Multi-factored gene-gene proximity measures exploiting biological knowledge extracted from Gene Ontology : application in gene clustering

Sudipta Acharya[1], Sriparna Saha[1], Prasanna Pradhan[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Department of Computer Applications, Sikkim Manipal Institute of Technology
Corresponding author: S. Acharya (email:sudiptaacharya.2012@gmail.com)

To describe the cellular functions of proteins and genes, a potential dynamic vocabulary is Gene Ontology (GO), which comprises of three sub-ontologies namely, Biological-process, Cellular-component and Molecular-function. It has several applications in the field of bioinformatics like annotating/measuring gene-gene or protein-protein semantic similarity, identifying genes/proteins by their GO annotations for disease gene and target discovery etc. To determine semantic similarity between genes, several semantic measures have been proposed in literature, which involve information content of *GO-terms*, GO tree structure or the combination of both. But most of the existing semantic similarity measures do not consider different topological and information theoretic aspects of *GO-terms* collectively. Inspired by this fact, in this article, we have first proposed three novel semantic similarity/distance measures for genes covering different aspects of GO-tree. These are further implanted in the frameworks of well-known multi-objective and single-objective based clustering algorithms to determine functionally similar genes. For comparative analysis, ten popular existing GO based semantic similarity/distance measures and tools are also considered. Experimental results on *Mouse genome*, *Yeast* and *Human genome* datasets evidently demonstrate the supremacy of multi-objective clustering algorithms in association with proposed multi-factored similarity/distance measures. Clustering outcomes are further validated by conducting some biological/statistical significance tests.

***Index Terms*—Gene Ontology (GO), Gene clustering, Semantic similarity/distance measure, Gene-gene similarity matrix, Multi-objective clustering.**

I. Introduction

GENE Ontology (GO) is a controlled and consistent global database where knowledge about gene functions for several world's major organisms for plant, animal and microbial genomes are stored in the form of directed acyclic graphs [2]. The overall biological knowledge is stored in the form of three controlled taxonomies - Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Each of them is a complete ontology containing several processes and sub-processes which are referred as *GO-terms* having direct and indirect relationships with each other. For various organism databases, their genes are annotated with a specific set of *GO-terms* (under BP, MF and CC) and this annotation information can be downloaded from the GO website (http://www.geneontology.org/). GO is used to serve different purposes like analysing gene products and their functionalities across different organisms. One such emerging application of GO is finding semantic similarity between genes. Two genes which are semantically related, represent that they are functionally connected and are involved in similar biological, molecular and cellular functions. GO annotation information and term similarities can intuitively measure the functional similarity between genes.

Here, one trivial question arises, *Why estimation of functional similarity between genes is needed?* In the post-genomic era, one of the important goals is to discover functions of genes. DNA microarray technology helps in monitoring the expression levels of thousands of genes during important biological processes and over the collection of related samples. Automatically uncovering functionally related genes, is a basic building block to solve various problems related to functional genomics [4]. But with the increasing number of genes, analysis of data has become a challenging task. To meet with this challenge, a potential alternative is to discover the interesting patterns lying within the dataset and this process can be assisted by some clustering [1] or bi-clustering [19] techniques. The effectiveness or accuracy of any clustering/bi-clustering technique highly depends on the underlying similarity/distance measure used. But automatically determining the suitable similarity measure/distance function for clustering of genes remains an open problem in the field of functional genomics.

The aim behind clustering of gene expression data is to find co-regulated genes or genes having similar expression values i.e., genes exhibiting similar functions. Though clustering of gene expression data over a set of samples helps in finding the co-regulated genes, but this does not consider the semantic relationships exist between them. Motivated by this, in the past, several researches have been carried out on developing various semantic similarity measures using GO to determine functionally similar genes. But most of the existing measures do not consider some important properties of *GO-terms* and GO. In this paper, we have identified various mutually exclusive *GO-term* properties and then developed two novel semantic similarity and a novel semantic distance measure utilizing different GO properties. These newly developed measures in turn help in correctly detecting groups of similar genes.

The main contributions of the current work are listed below:

1. We have proposed a new topological level based semantic similarity measure named as $SIM_{norm-struct_{depth}}$.

2. Two new multi-factored semantic similarity and distance measures, namely *Multi-SIM* and *Multi-DIST*, respectively are also proposed. These are called 'multi-factored', as they consider multiple information theoretic and topological/structural aspects of GO.
3. In order to show the efficacy of the newly developed similarity or distance measures, seven existing GO based similarity / distance measures and tools are considered for comparison.
4. Finally, these ten (out of which three are our proposed and rest seven are existing) proximity measures are used in association with two multi-objective and six single-objective based clustering algorithms to identify functionally similar genes from three benchmark gene datasets.

II. Existing works on identifying functionally similar genes

There are a number of existing semantic similarity measures proposed in the past literature. In this section, we will discuss some of them which are relevant to our work. A number of information theory based approaches exist in the literature to measure the semantic similarity between *GO-terms*. All of them try to maximize the similarity between two *GO-terms* which share more information. Therefore, to measure the shared information between *GO-terms*, the information content (IC) subsumed by them from GO can be utilized. For any *GO-term* t, $p(t)$ denotes probability to find a child of t in GO. Thus, according to principle of information theory, the Information content (IC) of term t is denoted by $-log(p(t))$.

One such existing information theory based similarity measure was proposed by Resnik [23]. According to this model the similarity between terms t_i and t_j is given as follows:

$$sim_{Resnik}(t_i, t_j) = max_{t \in S(t_i,t_j)}[-log(p(t))]] \quad (1)$$

where $S(t_i, t_j)$ denotes the set of parents *GO-terms* shared by both t_i and t_j. max is the maximum operator. Here, $sim_{Resnik}(t_i, t_j) \in [0, \infty]$. This measure provides the IC of Lowest Common Ancestor (LCA) of both terms t_i and t_j.

Another information-theory based similarity measure was proposed by Lin [13]. Here also similarity between two terms is calculated based on parent commonality. But here an additional information i.e., ICs of the query terms are also included in the similarity measure. Thus, given terms, t_i and t_j, their similarity may be calculated as:

$$sim_{Lin}(t_i, t_j) = \frac{2 \times max_{t \in S(t_i,t_j)}[log(p(t))]}{log(p(t_i)) + log(p(t_j))} \quad (2)$$

Here $sim_{Lin}(t_i, t_j) \in [0, 1]$. Lin's measure is normalized version of Equation 1.

One drawback of Resnik measure is that, it does not consider the distance of two terms from their LCA. As a result, pairs of terms that share the same LCA but reside in different levels of the GO tree will have same semantic similarity. But this is not a correct measure. Again in Lin's similarity measure, the relative distance of two terms from their LCA is taken into account but depth of LCA is not considered. To overcome the limitation of both Resnik and Lin's similarity measure, Schlicker et al. [25] proposed a similarity measure as defined below.

$$sim_{Schlicker}(t_i, t_j) = \frac{2 \times max_{t \in S(t_i,t_j)}[log(p(t))]}{log(p(t_i)) + log(p(t_j))} \times (1 - \frac{G_{LCA}}{G}) \quad (3)$$

Where, G_{LCA} and G are sets of genes annotated to corresponding LCA of t_i and t_j, and root of *GO-term*, respectively. The first part of the Equation 3 defines the relative distance of terms, t_i and t_j, from their LCAs. The second part represents the depth of LCA in the GO tree.

Similarity between terms can also be calculated in terms of distance measures. The more similar two terms is, the closer they would be in the distance space. Jiang's distance [10] is one such existing distance measure. It is defined as follows:

$$dist_{Jiang}(t_i, t_j) = 2 \times max_{t \in S(t_i,t_j)}[log(p(t))] - [log(p(t_i)) + log(p(t_j))] \quad (4)$$

The $dist_{Jiang}(t_i, t_j) \in [0, \infty]$ and it reflects the semantic dissimilarity or distance between two terms, t_i and t_j. The correlation between gene expression values and gene similarity measures computed by Resnik's, Jiang's and Lin's measures of semantic similarity [10], [13], [23] were analysed in Ref. [27]. More recently, authors of Ref. [26] proposed a IC based measure to compute similarity between *GO-terms* present in a GO, based on a combination of Lin's and Resnik's techniques.

