Mining Similar Aspects for Gene Similarity Explanation Based on Gene Information Network

Yidan Zhang, Lei Duan✉, Huiru Zheng, Jesse Li-Ling, Ruiqi Qin, Zihao Chen, Chengxin He, Tingting Wang

Abstract—Analysis of gene similarity not only can provide information on the understanding of the biological roles and functions of a gene, but may also reveal the relationships among various genes. In this paper, we introduce a novel idea of *mining similar aspects from a gene information network*, i.e., for a given gene pair, we want to know in which aspects (meta paths) they are most similar from the perspective of the gene information network. We defined a similarity metric based on the set of meta paths connecting the query genes in the gene information network and used the rank of similarity of a gene pair in a meta path set to measure the similarity significance in that aspect. A minimal set of gene meta paths where the query gene pair ranks the highest is a similar aspect, and the similar aspect of a query gene pair is far from trivial. We proposed a novel method, *SCENARIO*, to investigate minimal similar aspects. Our empirical study on the gene information network, constructed from six public gene-related databases, verified that our proposed method is effective, efficient, and useful.

Index Terms—similar aspect, gene information network, gene meta path

1 Introduction

Searching similar genes is a fundamental problem that has been studied in biological research such as gene clustering [1], [2], prediction and evaluation of protein interaction [3], [4], prediction of gene function [5], [6], and prioritization of disease-associated genes [7], [8]. Gene similarity analysis is an important task as it can provide abundant information for the understanding of their biological roles, functions, as well as the various relationships among them.

Intuitively, gene similarity analysis involves two important parts, i.e., the similarity measure and similarity comparison.

- Firstly, the similarity measure determines how similar two genes are, usually in a quantitative way.
- Secondly, the similarity comparison distinguishes each gene pair from the others by comparing their similarities, in order to tell how significantly similar a gene pair is compared with the others.

Generally, current research on the gene similarity analysis can be divided into three main categories: the sequence-based, the annotation-based, and the association-based approaches.

- *Sequence-based*: Some studies find similar genes by analysing their sequences for genes with similar sequences are likely to be similar in functions. Some methods align all the sequences to perform the analysis. For example, *BLAST* [9] is an alignment search tool for finding regions of similarity between gene sequences, which are used to identify members of gene families. *T-Coffe* [10] was designed for multiple sequence alignment, which can provide a dramatic improvement in accuracy. Some heuristics methods base the similarity analysis mainly on statistical characteristics of data. For example, *CPF* [11] employed compounded information to conduct DNA clustering, where word frequency, position and classification information of nucleotide bases from DNA sequences were used. Zhou *et al.* [12] constructed complex networks of DNA sequences for clustering analysis based on the central dogma. However, such methods tend to be time-consuming or inaccurate, and the data used are rather simple, and can provide very limited information. Besides, such methods can only handle a single data source.
- *Annotation-based*: Some studies conduct the gene similarity analysis by their Gene Ontology (GO) annotations [13], which systematically annotate gene functions. For example, Aurelien *et al.* proposed *ViSEA-GO* [14], an extension of classical functional GO analysis, which focuses on functional coherence through visualization, semantic similarity and enrichment analysis of Gene Ontology. Giri *et al.* [15] proposed a multi-view gene clustering approach containing two complementary views, one of which is based on Euclidean distance between gene expression values and the other is based on a GO-based gene-gene similarity measure. However, although GO is of large scale and covers a wide range of data, it is not accurate as it is half-manually sorted. Consequently,

- *Y. Zhang is with the School of Computer Science, Sichuan University, Chengdu 610065, China.*
- *L. Duan is with the School of Computer Science, Sichuan University, Chengdu 610065, China. E-mail: leiduan@scu.edu.cn.*
- *H. Zheng is with the School of Computing, Ulster University, Northern Ireland, United Kingdom.*
- *J. Li-Ling is with the State Laboratory of Biotherapy, Sichuan University, Chengdu 610041, China.*
- *R. Qin, Z. Chen, C. He and T. Wang are with the School of Computer Science, Sichuan University, Chengdu 610065, China.*

errors may occur and biased conclusions may be drawn.

- *Association-based*: Some studies analyse the gene similarity based on the associations between genes and gene-related entities. For example, *RWRB* [8] constructs an integrated gene similarity network based on five individual gene or protein similarity networks to infer causal genes of target diseases. *GAIN* [16] constructs bipartite networks of biological entities, where the interaction-profile similarities are calculated and compared, and the modules of genes with similar profiles are defined. *RGFSN* [17] is an integrated gene functional similarity network established by six different methods and is further refined by the PPI networks to perform the protein complex prediction. However, such methods are limited in giving explanations to the results, as they only conduct the computation of similarity and directly compare the results, lacking reasonability.

Unfortunately, current methods have the following limitations in terms of (1) similarity measure when evaluating the similarity of two query genes, or (2) similarity comparison when evaluating the distinctiveness of the query genes to the other genes, both of which have been pointed out as two critical parts in gene similarity analysis:

- *For similarity measure*, there are two major drawbacks: (i) the absolute numerical result produced by the similarity measure is directly used as the final evaluation of gene similarity, which may lead to biased conclusions that lack of reasonability. (ii) the adopted similarity measure for different queries is fixed rather than flexibly adjust itself to various queries according to their unique characteristics.
- *For similarity comparison*, two drawbacks also exist: (i) the conclusion is drawn either by comparing the quantitative results of the targets directly, or by using a threshold to judge whether the query genes are similar or not, while both of them are unreasonable. (ii) the query genes are invariably compared with the rest of the genes in the whole gene set, which lacks of reasonability and flexibility, and it is not always necessary to do so.

Consequently, such limitations result in the weakness in the explanations of the gene similarity. To address the aforementioned challenges, it is necessary to evaluate gene similarity with flexible similarity measures and a reasonable way of similarity comparison.

Intuitively, genes can be regarded as similar if, for example, they are annotated by the same GO term, they are targeted by the same miRNA, they can cause the same disease, or any over two of these situations combined. Therefore, the similarity of two genes can be reflected by the strongness of their relationships in different aspects. In other words, the *similar aspect* of a set of genes indicates they have a similar relationship. For example, three genes, *ABCC8*, *GCK* and *KCNJ11*, all can result in a type of diabetes called NIDDM (*non-insulin-dependent diabetes mellitus*). Thus, a similar aspect among the above three genes is the fact that they can cause the same disease NIDDM.

Formally, for query genes, we consider evaluating and explaining their similarity by their similar aspects, and thus propose a novel approach, *SCENARIO* (short for similar aspects for gene similarity explanation), for gene similar aspects detection. The characteristics of *SCENARIO* include: (1) it uses the *similar aspect* for the final evaluation of gene similarity rather than the quantitative results produced by the similar measures; (2) it uses flexible similarity measure which is able to make a slight adjustment to itself regarding to different query genes by their characteristics; (3) it uses rank statistics as the comparison between query genes with others without any manually predefined threshold; (4) it flexibly supports the comparison of two or more than two query genes with gene pairs in their neighborhood to discover the similar aspect for the query genes; (5) it computes gene similarity solely based on the associations between genes and their related biological entities, and detects similar aspects based on multiple data sources; (6) it is capable of explaining the similarity between any query genes by detecting their similar aspect.

The main contributions of this work are as follows:

- We introduce the novel problem of mining similar aspects of genes from a gene information network.
- We propose *SCENARIO*, a general framework for calculating aspect similarity for query genes to detect their similar aspect, which can be used to explain the similarity between them.
- We show how *SCENARIO* measures gene aspect similarity under a flexible meta path-based method, and compare the similarities in an adjustable scope.
- We evaluate *SCENARIO* on a random gene set to demonstrate its effectiveness in searching the similar aspect between query genes.

The rest of the paper is organized as follows. We review related work in Section 2, and formulate the problem of mining similar aspects in Section 3. In Section 4, we discuss the critical techniques of our method *SCENARIO*. We report a systematic empirical evaluation in Section 5, and conclude the paper in Section 6.

2 Related Work

2.1 Gene Similarity Analysis Under Multiple Data Sources

Technological advances in data generation from multiple levels of biology have driven the field of bioinformatics for the past decades, producing ever-increasing amounts of data as researchers strive to develop various databases. Analysis under multiple data sources methods are now emerging that aims to bridge the gap between our ability to generate vast amounts of data and our understanding of biology.

Aerts *et al.* [18] developed a software termed Endeavour that can find candidate genes underlying biological processes or diseases by fusing multiple data sources to rank unknown candidate test genes according to their similarity with known training genes. Genehopper [19] was developed as a search engine where a user can explore the neighborhood of a target gene by a gene-to-gene search as the weighted sum of nine normalized gene similarities based on multiple data sources. Besides, each weight can be adjusted by the user, allowing flexible customization of the gene search. Wang *et al.* [20] proposed a novel machine learning model, named *MFR*, for accurately measuring the similarity between gene expressions by incorporating features of gene ontology, transcriptomics, proteomics and so on. Bass *et al.* [21] represented biological processes among biological entities such as genes, tissues, proteins, and metabolites as networks, and utilized different association indices to integrate the networks to explore the similarity between genes.

