS_{g_i} \leftarrow P_{g_i \to u_j}$
      end if
    end for
  end while
end for
$PS = \bigcap_1^{|S|} PS_{g_i}$
return PS

Observation 2: For any integer $0 < k \leq k_{max}$, the k-core is contained by the $(k-1)$-core, where k_{max} is the maximum core number. In other words, A k-core community also satisfies the definition of $(k-1)$-core.

Notably, the number of k-core communities is monotonically decreasing as k increases by Observations 1 and 2, so we use a dichotomous approach to speed up the search process.

The community search process is shown in Algorithm 2. For a given number k, we can determine whether there is a k-core community in the time complexity of $\mathcal{O}(V)$, V denoting the number of links. The complexity can be proved in Appendix B of [**?**]. The query genes may be distributed in communities with different core numbers, and we select the largest core number, which is denoted as k_P, among them as the aggregation degree of the query genes on GIN.

4.3 Similarity Calculation based on Gene Community

To distinguish the meta paths of GINs for similarity computation, we add structure information to Eq.(1), which is only applicable to similarity analysis with specific meta path between two connected genes. *VENUS* inherits the iterative method of GSimRank [18] and updates the similarity score by the same kind of meta path instances, with each iteration referring only to the neighbors' similarity information in the community:

Algorithm 2 *Gene Community search*

Input: Gene information network $\mathcal{G}$, a set of query gene nodes $S = \{g_i\}$, meta path P

Output: The gene community $\mathcal{G}_P$ and its core number k_P

$maxdegree \leftarrow 1$, $mindegree \leftarrow 1$, $\mathcal{G}_P \leftarrow \emptyset$
for g_i in S **do**
  $maxdegree \leftarrow \max(maxdegree, deg_{\mathcal{G}}(g_i))$
end for
while $mindegree \leq maxdegree - 1$ **do**
  $d \leftarrow \lfloor (mindegree + maxdegree)/2 \rfloor$
  if $d - core$ gene community $\tilde{G}$ exist **then**
    $mindegree \leftarrow d, \mathcal{G}_P \leftarrow \tilde{\mathcal{G}}$
  else
    $maxdegree \leftarrow d$
  end if
end while
$k_P \leftarrow mindegree$
return $\mathcal{G}_P, k_P$

$$CSim(S_i, P) = \sum_{S_j}^{\Gamma(S_i,P)} \frac{CSim(S_j, P)}{|\Gamma(S_j, P)|} \tag{3}$$

where $\Gamma(S, P)$ contains all sets that are reachable by the S in a gene community $\mathcal{G}_P$ linked by meta path P, in which each element S' satisfies

$$S' = \{s_i' | Ins(P_{s_i \to s_i'}) \neq \emptyset, s_i \in S, s_i' \in S'\} \tag{4}$$

According to Eq.(3), the structural similarity of the query genes within the generated gene community can be obtained. For a gene similar aspect which may contain more than one meta path, the semantic similarity represented by each meta path should also be considered, Based on the core set, we assign weights to each meta path in aspect A:

$$W_{P_i} = \frac{k_{P_i}}{\sum_{j=1}^{|A|} k_{P_j}} \tag{5}$$

The similarity containing semantic and structural information is expressed as:

$$Sim(S_i, A) = \sum_{j=1}^{|A|} W_{P_j} \times CSim_(S_i, P_j) \tag{6}$$

The method of calculating similarity can be intuitively translated into a random walking model in which the speed of convergence starting with the query genes is calculated. Unlike meta path restrictions that exist in Eq.(1),

Table 1. Characteristics of the Gene Information Network

Relation	Type	# Nodes	# Edges	Source
Gene-Protein	gene	20209	20209	HGNC [4]
	protein	20075		
Drug-Protein	drug	5592	15567	DrugBank [17]
	protein	2796		
Gene-Disease	gene	17382	402329	OMIM [2] DisGeNET [9]
	disease	4284		
Disease-Phenotype	disease	4353	92153	OMIM [2]
	phenotype	44362		
miRNA-Gene	miRNA	2596	320372	miRNet [7]
	gene	14736		
miRNA-Disease	miRNA	684	5550	miRNet [7]
	disease	98		

VENUS reduces the space to be searched by incorporating the semantics into the adjacent edges in advance. In the case of equal cohesion in the community search, the community with smaller size is selected for similarity calculation.