Beside the above mentioned LCA based similarity/distance measures, Wang et al. [8], [30] proposed a similarity measure, which considers topological information of GO graph but does not consider gene annotation data. This measure takes into account all parent terms of two *GO-terms* instead of only considering their LCAs. For a *GO-term* t_i, let it's one parent be p. Then the semantic contribution of p to t_i, denoted as $SC_{t_i,p}$ measures the maximal semantic contribution of the paths from t_i to p. The semantic similarity between two *GO-terms* is measured as follows:

$$sim_{wang}(t_i, t_j) = \frac{\sum_{p \in P_i \cap P_j}(SC_{t_i,p} + SC_{t_j,p})}{\sum_{t \in P_i} SC_{t_i,t} + \sum_{t \in P_j} SC_{t_j,t}} \quad (5)$$

Where, P_i (or P_j) is the set of all parents of term t_i (or t_j). Utilizing this measure, authors have developed an on-line semantic similarity measurement tool named as *G-SESAME* [8], [30]. Another set based method to measure similarity between genes was proposed by Ref. [18], which is named as normalized term-overlap method [18]. The term overlap score of two genes g_1 and g_2 is calculated as the number of common *GO-terms* between annotation sets of these two genes. Then the term overlap score is divided by the annotation set size of the gene with the lower number of GO annotations. It is calculated as follows:

$$sim_{NTO}(g_1, g_2) = \frac{sim_{TO}(g_1, g_2)}{min(|annot_{g_1}|, |annot_{g_2}|)} \quad (6)$$

Here, $sim_{NTO}(g_1, g_2) \in [0, 1]$.
Utilizing the concept of Gloss Vector measure in natural language processing (NLP), the authors of Ref. [22] developed a semantic gene-gene similarity measure utilizing definitions of *GO-terms* in GO. Their proposed measure, namely simDEF - is an optimized version of Gloss Vector measure. Point wise

mutual information (PMI) is employed for the purpose of optimization. After constructing optimized definition-vectors of all *GO-terms*, the cosine of the angle between term's definition-vectors represents the degree of similarity between them.

Apart from the above discussed similarity measures, several hybrid similarity measures also exist in the literature which combine more than one aspect. For example in Shen et al. [28], authors proposed a similarity measure that takes into account both the path length between the terms as well as the IC of the ancestor terms. The distance between two terms is defined as:

$$dist_{Shen}(t_i, t_j) = \frac{arctan[\sum_{t_1 \in path_i} \frac{1}{IC[t_1]} + \sum_{t_2 \in path_j} \frac{1}{IC[t_2]}]}{\pi/2} \quad (7)$$

where $path_i$ and $path_j$ are the shortest paths that connect the terms t_i and t_j to the lowest common ancestors between them. As we move from more specific terms (higher IC) to general terms (lower IC) along the paths, the value of $\sum_{t_1 \in path_i} \frac{1}{IC[t_1]} + \sum_{t_2 \in path_j} \frac{1}{IC[t_2]}$ increases and so does the distance between two terms.

The corresponding semantic similarity between *GO-terms* was defined by Shen [28] as follows:

$$sim_{Shen}(t_i, t_j) = 1 - dist_{Shen}(t_i, t_j) \quad (8)$$

Another hybrid similarity measure was proposed in Ref. [20], named as NETSIM (network-based similarity measure). This measure incorporates information from gene co-function networks in addition of using the GO structures and annotations. Here authors have shown that incorporation of gene co-function network data clearly helps in improving the performance of the *GO-term* similarity measures. Another integrative approach to measure gene-gene similarity was proposed in Ref. [21] and the developed tool by the authors is named as *InteGO*. Here, the gene-gene similarity is measured by considering three well known existing similarity measures, namely, Yu [33], Wang [30] and Schlicker [25] and by integrating 'rank based gene-gene similarity' based on those measures. It automatically provides best integration policy (they named it as *seed measure integration*) to compute the similarity between given pair of genes. But in their proposed hybrid measure, some important topological and structural information of *GO-terms* are missing. For example, the weighted shortest path length factor between *GO-terms* by Shen [28], the level of *GO-terms* in GO tree (as considered by our proposed $SIM_{norm-struct_{depth}}$ measure), annotation sets commonality of genes [18] have not been considered by *InteGO* [21]. In the current paper, our proposed measures overcome the limitations of *InteGO* tool.

Another area of research in the field of bioinformatics is the identification of co-regulated genes [32]. As the availability of high throughput data like DNA microarray is increasing rapidly, biologists get access to a large collection of data. To get an insight of underlying biological processes, several analysis, statistical, biological tests are performed on this large volume of data. One of such broadly used analysis techniques is performing clustering on microarray gene expression data [1], [19]. The underlying hypothesis is that genes with similar expression patterns are involved in similar biological processes.

For any clustering algorithm, its efficiency majorly depends on the adopted underlying proximity measure to determine similarity between data points. Most of the existing works like Ref. [1], [19] identify co-regulated genes based on similar expression patterns. Here each gene is represented as a vector of expression values over different conditions. Similarity between two genes is calculated by using any standard distance measures like Euclidean, or symmetry based distance. Finding out similar genes in this way does not consider semantic relationships between genes. For this reason, in past few years, researchers have worked on grouping functionally similar genes based on known biological knowledge from GO.

In the field of gene clustering research using GO data, authors of Ref. [14] developed a set of methods for gene and *GO-term* analysis. In Ref. [9], authors proposed GOSim package within R environment for similarity computations of genes and for gene clustering. Here authors have represented genes as feature vectors and during clustering the similarity between feature vectors was considered as the normalized dot product between the vectors. In their developed R package for gene clustering, only hierarchical clustering can be applied on gene feature vectors. Also in gene clustering, authors of Ref. [31] used a graph based similarity measure (simUI implemented in Bioconductor in R) for computing gene similarity, and then PAM clustering algorithm is employed to group similar genes into clusters [31].

In Ref. [16], [17], authors developed a web tool named *DaGO-Fun* and later in Ref. [15] they developed its advanced version, namely, *A-DaGO-Fun* (ADaptable Gene Ontology semantic similarity based Functional analysis), which is a repository of python modules providing researchers the freedom to choose the most relevant measure for their specific application. It contains six main functions and implements 101 different functional similarity measures. Each of the eight annotation-based and three topology-based approaches, namely Resnik, XGraSM-Resnik, Nunivers, XGraSM-Nunivers, Lin, XGraSM-Lin, Relevance and Li et al. (2010) [11], Wang et al. (2007) [30], Zhang et al. (2006) [34], and GO-universal, is implemented with seven known term pairwise-based functional similarity measures: Avg, Max, ABM, BMA, BMM, HDF and VHDF. The tool *IT-GOM* within *A-DaGO-Fun* module calculates semantic similarity between *GO-terms* based on a selected semantic similarity measure as defined above. From the obtained term-term similarities using this tool, gene-gene similarities can be intuitively measured. One advantage of this tool is, it is organism and gene-ID independent. Another tool - *GOSP-FuCT* from *A-DaGO-Fun* module allows the partitioning of a gene or protein set into a set of biologically meaningful sub-classes using their functional closeness based on a selected semantic similarity measure as defined above. The tool has utilized three single-objective well known clustering techniques- Hierarchical, K-means and community detecting model [15], implemented

within it for the purpose of clustering of genes. The tool incorporates several measures (as mentioned above) within it but they do not consider various important properties of *GO-terms* collectively in a single measure. Also, as all of the incorporated clustering algorithms are based on single-objective optimization, therefore, *GOSP-FuCT* tool does not explore the utility of multi-objective optimization based clustering in identifying functionally similar genes.
In summary, different methodologies to measure similarity between genes based on existing biological knowledge from GO are available in the literature.