However, existing gene similarity analysis methods under multiple data sources focus on how to fuse multiple data sources or how to measure the similarity between genes, while failing to explain the gene similarity after data fusion.

2.2 Similarity Measure Based on Ontology

Ontology is widely used to measure the semantic similarity of two entities. Typically, Resnik [22] proposed a notion of information content (IC) to measure the semantic similarity of concepts in an IS-A taxonomy. Consequently, Lin [23] presented an information-theoretic definition of similarity which could be applicable in a number of different domains, *e.g.*, the biology domain, as long as probabilistic models are available.

In the bioinformatics field, GO is a standardized vocabulary of terms defined to represent gene product properties, which can annotate the function of genes. The GO terms are organized in a tree structure, thus some methods compute the similarity by utilizing the topological structure of GO terms in the graph [24]. In order to eliminate the misleadings from gene annotation statistics, Wang *et al.* [25] proposed a hybrid method considering not only the number of two GO terms' common ancestors, but also the distances between them and these ancestors.

GO-based approaches measure the gene similarity by using functional annotation. However, there are many non-functional relationships among genes, such as gene expression, transcription, and phenotype. As a result, the non-functional relationships among genes should be considered when measuring gene similarity.

2.3 Similarity Search on Heterogeneous Information Network

Heterogeneous information network (HIN) is a novel tool used to model the real-world in many scenarios, such as social networks and bibliographic networks. Due to the multiple types of objects and links involved in the network, the relations in HIN carry richer information and complex semantics than the relations in homogeneous networks do. Similarity search, which is a typical and important problem in data mining, has been studied by many researchers utilizing HIN.

PathSim [26] was proposed to measure the similarity between peer objects in a HIN, which was defined on a meta path framework. Considering the structure of meta path is relatively simple, Huang *et al.* [27] then proposed the concept of meta structure, which can describe more complex relationships between nodes, to measure the similarity between objects in HIN.

As an extension of similarity search, recommendation problem has also attracted the attention of researchers. Shi *et al.* [28] proposed the weighted HIN and weighted meta path to distinguish different attribute values of links, which can better capture the subtle semantics of paths for more accurate recommendations. Zhao *et al.* [29] proposed the concept of meta graph and introduced it to HIN to represent high-level semantics of recommendations, and based on the meta graph, the heterogeneous information in HIN was further fused to make better recommendations.

We tackled the problem of similar aspect mining in [30], a preliminary version of this paper. Compared to that work, in this paper, we provide a more flexible version of our method as well as a more detailed description of the key steps in the method, and perform more extensive empirical evaluations, including mining similar aspects from multiple genes.

3 Problem Formulation

We start with some preliminaries. Gene information network, which is a typical heterogeneous information network, can flexibly represent the relationships among gene-related biological objects [31]. Formally,

Definition 1 (Gene Information Network). A gene information network (GIN) is a graph $\mathbb{G} = (V, E)$ with an object mapping function $\phi : V \rightarrow \mathcal{A}$ and a link mapping function $\psi : E \rightarrow \mathcal{R}$ 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 relation types 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 type $\psi(e) \in \mathcal{R}$.

Definition 2 (Meta Path). A meta path P is a path defined on the gene information network and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, where $A \in \mathcal{A}$ and $R \in \mathcal{R}$. And $R = R_1 \circ R_2 \circ \cdots \circ R_l$ is a composite relation between object type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.

The length of the path is the number of objects contained in it, denoted by $|P|$.

A meta path P is a *gene meta path* if the two end nodes of P are two genes.

Example 1. An example of gene information network is illustrated in Figure 1. There are seven gene-related biological object types, i.e., $\{G, Pro, Dg, T, M, Dis, Phe\}$ and multiple gene meta paths. For example, the gene meta path "$G - T - G$" indicates two genes sharing the same GO terms, with gene path instances such as "$g_1 - t_2 - g_2$" and "$g_1 - t_3 - g_3$". And the gene meta path "$G - M - G$" indicates two genes are targeted by the same miRNAs, with gene path instances such as "$g_2 - m_2 - g_4$" and "$g_3 - m_3 - g_4$". The number of objects contained in "$G - T - G$" is 3, and so is "$G - M - G$". The gene meta path length of "$G - T - G$" and "$G - M - G$" are both 3.

In a GIN, for a gene meta path P, the *gene path instance set* between two genes g and g', denoted by $Ins(P_{g \rightarrow g'})$, is the set of paths which go from g to g' following P.

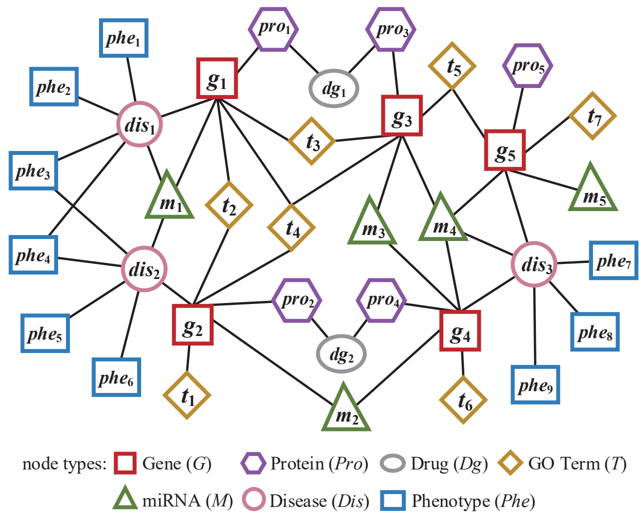


Fig. 1. An example of gene information network

As stated in [26], the similarity between any two peer nodes in a heterogeneous information network can be measured by the instances of a *symmetric meta path*. Correspondingly, given a GIN, for two genes g_1 and g_2, their similarity based on a symmetric gene meta path P can be defined as:

$$Sim_P(g_1, g_2) = \frac{2 \times |Ins(P_{g_1 \to g_2})|}{|Ins(P_{g_1 \to g_1})| + |Ins(P_{g_2 \to g_2})|} \tag{1}$$

Equation 1 is effective in computing the similarity based on a single gene meta path. However, please note that for two genes there may exist multiple gene meta paths. For example, the gene meta paths between g_3 and g_4 in the GIN shown in Figure 1 include: "$G - Dis - G$" and "$G - Pro - Dg - Pro - G$".

For a pair of genes g and g', we denote by

$$\mathcal{PS}(g, g') = \{P \mid Ins(P_{g \to g'}) \neq \emptyset\}$$

the set of gene meta paths between g and g'.

We consider a more general case in this study. For a pair of genes g_1 and g_2, let $\mathcal{P}$ be a subset of all gene meta paths between g_1 and g_2, i.e., $\mathcal{P} \subseteq \mathcal{PS}(g_1, g_2)$. Then, the similarity between g_1 and g_2 on the gene meta path set $\mathcal{P}$, denoted by $pSim_{\mathcal{P}}(g_1, g_2)$, is

$$pSim_{\mathcal{P}}(g_1, g_2) = \frac{2 \times \prod\limits_{P \in \mathcal{P}} |Ins(P_{g_1 \to g_2})|}{\prod\limits_{P \in \mathcal{P}} |Ins(P_{g_1 \to g_1})| + \prod\limits_{P \in \mathcal{P}} |Ins(P_{g_2 \to g_2})|} \tag{2}$$

Clearly, we have $0.0 \leq pSim_{\mathcal{P}}(g_1, g_2) \leq 1.0$.

Observation 1. In a gene information network, two genes are considered to be more similar if they are connected via shorter gene meta paths.

Example 2. In Figure 1, for gene pair (g_1, g_5), the instance of gene meta path "$G - T - G - T - G$" is "$g_1 - t_3 - g_3 - t_5 - g_5$" with the length of 5. For gene pair (g_1, g_3), the instances of gene meta path "$G - T - G$" are "$g_1 - t_3 - g_3$" and "$g_1 - t_4 - g_3$" with the length of 3. It is easy to find that the gene meta path instance "$g_1 - t_3 - g_3 - t_5 - g_5$" is an extension of instance "$g_1 - t_3 - g_3$", which indicates that the relationship between g_1 and g_5 is not as strong as the relationship between g_1 and g_3. In other words, g_1 is more similar to g_3 than to g_5.

Definition 3 (Gene Neighborhood). For a gene g, its neighborhood, denoted by $\mathcal{N}(g)$, is the set of all genes such that each gene is connected with g via at least one gene meta path, i.e.,

$$\mathcal{N}(g) = \{g' \mid \mathcal{PS}(g, g') \neq \emptyset\}$$

Please note that the genes in the group of gene pairs for similarity analysis is a subset of all genes in the GIN. Due to the huge number of genes in a typical GIN, it is unreasonable to consider all genes, since the semantic meaning between the genes with long path is insignificant.