5 Empirical Evaluation

5.1 Experimental settings

Datasets The gene information network is constructed by five different data sources. The main statistics of associations is listed in Table 1. The instance of each relationship constitutes a bipartite graph. In order to maintain the connectivity of the constructed network, we preprocess these extracted relations and entities to ensure that there is no outlier points. Finally, the gene information network is formed with the pre-processed information. In summary, there are 135,716 nodes and 856,180 edges.

Setup According to the experimental results, the optimal parameters are selected for each step. For the aforementioned methods, as different running parameter values corresponding to different performances, we apply the default parameters.

All experiments were conducted on a PC with four Intel Xeon E5-2698 GHz CPUs, four GeForce RTX 2080 Ti GPUs and 512 GB memory, running Ubuntu 20.04. The algorithms were implemented in Python and compiled by Python 3.7.

5.2 Effectiveness

To evaluate the performance of *VENUS*, we design the following steps to prove it. Firstly, we use Enzyme Comparison (EC) number [13] to reduce the effect on experimental results under different criteria of discrimination of gene similarity.

Table 2. Overall Performances of Comparison with Baselines

Valid data size	150			300			600			1200		
Metric	ACC	AUROC	PRC	ACC	AUROC	PRC	ACC	AUROC	PRC	ACC	AUROC	PRC
Lin's method	0.6933	0.7210	0.5719	**0.7967**	0.8251	0.6777	0.7833	0.8011	0.6701	0.7925	0.8124	0.6778
Resnik's method	0.6933	0.8081	**0.7718**	0.7367	0.9019	0.8726	0.7300	0.8694	0.8358	0.7408	0.8797	0.8519
Wang's method	0.6466	0.7090	0.5582	0.7400	0.8076	0.6503	0.7450	0.7956	0.6659	0.7450	0.8030	0.6637
SCENARIO	0.7200	0.8672	0.6849	0.7567	0.9371	0.8710	0.7616	0.9334	0.8848	0.7756	0.9351	0.8663
VENUS	**0.7600**	**0.8962**	0.7339	0.7700	**0.9405**	**0.8971**	**0.7920**	**0.9411**	**0.8977**	**0.7973**	**0.9371**	**0.8726**

According to the theory that genes annotated by the same EC number are functionally similar. Any two genes with the same EC number should be predicted to be associated, i.e., a positive example sample. Relatively, by randomly sampling from the whole gene set excluding nodes annotated by the EC number, we construct negative samples.

Table 2 compares the results of gene similarity measurement. Based on such results, we provide the following analysis.

We count the scores of gene similarity analysis methods under three criteria. Although each of these methods explained gene similarity from different information, all of them obtain relatively good scores, which indicates that gene similarity can be analyzed by multiple perspectives. The experiment result shows that *VENUS* outperforms all the baselines, which indicates that the gene community is helpful for gene similarity analysis. In conclusion, *VENUS* has good performance in analyzing the similarity of a pair of genes.

5.3 Case Study

We use the method of pathway enrichment analysis (PEA) to verify that the experimental results given by *VENUS* have relatively obvious biological significance. PEA is often used to show the involvement of different genes in biological pathways. Genes with close functional connections are often enriched in the same biological pathway. In this part, KOBAS is used as the main tool to perform the PEA. Notably, during the process of PEA the entire human genome will be used as a background gene.

A gene set, BRAF, PIK3CA, KRAS, EGFR, and MAP2K1, is randomly selected from HGNC that may be related to NSCLC (nonsmall-cell lung cancer) as the input data for *VENUS*. By searching gene communities from GIN, we find that among the communities where the five genes are involved, the genes participating in the Gene-Disease-Gene meta path constitute one community with the highest core number. Besides, by looking for gene nodes in the generated community with a high number of meta path instances, the following genes can be found: RASSF1, PLCG1, ARAF, CASP9, PKPD1 and EML4. All of these genes are related to NSCLC, which also shows the effectiveness of *VENUS* in searching for similar genes.

To further explain the biological significance of the program results, we use KOBAS to perform the PEA on these relating genes based on KEGG [10]. Ta-

Table 3. Pathways Significantly Enriched in NSCLC Pathogenic Genes

Gene Set	Pathway	p-value
(BRAF, PIK3CA, KRAS, EGFR, MAP2K1)	Endometrial cancer	9.16E-15
	Non-small cell lung cancer	1.68E-14
	Melanoma	2.55E-14
	Pancreatic cancer	3.10E-14
	Glioma	3.10E-14
(EML4, BRAF, KRAS, EGFR, MAP2K1, PIK3CA, PDPK1)	Non-small cell lung cancer	5.72E-20
	Endometrial cancer	1.03E-16
	Prostate cancer	1.97E-15
	FoxO signaling pathway	1.18E-14
	Proteoglycans in cancer	1.48E-13

ble 3 shows pathways with the five smallest p-value based on different gene sets. This demonstrates that the experimental results obtained by KOBAS are statistically significant. The five genes are of extremely importance in the pathogenesis of endometrial cancer, non-small cell lung cancer and other tumor diseases, which can prove the five genes have a synergistic or similar function. At the meantime, several genes with high similarity obtained by *VENUS*, such as above-mentioned EML4 and PDPK1, are added to the enrichment analysis, and it can also be demonstrated that these genes are involved in a pathway. This indicates that the genes found by *VENUS* have a high correlation with the query gene in reality.