III. Motivation

In the existing literature, as mentioned in section II, to calculate semantic similarity or distance between two genes, different authors have considered different *GO-terms* properties of GO tree. Some of them have considered IC based similarity/distance measures [10], [13], [23], [25]. No structural information of GO tree was considered in these works for measuring *GO-term* similarity/distance. In Ref. [28], to measure the similarity/distance, the shortest path between *GO-terms* is adopted which is a topological information of GO. In Ref. [18], a set based similarity measure was proposed. Authors of Ref. [22] utilize a familiar concept from NLP to measure semantic similarity between *GO-terms*. They have neither used information theoretic nor topological/structural information of GO tree while measuring similarity. Collectively, all of these existing measures as described in section II utilized different important properties of *GO-terms*, which are complementary to each other. But a single similarity measure which conjointly considers different important mutually exclusive *GO-terms* and GO tree properties is rare in the literature. There are limited attempts reported in the literature where some hybrid similarity measures are developed by considering various attributes related to information theory and topological information of GO [20], [21], [28]. But one or more important *GO-term* properties are not considered in these existing hybrid similarity measures.
Motivated by this fact, in the current paper firstly we have identified a topological level based *GO-term* property which has not been covered by any existing measure but plays an important role in discovering semantic similarity between *GO-terms* or genes. Next, along with our newly proposed level based *GO-term* property, some more existing mutually exclusive *GO-term* properties are combined to develop two 'multi-factored' semantic similarity and distance measures.

The existing works [25], [28] acknowledge that the depth of LCA in the GO tree provides an important information for measuring similarities between *GO-terms*. These works consider the effect of depth of LCA in terms of gene annotation count [25] or IC of *GO-terms* [28]. But while measuring 'depth', the topological level of LCA in the GO tree is also one important factor to be considered. If the LCA of two *GO-terms* is located at higher level (or lower level) of the GO tree, i.e., LCA is nearer to root, then the *GO-terms* tend to have less common functional properties. It is because they are separated at the higher level. The lower level of LCA (higher depth) indicates that its descendant *GO-terms* are more functionally similar. We have named this topological level information of *GO-term* as 'Structural-depth' or in short $struct_{depth}$. Inspired by this observation, in the current work, we have proposed a normalized $struct_{depth}$ based similarity measure. A possible good measure should be a combination of $struct_{depth}$ information of different *GO-terms*, their information contents and their annotation sets commonality. Along these lines, efforts are made in the current paper to develop a multi-factored semantic similarity and multi-factored semantic distance measures. The proposed similarity and distance measures are multi-factored because of the consideration of multiple factors related to information theory and topological structure of GO as given below:
1. IC based semantic similarity between *GO-terms* according to Ref. [13] and semantic distance according to Ref. [10]. It captures information theoretic properties of *GO-terms*.
2. Shortest weighted path distance based similarity between *GO-terms* according to Ref. [28]. It captures distance based topological information.
3. Term overlap based similarity between *GO-terms* according to Ref. [18]. It captures annotation sets commonality of genes.
4. Our proposed normalized $struct_{depth}$ based similarity between *GO-terms*. It captures topological level based information.

Considering all the above mentioned *GO-term* properties in the mentioned measures, we have proposed following three similarity and distance measures.
1. Normalized $struct_{depth}$ based semantic similarity named as $SIM_{norm-struct_{depth}}$.
2. Multi-factored semantic similarity named as *Multi-SIM*.
3. Multi-factored semantic distance named as *Multi-DIST*.

Please note that, the GO structure and it's organization are getting updated and changed with the advent of new *GO-terms* and their relationships with each other for different organisms. Accordingly, GO database is also getting updated periodically. Therefore, the properties of each *GO-term* like IC, annotation data, depth, path length between *GO-terms* etc., are biased because of data availability. So, for any existing similarity measure which utilizes any of information theoretic or structural properties of *GO-terms* in GO tree, for example Ref. [23], Ref. [13], Ref. [25], Ref. [21] etc., the similarity score between two *GO-terms* or genes will be changed/updated with the modernization of GO.

As our proposed $SIM_{norm-struct_{depth}}$, *Multi-SIM* and *Multi-DIST* measures also leverage the level or depth of *GO-terms*, therefore, the gene-gene similarity values will also keep on updated based on the amendments of GO. Note that with the advancements of GO, the proposed similarity values will be capable of capturing more fine grained information. The use of depth factor in the similarity measure enables it capturing the same. For the branches which are well-studied, the proposed $SIM_{norm-struct_{depth}}$, *Multi-SIM* and *Multi-DIST* measures will be able to extract the fine-grained similarity between

two genes. Thus with the advancement in GO, developed measures will become more robust and effective in capturing gene-gene similarity. Our main motivation behind the proposed measures is to consider different mutually exclusive properties of *GO-terms* which contain different important information to identify functionally co-related genes based on most recent GO tree.

Utilizing our proposed three different similarity/distance measures, three different gene-gene similarity matrices have been generated corresponding to three chosen datasets, based on which clustering has been performed on these datasets to identify functionally similar genes. For that purpose, eight clustering algorithms, among which two algorithms are multi-objective based and rest six are single-objective optimization based traditional clustering techniques, are deployed. Two well known popular multi-objective optimization techniques - AMOSA [3] and NSGA-II [7] have been utilized to develop two clustering approaches. According to the existing literature it has been observed that AMOSA or Archived Multi Objective Simulated Annealing optimization technique [3] excels in the field of multi-objective optimization with respect to other existing optimization techniques. Also NSGA-II [7] is very popular and widely used multi-objective optimization technique. Because of these reasons we have chosen AMOSA and NSGA-II to develop clustering algorithms utilizing their optimization capabilities. The six traditional single-objective clustering algorithms selected for our experiments are K-means, K-medoids, Single-linkage hierarchical algorithm, Complete-linkage hierarchical algorithm, Average-linkage hierarchical algorithm and DBSCAN clustering algorithm. Results reveal that AMOSA [3] based clustering works best in association with the proposed multi-factored similarity and distance measures over other existing similarity/distance measures and clustering techniques. To visualize the coherence between genes produced by AMOSA based clustering, cluster profile plots are also shown. The observations are further supported by strong biological and statistical significance tests.

IV. Proposed similarity and distance measures

In this section we have mathematically described each of the proposed similarity/distance measures in detail. The definitions of proposed measures are provided below.

A. Normalized $struct_{depth}$ based semantic similarity

The notion of normalized $struct_{depth}$ based gene-gene similarity can be understood using a simple example. In Figure 1, a snapshot[1] of sub-part of GO is shown. According to the basic property of GO, the terms located nearby the root (higher level) hold less specific information compared to terms located nearby the leaves (lower level). Thus the level of a *GO-term* in GO tree is an important parameter while measuring semantic similarity between terms. If two *GO-terms* get separated in higher level, it indicates that they share less functional properties, whereas if they get separated in lower level, it is an indication that they comparatively share more functional properties.

[1]http://www.geneontology.org/page/ontology-structure

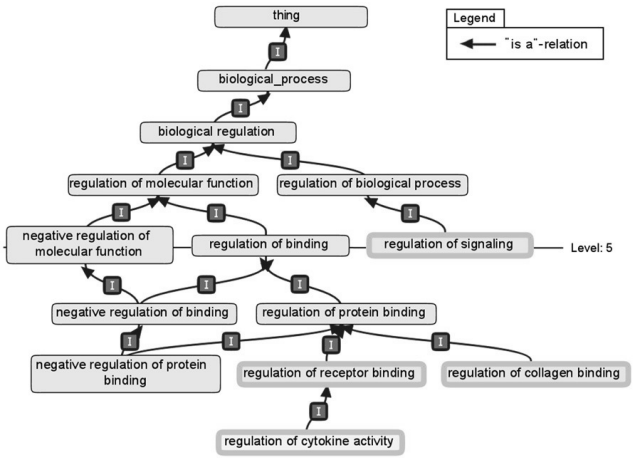


Fig. 1: A snapshot of sub graph of gene ontology

Thus the semantic similarity between two *GO-terms* is dependent on the level of their lowest common ancestors (LCA) in GO tree. We have named this level information of a *GO-term* as $struct_{depth}$ or structural depth of the term.