Definition 4 (Gene Pair Neighborhood). For a pair of genes (g_1, g_2), their gene pair neighborhood, denoted by $\mathcal{N}(g_1, g_2)$, is the set of gene pairs generated by a query gene and one of its neighbor genes on gene meta path set $\mathcal{PS}(g_1, g_2)$, i.e.,

$$\begin{aligned}\mathcal{N}(g_1, g_2) = \{&g_i \in \mathcal{N}(g_1) \cup \mathcal{N}(g_2) \mid \exists P \in \mathcal{PS}(g_1, g_2), \\ &P \in \mathcal{PS}(g_1, g_i) \vee P \in \mathcal{PS}(g_2, g_i)\}\end{aligned}$$

Example 3. As shown in Figure 1, the gene meta path set between g_1 and g_2 is $\mathcal{PS}(g_1, g_2) = \{$"$G - T - G$", "$G - Dis - M - Dis - G$", "$G - Dis - Phe - Dis - G$"$\}$. So the gene pair neighborhood of (g_1, g_2) is $\mathcal{N}(g_1, g_2) = \{g_1, g_2, g_3\}$. This is because g_1 and g_3 share gene meta path "$G - T - G$" with two gene path instances; g_2 and g_3 share gene meta path "$G - T - G$" with one gene path instance. Noting that g_4 does not belong to $\mathcal{N}(g_1, g_2)$, because none of the gene meta paths between g_2 and g_4 or between g_1 and g_4 exists in $\mathcal{PS}(g_1, g_2)$.

However, to answer the question "Is gene g_1 more similar to gene g_2 than to gene g_3?", it is unreasonable to compare $pSim_{\mathcal{P}}(g_1, g_2)$ and $pSim_{\mathcal{P}}(g_1, g_3)$ directly. The reason is that the gene meta paths connecting g_1 to g_2 may be different from those between g_1 and g_3.

To tackle this issue, we propose the solution of using rank statistics. Specifically, given a query pair of genes (g_1, g_2), and a gene meta path set $\mathcal{P}$, we rank the gene pairs generated from $\mathcal{N}(g_1, g_2)$ in their similarity descending order. The *similarity rank* of (g_1, g_2) on gene meta path set $\mathcal{P}$, denoted by $rank_{\mathcal{P}}(g_1, g_2)$, is

$$\begin{aligned}rank_{\mathcal{P}}(g_1, g_2) = |\{&(g_i, g_j) \mid pSim_{\mathcal{P}}(g_i, g_j) > pSim_{\mathcal{P}}(g_1, g_2), \\ &g_i \in \{g_1, g_2\}, g_j \in \mathcal{N}(g_1, g_2)\}| + 1\end{aligned} \tag{3}$$

The smaller the rank value is, the more significant of the similarity between g_1 and g_2 is comparing to the other gene pairs generated from $\mathcal{N}(g_1, g_2)$ on gene meta path set $\mathcal{P}$.

Definition 5 (Similar Aspect). Given two query genes, g_1 and g_2, gene meta path set $\mathcal{P}$ is a similar aspect for g_1 and g_2 returned by

$$\underset{\mathcal{P} \subseteq \mathcal{PS}(g_1, g_2)}{argmin} \quad rank_{\mathcal{P}}(g_1, g_2)$$

$\mathcal{P}$ is *minimal* if there does not exist another similar aspect $\mathcal{P}' \subset \mathcal{P}$ such that $rank_{\mathcal{P}'}(g_1, g_2) \leq rank_{\mathcal{P}}(g_1, g_2)$.

Given a pair of query genes (g_1, g_2), the problem of **mining similar aspects** is to find all minimal similar aspects between g_1 and g_2.

4 The Proposed *SCENARIO* Approach

To address the limitation in lack of explanations in the existing gene similarity analysis, we proposed a novel approach, *SCENARIO*, to mine the similar aspect between gene pair, which can provide the explanation for gene similarity. In this section, we introduce the framework of *SCENARIO*, which consists of four main parts:

- **Gene Information Network Construction**: Multiple gene-related data sources are fused to construct a gene information network, where the similarity analysis is performed. (Section 4.1)
- **Candidate Similar Aspect Generation**: The gene meta paths between query gene pair are searched to generate the candidate similar aspects on the constructed gene information network. (Section 4.2)
- **Mining Similar Aspects**: The rank statistics is used to mine the similar aspects between the query gene pairs (Section 4.3) and among multiple genes (Section 4.4).

4.1 Gene Information Network Construction

To address the problem of mining similar aspects between query genes, the key precondition of *SCENARIO* is the construction of the *gene information network*.

In this paper, we use data from six databases to construct the GIN, which are listed in Table 1. These six gene databases provide the associations between seven types of biomedical entities. We construct the GIN by connecting all seven relations via shared nodes.

- *Gene-Protein relation:* The protein is extracted from DrugBank [32], and the according gene-protein association is further obtained from HGNC [33]. The link between gene and protein denotes the “encoding” or “encoded-by” relations.
- *Drug-Protein relation:* The drug-protein relationship is obtained from DrugBank database. The link between drug and protein denotes the “targeting” or “targeted-by” relations.
- *GO term-Gene relation:* The GO term-gene relationship is obtained from GOA [34] and NCBI Gene [35]. The link between GO term and gene denotes the “annotating” or “annotated-by” relations.
- *Gene-Disease relation:* The gene-disease relationship is extracted from OMIM [36]. The link between gene and diseased denotes the “causing” or “caused-by” relations.
- *miRNA-Gene relation:* The miRNA-gene relationship is obtained from miRNet [37]. The link between miRNA and gene denotes the “targeting” or “Targeted-by” relations.
- *miRNA-Disease relation:* The miRNA-disease relationship is obtained from miRNet. The link between miRNA and disease denotes the “causing” or “caused-by” relations.

TABLE 1
Gene Database

Date Source	URL
DrugBank [32]	https://www.drugbank.ca/releases/latest
HGNC [33]	https://www.genenames.org/download/statistics-and-files
GOA [34]	http://geneontology.org/docs/downloads
NCBI Gene [35]	ftp://ftp.ncbi.nih.gov/gene/DATA
OMIM [36]	https://omim.org/downloads
miRNet [37]	https://www.mirnet.ca/miRNet/docs/Resources.xhtml

- *Disease-Phenotype relation:* The disease-phenotype relationship is extracted from OMIM. The link between disease and phenotype denotes the “including” or “included-by” relations.

4.2 Candidate Similar Aspect Generation

As stated in Definition 5, a similar aspect is a set of gene meta paths. Therefore, given a pair of query genes, before mining the similar aspect between them, it is necessary to find all gene meta paths between them. According to Section 3, the gene meta path which is used to measure the similarity between peer genes is symmetric.

Firstly, in order to accelerate the search for all gene meta paths between the query gene pair, the maximum gene meta path length needs to be given. Different from [30] that directly gives a certain maximum gene meta path, we use a flexible way to automatically determine the maximum gene meta path length between the given genes, which reduces unnecessary redundancy. For the sake of clarity, we denote by ℓ the maximum length of gene meta paths in $\mathcal{PS}(g, g')$. i.e.,

$$\ell = \max\{|P| \mid P \in \mathcal{PS}(g, g')\}$$

In order to optimize the maximum gene meta path length detection process, the tree structure is introduced. *SCENARIO* constructs two trees with two query genes as root nodes by scanning the GIN. Specifically, for two genes g_1 and g_2 taken as roots, by traversing the GIN, the adjacent nodes of g_1 and g_2 are gotten and added to the children nodes of g_1 and g_2, respectively. Then the adjacent nodes of these children nodes in the GIN, are added to their children nodes, and the added children nodes must satisfy that they are different from the previous nodes and are non-gene type nodes. Repeat the above steps until no children nodes can be added. The maximum length of the gene meta path is obtained by comparing whether the children nodes of the two trees have the same nodes. The depth of the deepest layer, which concludes the common children nodes of both trees, equals to $(\ell - 1)/2$.

For g_1 and g_2 in Figure 1, an example of the maximum gene meta path length detection is given in Figure 2, which is based on the tree structure.

Example 4. As shown in Figure 2, for g_1 and g_2, they are taken as roots to construct two trees by traversing the GIN shown in Figure 1. The adjacent nodes of g_1 and g_2 are added in these two trees as children nodes, respectively. In these two trees, the previous two layers of children

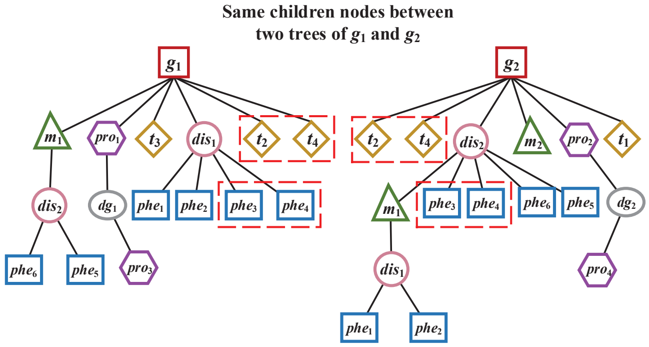


Fig. 2. An example of the maximum gene meta path length detection based on a tree structure.

nodes have the same nodes, which are $\{t_2, t_4\}$ in the first layer and $\{phe_3, phe_4\}$ in the second layer, and there is no same node in latter layers. So the maximum depth is 2, then we can get the maximum length of the gene meta path between g_1 and g_2 is $\ell = 5$, where $\ell = (2 \times 2) + 1$.