6 CONCLUSION

To discover gene similar aspect among query genes in multiple data sources, we propose a general method *VENUS* by gene community search from gene information networks. *VENUS* computes similarity by gene community which can reveal dense connections among multiply genes. It is novel in providing reasonable explanation of gene similar aspect by combining semantic information and gene community structure information.

Through the validation with the benchmark, the results of *VENUS* show high accuracy in discovering similar genes. In the case study, we further demonstrated the effectiveness of *VENUS* for a set of query gene similarity analysis by pathway enrichment analysis.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)

2. Amberger, J.S., Bocchini, C.A., Scott, A.F., Hamosh, A.: Omim.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Research **47**(D1), D1038–D1043 (2018)
3. Amelio, A., Pizzuti, C.: Overlapping community discovery methods: a survey. In: Social Networks: Analysis and Case Studies, pp. 105–125 (2014)
4. Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., Bruford, E.: Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Research **47**(D1), D786–D792 (2018)
5. Cavaleri, M., Enzmann, H., Straus, S., Cooke, E.: The european medicines agency's eu conditional marketing authorisations for covid-19 vaccines. The Lancet **397**(10272), 355–357 (2021)
6. Cui, W., Xiao, Y., Wang, H., Wang, W.: Local search of communities in large graphs. In: Proceedings of International Conference on Management of Data, SIGMOD, USA. pp. 991–1002 (2014)
7. Fan, Y., Siklenka, K., Arora, S.K., Ribeiro, P., Kimmins, S., Xia, J.: miRNet-dissecting mirna-target interactions and functional associations through network-based visual analysis. Nucleic Acids Research **44**(W1), W135–W141 (2016)
8. Girgis, H.Z., James, B.T., Luczak, B.B.: Identity: rapid alignment-free prediction of sequence alignment identity scores using self-supervised general linear models. NAR genomics and bioinformatics **3**(1), lqab001 (2021)
9. González, J.P., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I.: The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research **48**(Database-Issue), D845–D855 (2020)
10. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**(Database-Issue), D353–D361 (2017)
11. Ma, C.C., Wang, Z.L., Xu, T., He, Z.Y., Wei, Y.Q.: The approved gene therapy drugs worldwide: from 1998 to 2019. Biotechnology advances **40**, 107502 (2020)
12. Othman, R.M., Deris, S., Illias, R.M.: A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. Journal of biomedical informatics **41**(1), 65–81 (2008)
13. Peng, J., Wang, Y., Chen, J.: Towards integrative gene functional similarity measurement. BMC Bioinformatics **15**(S-2), S5 (2014)
14. Pesquita, C., Faria, D., Bastos, H., Falcao, A., Couto, F.: Evaluating GO-based semantic similarity measures. In: Proceedings of 10th Annual Bio-Ontologies Meeting. p. 38 (2007)
15. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: PathSim: meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment **4**(11), 992–1003 (2011)
16. Tian, Z., Guo, M., Wang, C., Xing, L., Wang, L., Zhang, Y.: Constructing an integrated gene similarity network for the identification of disease genes. Journal of biomedical semantics **8**(1), 27–41 (2017)
17. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research **46**(D1), D1074–D1082 (2017)
18. Zhang, C., Hong, X., Peng, Z.: GSimRank: a general similarity measure on heterogeneous information network. In: Proceedings of Asia-Pacific Web and Web-Age

Information Management Joint International Conference on Web and Big Data, China. pp. 588–602 (2020)
19. Zhang, Y., Duan, L., Zheng, H., Li-Ling, J., Hu, B., Qin, R., He, C.: SCENARIO: discovery of similar aspects for gene similarity explanation from gene information network. In: Proceedings of 2019 IEEE International Conference on Bioinformatics and Biomedicine. pp. 604–609 (2019)
20. Zhou, Y., Liu, L.: Social influence based clustering of heterogeneous information networks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA. pp. 338–346 (2013)