For example in Figure 1 the semantic similarity between *regulation of molecular function* and *regulation of biological process* is less than the semantic similarity between *regulation of receptor binding* and *regulation of collagen binding*. It is because the LCA of *regulation of molecular function* and *regulation of biological process*, which is *biological regulation*, is situated in higher level. Where as the LCA of *regulation of receptor binding* and *regulation of collagen binding*, which is *regulation of protein binding*, is located in lower level of GO tree. So intuitively the functional similarity between terms depends on positions of their LCAs in GO tree. Inspired by this observation we have proposed a $struct_{depth}$ based semantic similarity measure between genes. If there are two *GO-terms* t_i and t_j, the normalised-$struct_{depth}$ based semantic similarity is defined as,

$$sim_{norm-struct_{depth}}(t_i, t_j) = \frac{struct_{depth}(LCA(t_i, t_j))}{struct_{depth}(GO) - 1} \tag{9}$$

Where,
$struct_{depth}(LCA(t_i, t_j))$ is the lowest level of LCA of terms t_i and t_j in GO tree.
To make the value normalized we divide the above mentioned term by $struct_{depth}(GO)$ which is the lowest level or height of the GO tree.
We assume that the level of root term in GO tree is 1. The upper bound of $sim_{norm-struct_{depth}}$ is 1. Highest similarity value will be obtained between two *GO-terms* residing in the leaf level of GO tree, having their lowest common ancestors (LCA) residing in depth $(struct_{depth}(GO) - 1)$. The lowest value of $sim_{norm-struct_{depth}} > 0$. Lowest similarity value will be obtained between two *GO-terms*, when their LCA is root of the GO tree itself.

If we want to measure it in terms of distance then the above equation can be converted to the corresponding distance

function as follows,

$$dist_{norm-struct_{depth}}(t_i, t_j) = 1 - sim_{norm-struct_{depth}}(t_i, t_j) \quad (10)$$

Given two gene products g_i and g_j, which are annotated by a set of terms A_i and A_j where A_i and A_j contain m and n number of *GO-terms*, respectively, the normalized $struct_{depth}$ based semantic similarity between genes g_i and g_j, $SIM_{norm-struct_{depth}}(g_i, g_j)$ can be defined as the average inter-set similarity between two terms, A_i and A_j:

$$SIM_{norm-struct_{depth}}(g_i, g_j) = \left(\frac{1}{m \times n}\right) \times \sum_{t_k \in A_i, t_p \in A_j} sim_{norm-struct_{depth}}(t_k, t_p) \quad (11)$$

Where, $sim_{norm-struct_{depth}}(t_k, t_p)$ can be calculated using Equation 9.

B. Multi-factored semantic similarity

We have named this proposed semantic similarity measure as multi-factored as it covers different structural and information-theoretic properties of *GO-terms*. These factors are,

1. Shared IC based similarity between *GO-terms* according to Lin's model [13].
2. Shortest path based similarity between *GO-terms* based on Shen's model [28].
3. Normalized term overlap based similarity between genes based on Mistry's model [18].
4. Our proposed normalized $struct_{depth}$ based similarity between *GO-terms* according to Equation 9 in section IV-A.

These factors are mutually exclusive to each other and capture different structural and IC based aspects to measure similarity between *GO-terms*. Utilizing these factors the proposed multi-factored semantic similarity measure between term t_i and term t_j can be defined as,

$$\textit{Multi-sim}(t_i, t_j) = \frac{arctan[Y]}{\pi/2} \quad (12)$$

Where, $Y = [sim_{Lin}(t_i, t_j) + sim_{Shen}(t_i, t_j) + sim_{norm-struct_{depth}}(t_i, t_j)]$.
$sim_{Lin}(t_i, t_j)$ is calculated according to Equation 2.
$sim_{Shen}(t_i, t_j)$ is calculated according to Equation 8.
$sim_{norm-struct_{depth}}(t_i, t_j)$ is calculated according to Equation 9.
The arctan function is used to make the value of *Multi-sim*(t_i, t_j) normalized. Using the above equation the semantic similarity between gene products g_i and g_j can be defined as,

$$\textit{Multi-SIM}(g_i, g_j) = \frac{\frac{1}{m \times n} \sum_{t_k \in A_i, t_p \in A_j} \textit{Multi-sim}(t_k, t_p) + sim_{NTO}(g_i, g_j)}{2} \quad (13)$$

Where, $sim_{NTO}(g_i, g_j)$ is calculated according to Equation 6. We have taken the average so that *Multi-SIM*$(g_i, g_j) \in [0, 1]$.

C. Multi-factored semantic distance

Similar to multi-factored semantic similarity measurement we have also proposed a multi-factored distance measure. This is also a function of multiple factors like,

1. Shared IC based distance between *GO-terms* according to Jiang's model [10].
2. Shortest path based distance between *GO-terms* based on Shen's model [28].
3. Normalized term overlap based distance based on Mistry's model [18].
4. Normalized depth based distance between *GO-terms* according to Equation 10 in section IV-A.

Mathematically it can be defined as follows,

$$\textit{Multi-dist}(t_i, t_j) = \frac{arctan[X]}{\pi/2} \quad (14)$$

Where,
$X = [dist_{Jiang}(t_i, t_j) + dist_{Shen}(t_i, t_j) + dist_{norm-struct_{depth}}(t_i, t_j)]$.
$dist_{Jiang}(t_i, t_j)$ is calculated according to Equation 4.
$dist_{Shen}(t_i, t_j)$ is calculated according to Equation 7.
$dist_{norm-struct_{depth}}(t_i, t_j)$ is calculated according to Equation 10.
Using the above equation, the multi-factored semantic distance between gene products g_i and g_j is defined as,

$$\textit{Multi-DIST}(g_i, g_j) = \frac{\frac{1}{m \times n} \sum_{t_k \in A_i, t_p \in A_j} \textit{Multi-dist}(t_k, t_p) + dist_{NTO}(g_i, g_j)}{2} \quad (15)$$

Where, $dist_{NTO}(g_i, g_j)$ is the normalized term-overlap based distance which can be calculated from Equation 6 as follows,

$$dist_{NTO}(g_i, g_j) = 1 - sim_{NTO}(g_i, g_j) \quad (16)$$

The value of *Multi-DIST*$(g_i, g_j) \in [0, 1]$.

V. Proposed framework to identify functionally similar groups of genes based on semantic similarity and distance measures

We have divided our proposed framework into two modules.

1) Module 1: In this module three tasks are performed as follows,
   - We have chosen three datasets for experiment. The annotation information of all genes for all of three datasets are collected from GO database (http://www.geneontology.org/).
   - Each dataset is represented as *gene-GO-term* annotation matrix with the help of annotation information collected from previous step.
   - Utilizing the proposed three semantic similarity and distance measures, for all of the three datasets, corresponding gene-gene similarity matrices are extracted. We have also chosen seven existing similarity and distance measures and tools for comparison

purpose. For each of them, corresponding gene-gene similarity matrices for all of three datasets are generated too.

2) Module 2: Using the gene-gene similarity matrices produced at module 1, to identify functionally similar genes, clustering is performed on *gene-GO-term* annotation matrices or datasets using some multi-objective clustering algorithms (AMOSA [3] based clustering and NSGA-II [7] based clustering) along with six traditional single-objective clustering algorithms like K-means, K-medoids, Single-linkage hierarchical, Complete-linkage hierarchical, Average-linkage hierarchical and DBSCAN clustering algorithm.

Finally, a comparative study has been performed among all of the above chosen clustering algorithms utilizing existing and proposed similarity and distance measures. The proposed framework is shown in Figure 2 of supplementary file. Below we have described the different tasks of module 1 in detail.