After obtaining the maximum gene meta path length ℓ, *SCENARIO* searches for the gene path instance set between g_1 and g_2 starting from themselves, respectively, by traversing the GIN in a breadth-first manner under the constraint of ℓ, the gene meta paths and the corresponding gene path instances are obtained at the same time. During the traversal process, for each gene meta path $P \in \mathcal{PS}(g_1, g_2)$, the gene path instance sets $Ins(P_{g_1 \to g_1})$, $Ins(P_{g_2 \to g_2})$, and $Ins(P_{g_1 \to g_2})$ will be found.

For the obtained $\mathcal{PS}(g_1, g_2)$, a large number of subsets that are combined by various gene meta paths between g_1 and g_2 can be generated. In order to ensure all gene meta path sets between them can be found, *SCENARIO* adopts the set enumeration tree approach [38], which has been widely used in many data mining methods, to enumerate all gene meta path sets systematically. By doing so, all gene meta path sets between g_1 and g_2, i.e., their candidate similar aspects, can be obtained. According to Equation 2, the similarity between g_1 and g_2 in each candidate similar aspect will be calculated.

4.3 Mining Similar Aspects between Pairwise Genes

SCENARIO discovers similar aspects by comparing the similarity between the query gene pairs and their related gene pairs in their neighborhood. So, *SCENARIO* first finds the comparable gene pairs in the gene neighborhood for the query genes.

Specifically, given a pair of query genes (g_1, g_2), *SCENARIO* firstly travels all gene nodes to find their gene pair neighborhood according to Definition 4. Then, the genes in $\mathcal{N}(g_1, g_2)$ are paired with either one of the query genes to generate gene pairs, which are regarded as the comparable gene pairs.

Then, *SCENARIO* calculates the similarity of each gene pair on each candidate similar aspect by Equation 2. To improve the efficiency, the parallel strategy is adopted, where *SCENARIO* introduces the thread pool to accelerate the calculation speed.

Algorithm 1 SCENARIO($g_1, g_2, geneDB$)

```
Require: $g_1, g_2$: two query genes, $geneDB$: the gene-related data sources
Ensure: $\mathcal{P}$: minimal similar aspect between $g_1$ and $g_2$
$\mathbb{G} \leftarrow$ the gene information network constructed by $geneDB$;
$\mathcal{PS}(g_1, g_2) \leftarrow$ the set of gene meta paths between $g_1$ and $g_2$ on $\mathbb{G}$;
$\mathcal{N}(g_1, g_2) \leftarrow$ the gene pair neighborhood of $g_1$ and $g_2$;
$pairSet \leftarrow \{(g_i, g_j) \mid g_i \in \{g_1, g_2\}, g_j \in \mathcal{N}(g_1, g_2)\}$;
$minRank \leftarrow |pairSet|$;
for each candidate similar aspect $\mathcal{P}' \subseteq \mathcal{PS}(g_1, g_2)$ searched by traversing the set enumeration tree do
    for each gene pair $(g_i, g_j)$ in $pairSet$ do
        compute $pSim_{\mathcal{P}'}(g_i, g_j)$;
    end for
    if $rank_{\mathcal{P}'}(g_1, g_2) < minRank$ then
        $minRank \leftarrow rank_{\mathcal{P}'}(g_1, g_2)$;
        $\mathcal{P} \leftarrow \mathcal{P}'$;
    end if
    if $rank_{\mathcal{P}'}(g_1, g_2) = minRank$ and $\mathcal{P}' \subset \mathcal{P}$ then
        $\mathcal{P} \leftarrow \mathcal{P}'$;
    end if
end for
return $\mathcal{P}$
```

Since the similarity of each candidate similar aspect obtained is a score, it is not reasonable to compare gene pairs directly by their similarity scores, so we introduce the rank statistics. Then, *SCENARIO* compares the similarities of these gene pairs under the rank statistics, and for the candidate similar aspect $\mathcal{P}$, if the similarity rank of g_1 and g_2 on $\mathcal{P}$ is minimum and there is no subset of $\mathcal{P}$ in which the similarity rank is less than or equal to $\mathcal{P}$, then $\mathcal{P}$ is the similar aspect.

The details of *SCENARIO* are described in Algorithm 1.

4.4 Mining Similar Aspects among Multiple Genes

SCENARIO can be flexibly used to mine similar aspects among multiple genes. For a set of query genes, denoted by $\mathcal{G}$, *SCENARIO* mines the similar aspect of $\mathcal{G}$ based on the aspect similarity of its paired genes.

For a query gene set $\mathcal{G}$, any two genes in $\mathcal{G}$ are paired firstly. Then, *SCENARIO* searches all gene meta paths between each gene pair (g_i, g_j), where $g_i, g_j \in \mathcal{G}$, respectively. The set of all gene meta paths of $\mathcal{G}$ is

$$\mathcal{PS}(\mathcal{G}) = \bigcap_{g_i, g_j \in \mathcal{G}} \mathcal{PS}(g_i, g_j)$$

It is worth noting that if $\mathcal{PS}(\mathcal{G})$ is empty, *SCENARIO* will stop and indicate that there is no common gene meta path among $\mathcal{G}$.

And the gene pair neighborhood of $\mathcal{G}$ under the gene meta path set $\mathcal{PS}(\mathcal{G})$ is denoted as:

$$\mathcal{N}(\mathcal{G}) = \bigcup_{g_i, g_j \in \mathcal{G}} \mathcal{N}(g_i, g_j)$$

Example 5. As shown in Figure 1, for a set of query genes $\mathcal{G} = \{g_1, g_2, g_3\}$, the set of gene pairs generated from $\mathcal{G}$ is

TABLE 2
Characteristics of the Gene Information Network

Relation	Type of Nodes	# Nodes	# Edges	Source
Gene−Protein relation	genes	20209	20209	HGNC [33]
	proteins	20075		
Drug−Protein relation	drugs	5592	15567	DrugBank [32]
	proteins	2796		
GO term−Gene relation	genes	20629	287460	GOA [34] and NCBI Gene [35]
	GO terms	18265		
Gene−Disease relation	genes	3380	4627	OMIM [36]
	diseases	4284		
miRNA−Gene relation	miRNAs	2596	320372	miRNet [37]
	genes	14736		
miRNA−Disease relation	miRNAs	684	5550	miRNet [37]
	diseases	98		
Disease−Phenotype relation	diseases	4353	92153	OMIM [36]
	phenotypes	44362		

$\{(g_1,g_2),(g_1,g_3),(g_2,g_3)\}$. And $\mathcal{PS}(\mathcal{G}) = \{``G{-}T{-}G"\}$, $\mathcal{N}(\mathcal{G}) = \{g_1,g_2,g_3,g_5\}$.

Secondly, *SCENARIO* generates all candidate similar aspects from $\mathcal{PS}(\mathcal{G})$ by using the set enumeration tree approach. After that, *SCENARIO* computes the similarity of these gene pairs, which are generated from $\mathcal{N}(\mathcal{G})$, on each candidate similar aspect $\mathcal{P} \subseteq \mathcal{PS}(\mathcal{G})$. For each candidate similar aspect $\mathcal{P}$, the similarity ranks of $\mathcal{G}$ on it is

$$rank_{\mathcal{P}}(\mathcal{G}) = max\{rank_{\mathcal{P}}(g_i,g_j)\} \tag{4}$$

where $g_i, g_j \in \mathcal{G}$. Here, the highest rank is taken because by doing so, any two genes in $\mathcal{G}$ can surely share some similarities on $\mathcal{P}$.

Finally, the similar aspect for a query gene set $\mathcal{G}$ is returned by

$$\underset{\mathcal{P}}{argmin}\, rank_{\mathcal{P}}(\mathcal{G}). \tag{5}$$

where $\mathcal{P}$ is *minimal* if there does not exist another similar aspect $\mathcal{P}' \subset \mathcal{P}$ such that $rank_{\mathcal{P}'}(\mathcal{G}) \leq rank_{\mathcal{P}}(\mathcal{G})$.

5 Experiments, Results and Discussion

In this section, we evaluate the ability and performance of *SCENARIO* in answering the similar aspect between query genes in the complicated gene information network.

5.1 Experimental Setup

We constructed the gene information network based on seven relations from six gene databases. Table 2 lists the characteristics of the seven relations that were used to compose the gene information network. In summary, there are 135,716 nodes (including 21,726 gene nodes) and 745,875 edges.

All experiments were conducted on a PC with an Intel Xeon E5-2678 v3 2.50 GHz CPU and 64 GB main memory, running the Ubuntu 19.04. All algorithms were implemented in Python and compiled by Python 3.7.