A. Module 1: Preparing gene-GO-term annotation matrices and gene-gene similarity matrices based on existing and proposed semantic similarity/distance measures

The different tasks performed in this module are described as follows:

1) Task 1: Collecting annotation information from GO

We have chosen three datasets for experimental purpose which are prepared as follows,

- **Mouse genome dataset:** We have chosen set of genes randomly from *Mouse genome* as selected by Ref. [18]. Genes which are mapped to one or more *GO-terms* in Gene ontology Consortium (http://www.geneontology.org/) are chosen for further analysis. 14,599 number of such genes are mapped to one or more terms under one or more of three gene ontologies (BP, MF and CC). After genes get annotated by one or more *GO-terms*, within Gene ontology Consortium, there is provision to extract significant *GO-terms* having p-value ≤ 0.05 from the set of all mapped *GO-terms*. p-value is the probability or chance of seeing at least 'x' number of genes out of the total 'n' genes in the list annotated by a particular *GO-term*, given the proportion of genes in the whole genome that are annotated to that *GO-term*. That is, the *GO-terms* shared by the genes in the user's list are compared to the background distribution of annotation. The closer the p-value is to zero, the more significant the particular *GO-term* associated with the group of genes is (i.e., less likely the observed annotation of the particular *GO-term* to a group of genes occurs by chance). According to our chosen GO tool (http://www.geneontology.org/), we have selected significant *GO-terms* having p-value ≤ 0.05 in this experiment. For *Mouse genome* dataset, the total number of significant *GO-terms* obtained is 792 (number of *GO-terms* under biological process is 539, under Molecular function is 122, and under cellular component is 131).
- **Yeast dataset:** We have also chosen the *Yeast* [2] dataset [29] for experiments. Similar to *Mouse genome* dataset, genes of *Yeast* dataset also get mapped to one or more *GO-terms* in Gene ontology Consortium (http://www.geneontology.org/). Originally in *Yeast* dataset 2260 number of genes are mapped to one or more *GO-terms* under one or more ontologies (BP, MF and CC). For *Yeast* dataset, the number of obtained significant *GO-terms* is 166 (number of *GO-terms* under biological process is 100, under Molecular function is 43, and under cellular component is 23).
- **Human genome dataset:** Apart from *Mouse genome* and *Yeast* dataset, we have also chosen *Human genome* [3] dataset for our experiments. Same strategy is followed here like other two datasets to obtain gene annotation information. From the original set of genes, 18,244 number of genes get mapped to one or more *GO-terms* of Gene ontology Consortium under one or more ontologies (BP, MF and CC). For this dataset, number of obtained significant *GO-term* is 273 (number of *GO-terms* under biological process is 172, under Molecular function is 35, and under cellular component is 66).

Also we retrieve the full GO tree [4] containing all *GO-terms* and their corresponding relationships from Gene ontology Consortium. Among all types of GO relationships, we have considered 'is-a' and 'part-of' relationships as they are most dominating.

2) Task 2: Generating Binary gene-GO-term annotation matrix corresponding to each datasets

After retrieving the annotation information of genes from GO (previous task), it is utilized to generate binary *gene-GO-term* annotation matrices for all three datasets. These *gene-GO-term* annotation matrices are developed solely based on available GO information. No gene expression data is utilized for this experiment as that does not capture semantic information about genes. The binary *gene-GO-term* annotation matrices are generated in the following way:
Suppose for any of our chosen datasets, if there are n number of annotated genes, the number of total significant *GO-terms* for biological process, cellular component, molecular function ontology are x, y and z, respectively. Then the size of the *gene-GO-term* matrix will be $n \times (x + y + z)$. Here for *Mouse genome* dataset, *Yeast* and *Human genome* dataset, number of significant total *GO-terms* are 792, 166 and 273, respectively. So the dimensions of *gene-GO-term* annotation matrix for *Mouse genome*, *Yeast* and *Human genome* dataset are $14,599 \times 792$, 2260×166 and $18,244 \times 273$ respectively. Entries in the matrix are binary values based on whether the gene is mapped to that particular *GO-term* or not. Mathematically it can be described as follows,
If $\exists$ n genes and x, y, z number of significant Biological function *GO-terms*, Molecular function *GO-terms* and Cellular component *GO-terms*, respectively, then $|M| = n \times (x + y + z)$.

[2] http://arep.med.harvard.edu.
[3] http://120.77.47.2:8081/download
[4] http://purl.obolibrary.org/obo/go/go-basic.obo

Suppose G_i represents i^{th} gene where $i \in [1, n]$.
Bio_GO_k represents k^{th} significant term of Biological process ontology, where $k \in [1, x]$.
MF_GO_l represents l^{th} significant term of Molecular function ontology, where $l \in [1, y]$.
CC_GO_m represents m^{th} significant term of Cellular component ontology, where $m \in [1, z]$.
Let M[n][x+y+z] be the binary annotation matrix of size $n \times (x + y + z)$.
The matrix is generated as follows,

$$M[i][Bio\_GO_k] = \begin{cases} 1, & \text{if } G_i \text{ annotated with } Bio\_GO_k \\ 0, & \text{otherwise} \end{cases}$$

where $i \in [1, n]$ and $k \in [1, x]$.

$$M[i][MF\_GO_l] = \begin{cases} 1, & \text{if } G_i \text{ annotated with } MF\_GO_l \\ 0, & \text{otherwise} \end{cases}$$

where $i \in [1, n]$ and $l \in [1, y]$.

$$M[i][CC\_GO_m] = \begin{cases} 1, & \text{if } G_i \text{ annotated with } CC\_GO_m \\ 0, & \text{otherwise} \end{cases}$$

where $i \in [1, n]$ and $m \in [1, z]$. In Figure 3 of supplementary file, a sample binary *gene-GO-term* annotation matrix has been shown.

3) Task 3: Generating gene-gene similarity matrices for all three datasets

To perform clustering on genes of produced *gene-GO-term* annotation datasets by task 2, similarity/distance is required to be calculated between genes. For that, we have generated gene-gene similarity/distance matrices based on which genes are allocated to different groups. In addition to our proposed similarity and distance measures, we have also chosen seven existing state-of-the-art similarity and distance measures and tools to measure gene-gene similarity and prepare similarity matrix corresponding to each of them. The existing chosen measures and tools are, Jiang's distance measure [10], Lin's similarity measure [13], Shortest path based similarity measure by Shen [28], Normalized term overlap based similarity measure by Mistry [18], *InteGO* [21], *G-SESAME* [30], *IT-GOM* under *A-DaGO-Fun* package [15]. Within *A-DaGO-Fun* package, as *IT-GOM* tool is independent of organisms and gene-IDs, therefore, we have chosen this tool to measure term-term similarities and hence from those intuitively gene-gene similarities are calculated. Using *IT-GOM*, we have calculated term-term similarities using annotation based XGraSM-Nunivers similarity measure approach. For *InteGO* [21], *G-SESAME* [30], *IT-GOM* [15] tools, as gene-gene or term-term similarity is calculated separately for each ontology (BP, MF, CC), therefore, we have taken average similarity over three ontologies to create final gene-gene similarity matrix.
In a dataset if there are n number of genes, then the dimension of gene-gene similarity matrix corresponding to each of the proposed and chosen similarity and distance measures is $n \times n$.
Suppose S[n][n] is the similarity matrix of size $n \times n$. For our three proposed similarity/distance measures, the matrix is generated as follows,

$$S[i][j] = \begin{cases} SIM_{norm-struct_{depth}}(g_i, g_j), \\ \text{if Normalized depth based semantic similarity} \\ \text{is used.} \\ \textit{Multi-SIM}(g_i, g_j), \\ \text{if Multi-factored semantic similarity is used.} \\ (1 - \textit{Multi-DIST}(g_i, g_j)), \\ \text{if Multi-factored semantic distance is used.} \end{cases}$$

where $i \in [1, n]$ and $j \in [1, n]$.

For, other chosen proximity measures, the corresponding gene-gene similarity matrices are generated similarly.

B. Module 2: Application of clustering algorithms on gene-GO-term annotation matrices using proposed and existing similarity and distance measures

This is module 2 of our proposed framework where different clustering algorithms (both multi-objective and single-objective) have been applied on *gene-GO-term* annotation datasets utilizing their corresponding similarity/distance matrices obtained from Module 1 to uncover functionally related genes. Genes which are functionally similar with respect to the proximity measures are clustered together. Existing literature [1] proved the utility of multi-objective optimization over single-objective optimization in solving different real-life problems. Inspired by this, in recent years several multi-objective optimization based clustering techniques are also developed in the literature Ref. [1], [12]. These approaches work better than their single-objective counter parts. Motivated by this, in the current study we have executed two multi-objective based clustering techniques on *gene-GO-term* annotation matrices of all three chosen datasets utilizing ten different similarity and distance measures (as mentioned in Sections II and IV) to partition the set of available genes into some functionally similar groups. Two popular multi-objective optimization(MOO) strategies, AMOSA [3] and NSGA-II [7] based clustering, along with six traditional single-objective clustering techniques are utilized for this purpose. All of these clustering algorithms are well-known and widely used in the field of pattern recognition as well as bioinformatics. The detailed descriptions of chosen clustering techniques are provided in Section I of supplementary file. We have conducted a comparative study of the performance of these clustering algorithms in the current context.

VI. Results and discussion

After application of all the selected clustering algorithms on *gene-GO-term* annotation matrices for all of three datasets by varying different similarity and distance measures (both of our proposed and existing measures), we obtained different clustering solutions. In this section we have analysed our obtained solutions using two different performance metrics, namely Silhouette index [24] and DB index [6], to quantify the goodness of obtained partitions. A clustering solution having higher Silhouette index and lower DB index is considered to

have good quality. These validity measures are described in detail in Section II of supplementary file.

A. Setting of input parameters for clustering algorithms

1) Input parameters of AMOSA based clustering

We have executed AMOSA based clustering technique with the following parameter combinations:
T_{min} = 0.0001, T_{max} = 100, α = 0.9, HL = 50, SL = 100 and $iter$ = 100. K_{min} or minimum number of clusters = 2 and K_{max} or maximum number of clusters = $\sqrt{n}$, where n number of genes/points to be clustered.
The parameter values are determined after conducting a thorough sensitivity study. The number of clusters is determined automatically by the algorithm within given ranges.

2) Input parameters for NSGA-II based clustering

We have executed NSGA-II based clustering using following parameter combinations:
Number of generation: 50. Population size: 50. Crossover probability: 0.8. Mutation strength 0.2. K_{min} or minimum number of clusters = 2 and K_{max} or maximum number of clusters = $\sqrt{n}$, where n is the number of genes/points to be clustered. Here also the number of clusters is determined automatically within the given range.

3) Input parameters for K-means, K-medoids and hierarchical clustering

Unlike AMOSA or NSGA-II based clustering, for K-means/K-medoids clustering technique, prior information about the number of clusters (K) is needed. Also, we have applied agglomerative type single, complete and average linkage hierarchical clustering algorithms. Similar to K-means/K-medoids, here also the number of clusters 'K' is required to be initialized as terminating point during agglomeration. It is known that if no information about the number of clusters is given, then for n number of data points, the maximum number of clusters can be chosen as $\sqrt{n}$ [5]. According to that, for *Yeast*, *Mouse genome* and *Human genome* datasets, the maximum number of clusters can be $\sqrt{2260}$ or 48, $\sqrt{14,599}$ or 120 and $\sqrt{18,244}$ or 135 , respectively. To determine the optimal value of 'K', we have varied the value of K in the range 2 to 50 as shown in Table I. For *Yeast*, *Mouse genome* and *Human genome* datasets the results were deteriorated rapidly with the increase in the value of K above 20. Therefore, we stopped the experiment with value of K beyond 50. The optimal value of K for which best clustering solution is obtained is reported in Table II for each clustering algorithms.

Datasets	K													
Yeast	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Mouse genome	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Human genome	2	3	4	5	6	7	8	9	10	15	20	30	40	50

TABLE I: Chosen K values for K-means, K-medoids and Hierarchical clustering algorithms

4) Input parameters for DBSCAN

In DBSCAN, two parameters ξ or radius and minPts or minimum points in a cluster are needed to be set at the beginning of its execution. We have thoroughly done a sensitivity study on the value of minPts and ξ in such a way so that total number of outliers should be minimum.

B. Discussion on results

In Table II, we have reported our obtained results for *Mouse genome*, *Yeast* and *Human genome* datasets. For each chosen similarity measure, the best obtained Silhouette/DB index values across all chosen clustering algorithms are highlighted in bold font. Similarly, for each selected clustering algorithm, the best of Silhouette/DB index values across all mentioned similarity measures are highlighted with blue and pink box, respectively. Beside each Silhouette index value, in bracket, we have mentioned the optimal number of clusters (K) corresponding to that clustering solution. The comparative results are also graphically plotted in Figures 3, 4, 5.

Please note that we could not get results for *Mouse genome* dataset using *InteGO* tool as the corresponding *Mus musculus* organism is beyond the scope of available organisms in the tool. The missing values are represented using '–' in Table II. While using *IT-GOM* tool under *A-DaGO-Fun* package, we have calculated gene-gene similarities via calculating term-term similarities using annotation based XGraSM-Nunivers similarity measure.

We have carried out a thorough comparative analysis of the obtained results as shown in Table II and Figures 3, 4 and 5. We observed that for *Mouse genome* dataset, for all nine (except *InteGO*) chosen similarity measures (six of them are existing and other three are our proposed measures), multi-objective clustering algorithms, i.e., AMOSA and NSGA-II based clustering algorithms perform better than other chosen algorithms with respect to both Silhouette and DB index values in almost all cases. Again, while comparing AMOSA based clustering with NSGA-II based clustering, we have observed that for almost all cases AMOSA based clustering performs better than NSGA-II based clustering with respect to both Silhouette and DB indices.

Also, the performance of the proposed and existing similarity and distance measures can be compared with each other with the help of the obtained results. It is quite clear from Table II that all of the chosen clustering algorithms perform well when any of our proposed *Multi-SIM* and *Multi-DIST* measures is considered as proximity measure in almost all cases. It has been observed that, *G-SESAME* outperforms chosen existing measures like $dist_{Jiang}$, sim_{Shen}, sim_{NTO}, sim_{Lin}, *A-DaGO-Fun* as well as our proposed $SIM_{norm-struct_{depth}}$ measure, in most of the cases with respect to both Silhouette index and DB index values, but both of our proposed multi-factored proximity measures perform better than *G-SESAME* for all the cases. The same trend has also been observed if we compare outcomes of *IT-GOM* tool (under *A-DaGO-Fun* package) with our proposed multi-factored measures. Though $SIM_{norm-struct_{depth}}$ alone could not perform better than most of the existing and proposed measures but considering this factor in both *Multi-SIM* and *Multi-DIST* measures, makes them superior to other measures. It proves that $struct_{depth}$ is an important topological factor which should not be ignored while determining functional similarity between genes.

We have also compared the efficiencies of *Multi-SIM* and *Multi-DIST* measures with respect to each other and it was found that both of them are complementary to each other,

i.e., no one is better than the other. Investigations of overall obtained results for *Mouse genome* dataset reveal that best (with respect to Silhouette and DB index values) clustering solutions are obtained when AMOSA based clustering is applied in association with *Multi-SIM* or *Multi-DIST* as the underlying proximity measure.

Similar to *Mouse genome* dataset, we have also thoroughly analysed the results for *Yeast* dataset in Table II and Figure 4. Here also, the superiority of AMOSA based clustering over other multi and single-objective based clustering techniques and the supremacy of proposed *Multi-SIM* and *Multi-DIST* similarity measures over other chosen similarity/distance measures are established based on the obtained Silhouette and DB index values. Though the results of *InteGO* tool are better than other existing measures like $dist_{Jiang}$, sim_{Shen}, sim_{NTO}, sim_{Lin}, *G-SESAME*, *A-DaGO-Fun* and also $SIM_{norm-struct_{depth}}$, but it fails to defeat newly proposed $Multi-SIM$ and $Multi-DIST$ measures.