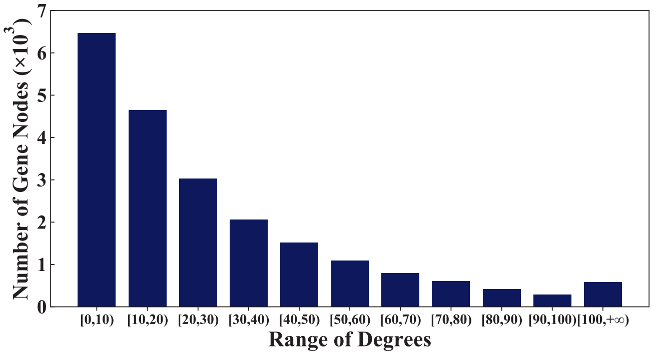


Fig. 3. The distribution of the degrees of all gene nodes.

5.2 Network Quality Analysis

To evaluate the quality of the gene information network, we used the following two measures.

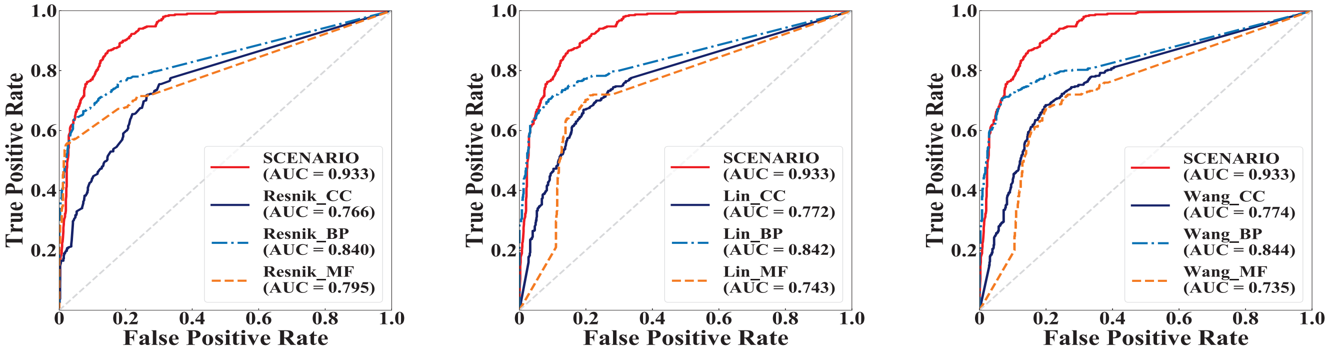
- *Distribution of gene node degrees* : the degree of a node in a gene information network represents the information content. The larger the degree of a gene node, the more informative the network is.
- *Number of gene meta paths*: the number of gene meta paths in a gene information network evaluates the richness of semantic information contained in the gene information network. The larger number of gene meta paths in a gene information network is, the more informative the network is.

Figure 3 shows the distribution of the degrees of all gene nodes in the network. We can see that the degrees of 29.7% gene nodes are less than 10 and the degrees of 25.4% gene nodes are larger than 50. In other words, most gene nodes have at least one edge connecting other nodes.

Table 3 lists the gene meta paths found by *SCENARIO* from the gene information network. There are 10 gene meta paths in total. We can see that the number of gene pairs and the number of meta path instances change a lot. For "*Gene-Disease-Gene*", there are only 18 gene pairs and 20 path instances. However, for "*Gene-miRNA-Disease-Phenotype-Disease-miRNA-Gene*", there are 300,333 gene pairs and 67,060,250 path instances. From Table 3, we can also see that it is unfair to compare the similarities of gene pairs based on different meta paths. Thus, as introduced in Section 3, *SCENARIO* used similarity rank to find similar aspects for the query gene pair.

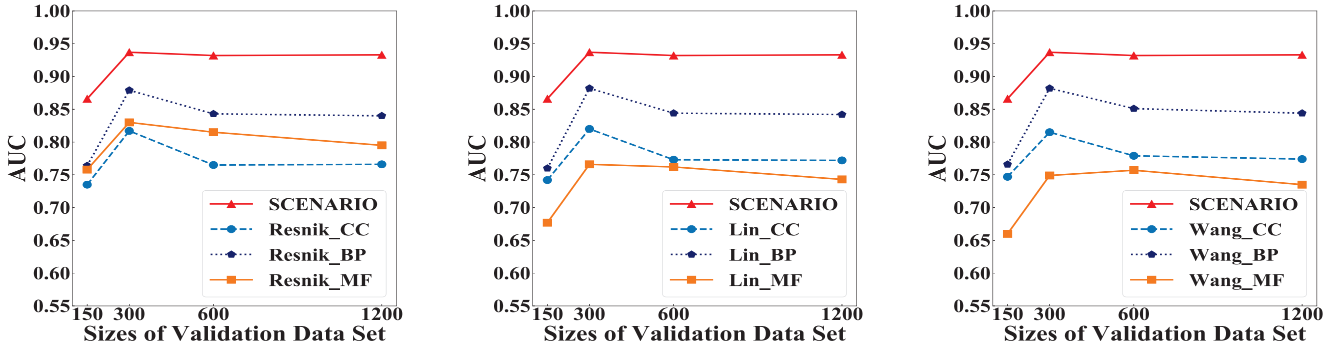
5.3 Effectiveness

As stated in [39], using EC (Enzyme Commission) number to evaluate the gene similarity is a fast and effective way, as genes annotated by the same EC number are functionally similar. In this study, the EC number was used as the label of genes, and from those with the same EC number, 400 gene pairs were randomly selected as the positive samples, i.e., similar gene pairs. And by randomly sampling from the whole gene set excluding nodes annotated by the EC number, a random gene set containing 800 gene pairs was produced, regarded as the negative samples, i.e., dissimilar



(a) *SCENARIO* compared with Resnik (b) *SCENARIO* compared with Lin (c) *SCENARIO* compared with Wang

Fig. 4. Performance analysis of *SCENARIO* compared with Resnik, Lin and Wang on three Gene Ontologies.



(a) *SCENARIO* compared with Resnik (b) *SCENARIO* compared with Lin (c) *SCENARIO* compared with Wang

Fig. 5. Performance analysis of *SCENARIO* compared with Resnik, Lin and Wang on different sizes of validation data set.

TABLE 3
The statistics of meta paths found from SCENARIO

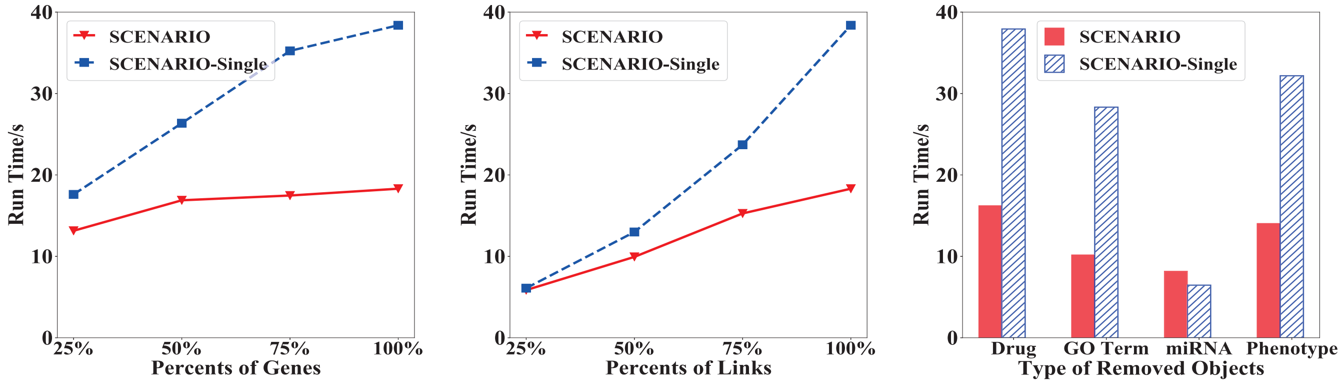
Gene Meta Path	# Pairs	# Instances
"Gene-GO Term-Gene"	250845	762948
"Gene-miRNA-Gene"	272767	2248693
"Gene-Disease-Gene"	18	20
"Gene-Protein-Drug-Protein-Gene"	1919	2477
"Gene-miRNA-Disease-miRNA-Gene"	308640	45477398
"Gene-Disease-miRNA-Disease-Gene"	67	5986
"Gene-Disease-Phenotype-Disease-Gene"	6627	17662
"Gene-miRNA-Disease-Phenotype-Disease-miRNA-Gene"	300333	67060250

gene pairs, since the coincidence that two genes sampled randomly just happen to be similar is at a very low probability. Throughout all the experiments, *SCENARIO* was applied on the complete gene information network of all the 21,726 gene nodes, and we used the validation data set of 1,200 gene pairs containing the aforementioned 400 positive samples and 800 negative samples to verify the effectiveness of *SCENARIO*.