For *Human genome* dataset also we have performed similar analysis of the obtained results (reported in Table II and shown in Figure 5). Here also, the superiority of AMOSA-based clustering in association with all chosen similarity measures over other clustering algorithms is conserved. Comparison between capabilities of all ten (chosen and proposed) similarity measures reveals that our proposed $Multi-SIM$ and $Multi-DIST$ measures perform better than the others. Here also, *InteGO* tool performs better than other chosen existing measures as well as our proposed $SIM_{norm-struct_{depth}}$, but it is not able to defeat newly proposed $Multi-SIM$ and $Multi-DIST$ measures. For all three chosen datasets, the best clustering solution is obtained by AMOSA based clustering algorithm with *Multi-SIM* or *Multi-DIST* as the underlying proximity measure.

To validate the above observation through visualization, we have shown the cluster profile plots of obtained clusters for *Yeast* dataset. For this dataset, the number of clusters obtained by AMOSA based clustering with *Multi-SIM* as the underlying proximity measure is six (also indicated in Table II). We have plotted expression profiles of the genes for first two obtained clusters. The expression value of each gene is extracted from gene expression dataset of *Yeast*. The plots are shown in Figure 2. For each plot, the X-axis represents the samples or time points of each gene, while Y-axis represents the genes within the corresponding cluster. The log normalized expression values of genes within a cluster for some given time points are plotted. From this plot (shown in Figure 2) it can be inferred that genes within the same cluster are nicely coherent to each other. It signifies that they are strongly co-related. Similarly, other clusters can also be plotted through cluster profile plots. These plots also support our observations from obtained results in Table II.

The major observations concluded from the obtained results are summarized below:

1) Our proposed $Multi-SIM$ and $Multi-DIST$ similarity measures can be treated as best proximity measures in terms of their capability in identifying functionally similar genes compared to other chosen similarity/distance measures. The obtained Silhouette and DB index values of Table II justify this observation.
2) AMOSA [3] based clustering performs the best in identifying functionally similar groups of genes, compared to other mentioned single and multi-objective based clustering algorithms for all ten chosen similarity/distance measures.
3) AMOSA based clustering with $Multi-SIM$ or $Multi-DIST$ as the underlying proximity measure, are two best combinations in terms of determining groups of most functionally similar genes. The supremacies of the clustering solutions produced by these two combinations are established by the thorough experimental results (presented in Table II and Figure 2).

C. Statistical significance test

In Table II, we have shown the results on all three datasets. From the results the supremacy of AMOSA based clustering and $Multi-SIM$ and $Multi-DIST$ measures were observed. Now in order to prove these observations statistically, a statistical significance test (with respect to DB index) is conducted (also known as t-test) at 5% significance level for *Yeast* dataset. Instead of DB index, Silhouette index also could be chosen for this test. As null hypothesis we assume that there are insignificant differences between mean values of two groups. According to alternative hypothesis there are significant differences in the mean values of two groups.

To prove the supremacy of AMOSA based clustering, eight groups corresponding to eight chosen clustering algorithms are formed for each chosen similarity/distance measure. The p-values produced between each two groups(one corresponding to AMOSA and other corresponding to any one of seven other clustering algorithms) by t-tests for *Yeast* dataset are reported in Table I of supplementary file. Similarly, to prove superiority of *Multi-SIM*, nine groups corresponding to nine similarity/distance measures(except *Multi-DIST* as we have seen that *Multi-DIST* has equal potential to *Multi-SIM*) are formed. In Table II of supplementary file, statistical superiority of *Multi-SIM* over other existing similarity/distance measures (except *Multi-DIST*) is shown. Similar to Table II, supremacy of *Multi-DIST* can also be proved statistically.

From both of the tables, it can easily be seen that the p-value in each case is less than 0.05. This outcome supports the alternative hypothesis i.e., the supremacy of AMOSA based clustering algorithm over other algorithms and *Multi-SIM* over other measures (except *Multi-DIST*). Finally, the superiority of AMOSA based clustering along with *Multi-SIM* as underlying proximity measure is proved by this statistical test. For *Mouse genome* and *Human genome* dataset similar statistical significance tests can be performed.

D. Biological significance test

From Table II and Figure 3, 4 and 5, it is evident that the performance of AMOSA based clustering with *Multi-SIM* or *Multi-DIST* proximity measures was best among other possible combinations. To support the obtained results statistically we have also performed statistical significance test which is shown

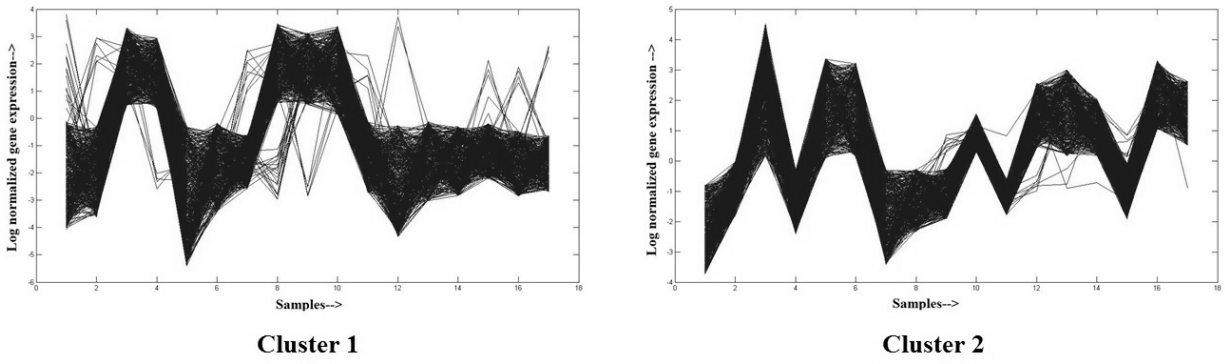


Fig. 2: Cluster profile plots for first two obtained clusters by AMOSA-*Multi-SIM* algorithm for *Yeast* dataset