As GO provides the precise and informative description of gene terms, *SCENARIO* was further compared with three ontology-based methods on the same dataset GO, including two information content-based methods, i.e., Lin's method [23] and Resnik's method [22], and one graph-based method, i.e., Wang's method [25]. It should be noted that GO consists of three orthogonal ontologies, i.e., the cellular component (CC), biological process (BP), and molecular function (MF), each containing an independent system of ontological terms. These three methods under comparison are all based on the relationships among GO terms, and terms in different ontologies are not comparable. Thus, the experiments were carried out on the three ontologies, respectively, when compared with these methods. All the comparison experiments were conducted on the same 1200 gene pairs, and the similarity scores of the three methods were obtained by the tool GOSemSim [40].

Figure 4 shows the ROC (Receiver Operating Characteristic) curves drawn for *SCENARIO* and the other three methods. For the results produced by Lin's method, we referred "Lin_CC" to the result from the "cellular component" ontology, "Lin_BP" to the result from the "biological process" ontology and "Lin_MF" to the result from the "molecular function" ontology, and the same was to Resnik's method and Wang's method. As shown in Figure 4, *SCENARIO* achieved the best performance among all methods with the AUC (Area Under roc Curve) score of 0.933.

To assess the influence of random sampling and the size of validation set, we further generated three validation sets in different sizes as of 150 gene pairs, 300 gene pairs and 600 gene pairs, respectively, by the same sampling strategy as mentioned in the previous context, while the



(a) Runtime w.r.t. the number of genes

(b) Runtime w.r.t. the number of links

(c) Runtime w.r.t. the number of object types

Fig. 6. The efficiency of *SCENARIO* w.r.t. the parallel strategy.

other procedure remained the same. Figure 5 illustrates that with the increase of the validation data set size, the performance of every method tends to be stable. We can see that the fluctuation brought by random sampling is decayed when the size of the validation set is larger than 600. It is worth noting that *SCENARIO* achieves the best AUC scores compared with other methods.

5.4 Efficiency

As stated in Section 4.3, to improve the efficiency of *SCENARIO*, we adopted the parallel strategy. Specifically, we set the number of threads as 16 to run *SCENARIO*. Then, we compared the runtime of *SCENARIO* with the runtime of *SCENARIO* without the parallel strategy, which is called as *SCENARIO-single*.

First, we evaluated how the runtime changes with the number of genes in GIN. We randomly chose 25%, 50%, 75% and 100% genes to construct GIN, respectively. For each GIN, we randomly chose 50 pairs of genes as query gene pairs, and reported average runtime for all results. As shown in Figure 6(a), as the number of genes increases, the scale of GIN grows, so the runtime of *SCENARIO* and *SCENARIO-single* both increases. However, *SCENARIO* is substantially faster than *SCENARIO-single*.

Second, we evaluated how the runtime changes with the number of links in GIN. For GIN constructed in Section 4.1, we randomly removed 75%, 50%, 20% and 0% of all links (reserved 25%, 50%, 75% and 100% of all links). Similarly, we randomly chose 50 pairs of genes as query gene pairs, and reported average runtime for all results. As shown in Figure 6(b), as the number of removed genes increases, the number of gene meta path instances decreases, thus the runtime of *SCENARIO* and *SCENARIO-single* both decreases. However, *SCENARIO* is substantially faster than *SCENARIO-single*.

Third, we evaluated how the runtime changes with the number of object types in GIN. For GIN constructed in Section 4.1, we deleted drug-type, GO term-type, miRNA-type and phenotype-type nodes, respectively. Once again, we randomly chose 50 pairs of genes as query gene pairs, and reported average runtime for all results. As shown in Figure 6(c), compared to GIN with all object types, due to the decrease of gene meta paths, *SCENARIO* and *SCENARIO-single* both run faster on GIN without a object type. However, *SCENARIO* is substantially faster than *SCENARIO-single*. It is worth noting that when miRNA objects are deleted, *SCENARIO* has no obvious advantage over *SCENARIO-single*. The main reason is that the gene pairs associated with miRNA objects account for the majority on the GIN, which can be found in Table 3. So when miRNA objects are removed, the search space will be greatly reduced, thus the runtime will also be greatly decreased, but *SCENARIO* needs thread scheduling time compared to *SCENARIO-single*, which causes *SCENARIO* to spend more time in thread scheduling than the reduced search time by using multi-thread.

5.5 Case Study

Three genes *GFER*, *QSOX2*, *QSOX1* were selected as the query genes for evaluating the usefulness of similarity aspect mining. Table 4 presents the candidate similarity aspects as well as the similarity ranks for the three query genes.

For *GFER* and *QSOX2*, the minimal similar aspect is $\{``G-T-G", ``G-M-G"\}$ with similarity rank = 8. It indicates that the similarity between *GFER* and *QSOX2* is the most significant in both “$G-T-G$” and “$G-M-G$”. The reason is that they are targeted by the same miRNA, and they are annotated by the same GO term at the same time. It is interesting to see that the similarity ranks are 41 and 57 when the similarity between *GFER* and *QSOX2* are evaluated by “$G-T-G$” and “$G-M-G$”, respectively. Thus, similarity aspect discovery can give more insights into the gene relations compared with similarity comparison on a single meta path.

For *QSOX1* and *QSOX2*, the minimal similar aspect is “$G-T-G$” with similarity rank = 2. They are similar because that they are annotated by the same GO term. As stated in GeneCards [41], *QSOX2* is an important paralog of *QSOX1*, and vice verse. We can see that for similar gene pairs, *SCENARIO* can easily capture their similarity.

For *GFER* and *QSOX1*, the minimal similar aspect is “$G-T-G$” with similarity rank = 85. Clearly, 85 is a considerably large similarity rank, which means that many gene pairs are more similar than the query genes in this aspect. However, *SCENARIO* still finds the aspect in which *GFER*

TABLE 4
Statistics on the similarity ranks w.r.t. genes *GFER*, *QSOX2* and *QSOX1*

Query Genes	Candidate Similar Aspects	Similarity Rank	Size of Gene Pair Neighborhood
(*GFER*, *QSOX2*)	{*G-T-G*}	41	17672
	{*G-M-G*}	57	
	{*G-M-Dis-M-G*}	2905	
	{*G-T-G, G-M-G*}	**8**	
	{*G-T-G, G-M-Dis-M-G*}	3844	
	{*G-M-G, G-M-Dis-M-G*}	2314	
	{*G-T-G, G-M-G, G-M-Dis-M-G*}	580	
(*QSOX2*, *QSOX1*)	**{*G-T-G*}**	**2**	16312
	{*G-M-G*}	1097	
	{*G-M-Dis-M-G*}	581	
	{*G-M-Dis-Phe-Dis-M-G*}	548	
	{*G-T-G, G-M-G*}	320	
	{*G-T-G, G-M-Dis-M-G*}	157	
	{*G-M-G, G-M-Dis-M-G*}	1837	
	{*G-T-G, G-M-Dis-Phe-Dis-M-G*}	163	
	{*G-M-G, G-M-Dis-Phe-Dis-M-G*}	815	
	{*G-M-Dis-M-G, G-M-Dis-Phe-Dis-M-G*}	432	
	{*G-T-G, G-M-G, G-M-Dis-M-G*}	589	
	{*G-T-G, G-M-G, G-M-Dis-Phe-Dis-M-G*	232	
	{*G-T-G, G-M-Dis-M-G, G-M-Dis-Phe-Dis-M-G*}	129	
	{*G-M-G, G-M-Dis-M-G, G-M-Dis-Phe-Dis-M-G*}	741	
	{*G-T-G, G-M-G, G-M-Dis-M-G, G-M-Dis-Phe-Dis-M-G*}	215	
(*GFER*, *QSOX1*)	**{*G-T-G*}**	**85**	16757
	{*G-M-Dis-M-G*}	2962	
	{*G-T-G, G-M-Dis-M-G*}	2047	
{*GFER*, *QSOX1*, *QSOX2*}	**{*G-T-G*}**	**85**	17903
	{*G-M-Dis-M-G*}	2962	
	{*G-T-G, G-M-Dis-M-G*}	3844	

and *QSOX1* are similar to each other. Thus, *SCENARIO* is flexible to find the similarity between somehow dissimilar gene pairs.

For the gene set consists of *GFER*, *QSOX1* and *QSOX2*, the similar aspect among them is $\{“G - T - G”\}$ with similarity rank = 85, which is a considerably large similarity rank. This is because the similarity between *GFER* and *QSOX1* is not significant, which affects the similarity between the three genes. It indicates that *SCENARIO* does not ignore the influence of the dissimilar aspects between query genes. However, the mined similar aspect $\{“G - T - G”\}$ also indicates that there are a certain of same GO terms annotating *GFER*, *QSOX1* and *QSOX2* at the same time. Thus, *SCENARIO* is flexible to make some slight adjustment to itself regarding to different query genes by their characteristics.

TABLE 5
The similar aspects w.r.t. gene pair (*ABCC8*, *KCNJ11*) and gene set {*ABCC8*, *KCNJ11*, *GCK*}

Query Genes	Similar Aspects	Similarity Rank	Size of Gene Pair Neighborhood
(*ABCC8*, *KCNJ11*)	{*G-Dis-G, G-Pro-Dg-Pro-G*}	1	15003
	{*G-Dis-G, G-T-G*}	1	
	{*G-Dis-Phe-Dis-G, G-Pro-Dg-Pro-G*}	1	
	{*G-Dis-Phe-Dis-G, G-T-G*}	1	
	{*G-Dis-G, G-T-G G-Dis-Phe-Dis-Phe-Dis-G*}	1	
{*ABCC8*, *KCNJ11*, *GCK*}	{*G-Dis-G*}	3	17503
	{*G-Dis-Phe-Dis-G*}	3	

Based on above, we can see that similar aspect mining is useful to capture and explain the similarities among genes, and the proposed method, *SCENARIO*, is effective.

5.6 Pathway Enrichment Analysis

We also used pathway enrichment analysis as part of our evaluation of the meaningfulness of our method. Known pathways from KEGG [42] were used. We performed pathway enrichment analysis at KOBAS 3.0[1] [43], which is a widely used gene set enrichment analysis tool.

A NIDDM-related gene set, consists of *ABCC8, GCK, IRS1, IRS2, INSR, PDX1, KCNJ11, SLC2A2, SLC2A4* and *SOCS1*, was selected to perform *SCENARIO* and pathway enrichment analysis. Please note that the whole genome of human was used as background genes in pathway enrichment analysis.

From the results of *SCENARIO*, there are some gene pairs with rank-1 similar aspects in the NIDDM-related gene set, such as gene pair (*ABCC8*, *KCNJ11*). These discovered strong relationships can be their similarity explanation. As listed in Table 5, the gene pair (*ABCC8*, *KCNJ11*) has five rank-1 similar aspects. This shows that the similarity between genes *ABCC8* and *KCNJ11* can be explained from five perspectives. For example, the similar aspect $\{“G - Dis - G”, “G - Pro - Dg - Pro - G”\}$ indicates one of the most significant relationships between *ABCC8* and *KCNJ11*. That is, they not only cause the same disease, e.g, NIDDM, but also have the same targeted drug, e.g., glimepiride. And this relationship can be used to explain the similarity between *ABCC8* and *KCNJ11*.

By searching the GIN, we found that there is a common disease caused by *ABCC8* and *KCNJ11*, which is PDMI (*permanent neonatal diabetes mellitus*). And we also found three common targeted drugs between *ABCC8* and *KCNJ11*, i.e., glimepiride, glyburide and tolazamide. According to DrugBank [32], all of these three drugs are used for the management of type 2 diabetes mellitus.

In addition, Table 6 lists the results of pathway enrichment analysis on the NIDDM-related gene set. We can see

[1] http://kobas.cbi.pku.edu.cn/kobas3

TABLE 6
Pathways significantly enriched in NIDDM pathogenic genes

Pathway	Genes	p-value
Type II diabetes mellitus	***KCNJ11***, *IRS1, IRS2,* *SOCS1, PDX1* *INSR,* ***ABCC8***, *SLC2A4, SLC2A2*	4.03E-27
Insulin resistance	*INSR, SLC2A4, IRS1,* *IRS2, SLC2A2*	7.41E-12
Insulin signaling pathway	*INSR, IRS1, IRS2,* *SLC2A4, SOCS1*	2.36E-11
Insulin secretion	*PDX1,* ***KCNJ11***, ***ABCC8***, *SLC2A2*	1.27E-9
AMPK signaling pathway	*INSR, SLC2A4,* *IRS1, IRS2*	4.64E-9
FoxO signaling pathway	*INSR, SLC2A4,* *IRS1, IRS2*	6.74E-9
Regulation of lipolysis in adipocytes	*INSR, IRS1, IRS2*	1.24E-7
Longevity regulating pathway - multiple species	*INSR, IRS1, IRS2*	1.75E-7
Adipocytokine signaling pathway	*SLC2A4, IRS1, IRS2*	2.38E-7
Longevity regulating pathway	*INSR, IRS1, IRS2*	5.01E-7
Non-alcoholic fatty liver disease (NAFLD)	*INSR, IRS1, IRS2*	2.28E-6
cGMP-PKG signaling pathway	*INSR, IRS1, IRS2*	3.19E-6
Maturity onset diabetes of the young	*PDX1, SLC2A2*	2.88E-5
MicroRNAs in cancer	*SOCS1, IRS1, IRS2*	1.78E-5
Aldosterone - regulated sodium reabsorption	*INSR, IRS1*	2.09E-5
Prolactin signaling pathway	*SOCS1, SLC2A2*	7.20E-5
Autophagy - animal	*IRS1, IRS2*	2.35E-4
mTOR signaling pathway	*INSR, IRS1*	3.33E-4
MAPK signaling pathway	*INSR,* ***GCK***	1.21E-3
PI3K-Akt signaling pathway	*INSR, IRS1*	1.73E-3
Carbohydrate digestion and absorption	*SLC2A2*	7.96E-3
ABC transporters	***ABCC8***	8.13E-3
Ovarian steroidogenesis	*INSR*	8.84E-3
Central carbon metabolism in cancer	*SLC2A2*	1.24E-2
Adherens junction	*INSR*	1.29E-2
Glucagon signaling pathway	*SLC2A2*	1.88E-2
HIF-1 signaling pathway	*INSR*	1.94E-2
Toxoplasmosis	*SOCS1*	2.00E-2
Neurotrophin signaling pathway	*IRS1*	2.11E-2
Osteoclast differentiation	*SOCS1*	2.27E-2
Ubiquitin mediated proteolysis	*SOCS1*	2.42E-2
Phospholipase D signaling pathway	*INSR*	2.61E-2
Jak-STAT signaling pathway	*SOCS1*	2.86E-2
Rap1 signaling pathway	*INSR*	3.68E-2
Ras signaling pathway	*INSR*	4.06E-2

that the pathway contains genes *ABCC8* and *KCNJ11*, called *type II diabetes mellitus*, has the lowest p-value. According to KEGG [42], type 2 diabetes mellitus is a related disease of *type II diabetes mellitus* pathway. This indicates that *ABCC8* and *KCNJ11* are strongly associated with type 2 diabetes mellitus.

Based on the above observations, the results of *SCENARIO* are meaningful in similar aspects mining for gene similarity explanation.

Moreover, from Table 5, it is interesting to see that there are two similar aspects among genes *ABCC8*, *KCNJ11*, and *GCK* with high similarity rank i.e., $\{``G - Dis - G"\}$ and $\{``G - Dis - Phe - Dis - G"\}$. This shows that there are two relationships that can be used to explain the similarity among genes *ABCC8*, *KCNJ11*, and *GCK*, which are causing the same disease and the same phenotype. However, it is difficult to find the similarity among genes *ABCC8*, *KCNJ11*, and *GCK* in causing the same phenotype from the result of pathway enrichment analysis, since *GCK* is not included in any pathway containing either *ABCC8* or *KCNJ11* in Table 6.

It is worth noting that the disease PDMI, mentioned above as the common disease caused by genes *ABCC8* and *KCNJ11*, has material basis in homozygous mutation in the glucokinase gene (*GCK*), heterozygous mutation in the *KCNJ11* and *INS* genes, or by heterozygous or homozygous mutation in the *ABCC8* gene [36]. This verifies the effectiveness of *SCENARIO*.

6 Conclusion

In recent years, gene similarity has become a hotspot in biology research, and it is informative for understanding the biological roles and functions of genes. However, most of the current methods searching similar genes lack an explanation for gene similarity. Besides, many of them evaluate gene similarity only under a single metric and only from a single data source, which lack full consideration of multiple aspects.

Here, we proposed *SCENARIO*, a general similar framework for calculating aspect similarity for a set of genes to detect their similar aspect, which can be used to explain the similarity between them. *SCENARIO* computes gene aspect similarity under the meta path-based measure, and it is novel in providing reasonable explanation of similarity between genes by exploiting the associations among multiple gene-related data.

The performance of *SCENARIO* was evaluated by the EC number metric. The high AUC suggested that *SCENARIO* is effective in discovering similar genes and the efficiency test showed that the parallel strategy employed in *SCENARIO* significantly improves the efficiency. *SCENARIO* also had good extensibility, not only can detect the similar aspect of two genes, but also mine similar aspects among multiple genes. The details of our experiment study and source codes of *SCENARIO* are available on https://github.com/ZhangYid/SCENARIO.

As for future work, we intend to focus on the following tasks. First, *SCENARIO* will be applied to other problems to further test its performance. The scalability of *SCENARIO* should be further improved so that more data sources can be integrated to promote the precision in gene similarity explanation on large-scale datasets. Second, novel similarity measures can be designed to better balance the importance of metrics and to better utilize rich multi-source information. Third, we will improve the gene information network to incorporate the heterogeneity of relationships between different biological entities into the gene similarity measure.

7 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61972268, 61572332, 81473446), the Sichuan Science and Technology Program (2020YFG0034, 2020YFS0574), and COST ACTION Open Multiscale Systems Medicine (CA15120).

References

[1] M. Liu and P. D. Thomas, "GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness," *BMC Bioinformatics*, vol. 20, no. 1, pp. 155:1–155:15, 2019.

[2] M. Brameier and C. Wiuf, "Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps," *Journal of Biomedical Informatics*, vol. 40, no. 2, pp. 160–173, 2007.

[3] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinformatics*, vol. 11, p. 562, 2010.

[4] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in Bioinformatics*, vol. 13, no. 5, pp. 569–585, 2012.

[5] G. Yu, H. Zhu, C. Domeniconi, and J. Liu, "Predicting protein function via downward random walks on a gene ontology," *BMC Bioinformatics*, vol. 16, pp. 271:1–271:13, 2015.

[6] G. Fu, J. Wang, B. Yang, and G. Yu, "NegGOA: negative GO annotations selection using ontology structure," *Bioinformatics*, vol. 32, no. 19, pp. 2996–3004, 2016.

[7] L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, "SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association," *PLOS ONE*, vol. 9, 06 2014.

[8] Z. Tian, M. Guo, C. Wang, L. Xing, L. Wang, and Y. Zhang, "Constructing an integrated gene similarity network for the identification of disease genes," *Journal of Biomedical Semantics*, vol. 8-S, no. 1, pp. 27–41, 2017.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[10] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

[11] J. Bao, R. Yuan, and Z. Bao, "An improved alignment-free model for dna sequence similarity metric," *BMC Bioinformatics*, vol. 15, p. 321, 2014.

[12] J. Zhou, P. Zhong, and T. Zhang, "A novel method for alignment-free dna sequence similarity analysis based on the characterization of complex networks," *Evolutionary Bioinformatics*, vol. 12, pp. EBO–S40 474, 2016.

[13] G. K. Mazandu, E. R. Chimusa, and N. J. Mulder, "Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 886–901, 07 2016.

[14] A. Brionne, A. Juanchich, and C. Hennequet-Antier, "ViSEAGO: a bioconductor package for clustering biological functions using gene ontology and semantic similarity," *BioData Mining*, vol. 12, no. 16, 2019.

[15] S. J. Giri and S. Saha, "Multi-view gene clustering using gene ontology and expression-based similarities," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, July 2020, pp. 1–8.

[16] J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. Walhout, "Using networks to measure similarity between genes: association index selection," *Nature Methods*, vol. 10, pp. 1169–1176, 2013.

[17] Z. Tian, M. Guo, C. Wang, X. Liu, and S. Wang, "Refine gene functional similarity network based on interaction networks," *BMC Bioinformatics*, vol. 18, no. 16, p. 550, 2017.

[18] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[19] M. Munz, S. Tönnies, W.-T. Balke, and E. Simon, "Multidimensional gene search with Genehopper," *Nucleic Acids Research*, vol. 43, no. W1, pp. W98–W103, 05 2015.

[20] Y. Wang, S. Yang, J. Zhao, W. Du, Y. Liang, C. Wang, F. Zhou, Y. Tian, and Q. Ma, "Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model," *Scientific Reports*, vol. 9, no. 1, pp. 2045–2322, 2019.

[21] F. B. I., A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout, "Using networks to measure similarity between genes: association index selection," *Nature Methods*, vol. 10, no. 12, pp. 1167–1176, 2013.

[22] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pp. 448–453.

[23] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pp. 296–304.

[24] H. Wang, H. Zheng, and F. Azuaje, "Ontology-and graph-based similarity assessment in biological networks," *Bioinformatics*, vol. 26, no. 20, pp. 2643–2644, 2010.

[25] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[26] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: meta path-based top-k similarity search in heterogeneous information networks," *PVLDB*, vol. 4, no. 11, pp. 992–1003, 2011.

[27] H. Zhipeng, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, "Meta Structure: computing relevance in large heterogeneous information networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1595–1604.

[28] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, "Semantic path based personalized recommendation on weighted heterogeneous information networks," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19–23, 2015*, pp. 453–462.

[29] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, "Meta-graph based recommendation fusion over heterogeneous information networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13-17 , 2017*, pp. 635–644.

[30] Y. Zhang, L. Duan, H. Zheng, J. Li-Ling, B. Hu, R. Qin, and C. He, "SCENARIO: discovery of similar aspects for gene similarity explanation from gene information network," in *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019*, pp. 604–609.

[31] R. Qin, L. Duan, H. Zheng, J. Li-Ling, K. Song, and Y. Zhang, "An ontology-independent representation learning for similar disease detection based on multi-layer similarity network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.

[32] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 11 2017.

[33] B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, B. Yates, and E. Bruford, "Genenames.org: the HGNC and VGNC resources in 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D786–D792, 10 2018.

[34] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D262–D266, 01 2004.

[35] G. R. Brown, V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott, and T. D. Murphy, "Gene: a gene-centered information resource at NCBI," *Nucleic Acids Research*, vol. 43, no. D1, pp. D36–D42, 10 2014.

[36] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "OMIM.org: leveraging knowledge across phenotypegene relationships," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1038–D1043, 11 2018.

[37] Y. Fan, K. Siklenka, S. K. Arora, P. Ribeiro, S. Kimmins, and J. Xia, "miRNet - dissecting miRNA-target interactions and functional

associations through network-based visual analysis," *Nucleic Acids Research*, vol. 44, no. W1, pp. W135–W141, 04 2016.

[38] R. Rymon, "Search through systematic set enumeration," in *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning*, ser. KR '92, 1992, pp. 539–550.

[39] J. Peng, Y. Wang, and J. Chen, "Towards integrative gene functional similarity measurement," *BMC Bioinformatics*, vol. 15, no. S-2, p. S5, 2014.

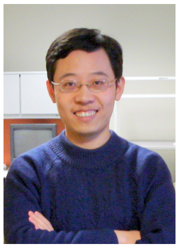
[40] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 02 2010.

[41] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug *et al.*, "GeneCards Version 3: the human gene integrator," *Database*, vol. 2010, 2010.

[42] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 11 2016.

[43] C. Xie, X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C.-Y. Li, and L. Wei, "KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W316–W322, 06 2011.

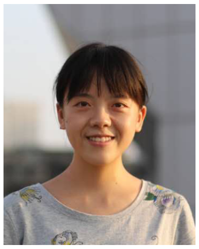
Yidan Zhang received her bachelor degrees in Biological Sciences and Software Engineering from Sichuan University in 2018. She is currently working towards her Ph.D. degree in the School of Computer Science at Sichuan University. Her research interests include bioinformatics and biomedicine.



Lei Duan received his B.Sc. and Ph.D. degrees both in Computer Science from Sichuan University in 2003 and 2008, respectively. He was a visiting Ph.D. student in the Department of Computer Science and Engineering at Wright State University from 2007 to 2008, and was a visiting scholar in the School of Computing Science at Simon Fraser University from 2012 to 2013. He is currently a Professor in the School of Computer Science at Sichuan University. His research interests include data mining, knowledge management, evolutionary computation, bioinformatics and health-informatics.

Huiru Zheng IEEE Senior Member, is a Professor of Computer Science with School of Computing at Ulster University, UK; and a Fellow of the UK Higher Education Academy. She was awarded a Ph.D. in Bioinformatics in 2003 and a Postgraduate Certificate in Teaching in Higher Education in 2005 from Ulster University. Prof. Zheng is an active researcher in bioinformatics and healthcare informatics. Within her broad interests in data mining, data integration, machine learning and healthcare decision support, Prof. Zheng has a particular research interest and expertise in integrative data analytics in the field of systems biology, and intelligent data analysis and assistive technology to support healthcare and independent living. She has published over 250 peer reviewed scientific research papers.

Jesse Li-Ling received his M.D. degree from West China University of Medical Sciences in 1993 and Ph.D. degree from University of Newcastle upon Tyne, U.K. in 2000. He conducted his postdoctoral research at Tsinghua University from 2001 to 2003. In 2007, he was promoted to full professor. Prof. Li-Ling is currently working at Sichuan University (State Key Laboratory of Biotherapy) and his main research interests include medical genetics, bioinformatics and Traditional Chinese Medicine.



Ruiqi Qin received her master degree in the School of Computer Science at Sichuan University in 2020. Her research interests include bioinformatics and data mining.

Zihao Chen received his bachelor degree in the School of Computer Science at Sichuan University in 2020. He is currently working towards his master degree in the School of Computer Science at Sichuan University. His research interests include data mining and bioinformatics.

Chengxin He received his bachelor degree in Computer Science from Sichuan University in 2018. He is currently working towards his Ph.D. degree in the School of Computer Science, Sichuan University. His research interests include data mining, bioinformatics and biomedicine.

Tingting Wang received her master degree in Computer Science from Sichuan University in 2020. She is currently working towards her Ph.D degree in the School of Computer Science & IT, RMIT University. Her research interests include data mining and database.