Dataset	Algorithms	$dist_{Jiang}$		sim_{Shen}		sim_{NTO}		sim_{Lin}		*InteGO*		*G-SESAME*		*A-DaGOFun*		$SIM_{norm-struct_{depth}}$		*Multi-SIM*		*Multi-DIST*	
		Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB	Sil	DB
Mouse	AMOSA - clus	**0.61**(8)	**1.234**	**0.59**(9)	**1.19**	0.55(8)	**1.342**	**0.59**(9)	**1.245**	–	–	**0.73**(10)	**0.99**	**0.69**(10)	1.21	0.568(8)	**1.278**	**0.7802**(11)	**0.9509**	**0.79**(11)	0.99
	NSGA-II - clus	0.56(7)	1.49	0.56(7)	1.19	**0.57**(7)	1.38	0.574(7)	1.27	–	–	**0.73**(9)	1.08	0.68(10)	**1.19**	0.567(7)	1.31	0.75(9)	0.97	0.752(10)	**0.98**
	K-means	0.48(5)	2.13	0.49(5)	2.3	0.48(5)	2.3	0.465(5)	2.3	–	–	0.52(8)	1.98	0.59(8)	1.63	0.48(6)	2.005	0.58(6)	1.69	0.51(8)	1.61
	K-medoids	0.49(5)	2.132	0.499(5)	2.31	0.48(5)	2.32	0.47(5)	2.3	–	–	0.53(8)	1.9	0.59(8)	1.62	0.49(6)	2.002	0.6(8)	1.68	0.54(8)	1.6
	Hie-single	0.52(6)	1.53	0.535(6)	1.2	0.516(6)	1.48	0.55(8)	1.42	–	–	0.61(8)	1.36	0.6(8)	1.37	0.53(7)	1.29	0.65(10)	1.23	0.66(10)	1.21
	Hie-average	0.54(6)	1.54	0.543(6)	**1.19**	0.537(6)	1.42	0.56(8)	1.4	–	–	0.612(8)	1.21	0.61(8)	1.36	0.546(8)	1.3	0.657(10)	0.98	0.656(10)	1
	Hie-complete	0.56(8)	1.56	0.546(6)	1.198	0.56(8)	1.35	0.57(8)	1.43	–	–	0.63(10)	1.23	0.62(8)	1.35	**0.58**(9)	1.5	0.66(10)	0.982	0.659(10)	1.003
	DBSCAN	0.53(6)	1.61	0.53(6)	1.52	0.5(8)	1.4	0.55(8)	1.44	–	–	0.63(9)	1.39	0.61(8)	1.42	**0.58**(8)	1.54	0.634(9)	1.31	0.64(9)	1.33
Yeast	AMOSA - clus	**0.599**(5)	**1.91**	**0.529**(5)	**1.86**	**0.476**(4)	2.32	**0.62**(5)	**1.97**	**0.78**(6)	**1.14**	**0.74**(6)	1.21	**0.73**(6)	**1.29**	**0.49**(4)	**1.65**	**0.81**(6)	**1.031**	**0.738**(6)	**0.99**
	NSGA-II - clus	0.594(5)	**1.91**	0.521(5)	1.88	0.47(4)	**2.28**	0.58(5)	1.99	0.72(6)	1.24	0.69(6)	1.28	**0.73**(6)	1.28	0.489(4)	1.68	0.78(6)	1.13	0.715(5)	1.109
	K-means	0.579(8)	1.99	0.479(9)	2	0.393(10)	2.49	0.48(8)	2.03	0.6(5)	1.73	0.57(8)	1.9	0.56(8)	1.71	0.461(9)	1.79	0.59(5)	1.53	0.608(5)	1.62
	K-medoids	0.581(8)	1.95	0.488(9)	1.99	0.398(10)	2.48	0.5(8)	2.01	0.62(5)	1.71	0.59(8)	1.8	0.57(8)	1.7	0.467(9)	1.78	0.61(5)	1.50	0.615(5)	1.618
	Hie-single	0.589(8)	1.93	0.489(9)	1.91	0.418(9)	2.39	0.55(8)	1.99	0.62(5)	1.17	0.58(8)	**1.2**	0.6(8)	1.61	0.486(9)	1.69	0.628(5)	1.158	0.624(5)	1.032
	Hie-average	0.593(8)	1.92	0.510(9)	1.89	0.428(9)	2.31	0.56(8)	1.98	0.62(5)	1.31	0.59(8)	1.38	0.62(8)	1.59	0.489(9)	1.67	0.620(5)	1.26	0.629(5)	1.12
	Hie-complete	0.591(8)	1.93	0.499(9)	1.9	0.429(9)	2.29	0.57(8)	1.98	0.64(5)	1.4	0.62(8)	1.52	0.61(8)	1.6	0.479(9)	1.71	0.658(5)	1.3	0.649(5)	1.297
	DBSCAN	0.578(8)	1.94	0.5(9)	2.03	0.417(9)	2.34	0.59(8)	2	0.63(5)	1.18	0.6(5)	1.21	0.59(8)	1.62	0.469(9)	1.67	0.63(7)	1.31	0.621(5)	1.32
Human	AMOSA - clus	**0.67**(14)	**1.87**	**0.65**(14)	**1.88**	0.58(16)	**2.13**	**0.66**(14)	1.93	**0.76**(13)	**1.23**	**0.71**(13)	1.64	**0.73**(13)	**1.24**	**0.59**(16)	**2.03**	**0.87**(11)	**1.15**	**0.89**(11)	**1.09**
	NSGA-II - clus	0.66(14)	1.88	**0.65**(14)	1.89	**0.59**(15)	2.15	0.65(14)	**1.91**	0.75(14)	1.25	0.69(14)	**1.63**	0.72(13)	1.25	0.58(16)	2.05	**0.87**(11)	1.17	0.87(11)	1.1
	K-means	0.56(15)	2.23	0.58(15)	2.27	0.49(20)	2.32	0.57(15)	2.27	0.68(15)	1.34	0.61(15)	1.78	0.64(15)	1.58	0.49(20)	2.19	0.77(10)	1.23	0.76(10)	1.2
	K-medoids	0.57(15)	2.21	0.59(15)	2.26	0.51(20)	2.32	0.59(15)	2.25	0.68(15)	1.34	0.61(15)	1.76	0.65(15)	1.56	0.5(20)	2.15	0.79(10)	1.23	0.76(10)	1.21
	Hie-single	0.6(15)	2.14	0.61(15)	2.21	0.52(20)	2.3	0.62(15)	2.23	0.7(10)	1.32	0.62(15)	1.72	0.67(15)	1.45	0.53(20)	2.12	0.83(10)	1.2	0.84(10)	1.24
	Hie-average	0.61(15)	2.12	0.61(15)	2.19	0.52(20)	2.3	0.61(15)	2.22	0.71(10)	1.31	0.63(15)	1.7	0.69(15)	1.43	0.52(20)	2.07	0.83(10)	1.19	0.84(10)	1.22
	Hie-complete	0.61(15)	2.13	0.62(15)	2.18	0.53(20)	2.29	0.62(15)	2.23	0.71(10)	1.3	0.62(15)	1.69	0.68(15)	1.43	0.52(20)	2.05	0.84(10)	1.19	0.83(10)	1.1
	DBSCAN	0.6(15)	2.11	0.61(15)	2.18	0.51(20)	2.27	0.63(15)	2.21	0.7(10)	1.3	0.6(15)	1.68	0.68(15)	1.41	0.53(20)	2.07	0.82(10)	1.18	0.83(10)	1.12

TABLE II: The comparative results of different clustering algorithms applied with our proposed and existing similarity and distance measures on *Mouse genome*, *Yeast* and *Human genome* datasets. '–' represents missing values. The number of clusters in optimal solution is mentioned within bracket beside the Silhouette index for each clustering technique.

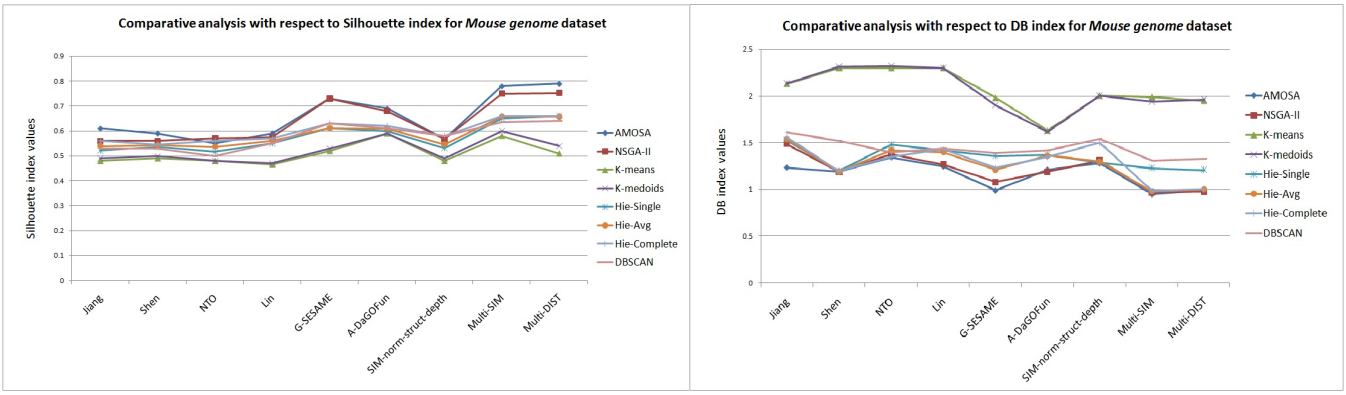


Fig. 3: Comparative results for *Mouse Genome* dataset

in Tables I and II of supplementary file. In this section, we have performed biological significance test to prove that obtained clusters by AMOSA based clustering with *Multi-SIM* or *Multi-DIST* proximity measures for *Yeast* dataset are biologically enriched. We have performed the study with the help of GOTERMMAPPER (http://go.princeton.edu/cgi-bin/GOTermMapper) under Biological process category. The best solution obtained for *Yeast* dataset after performing AMOSA based clustering with *Multi-SIM* proximity measure contains six clusters. For first cluster, the outcome of biological significance test is shown in Table III. Similarly for other five clusters we have performed the same test and the corresponding result tables are give in supplementary file (Table III, IV, V, VI, VII of supplementary file respectively). In each table we have summarized significant *GO-terms* shared by genes of corresponding cluster.

For each *GO-term*, the percentage of genes sharing that term among the genes of that cluster and among the whole

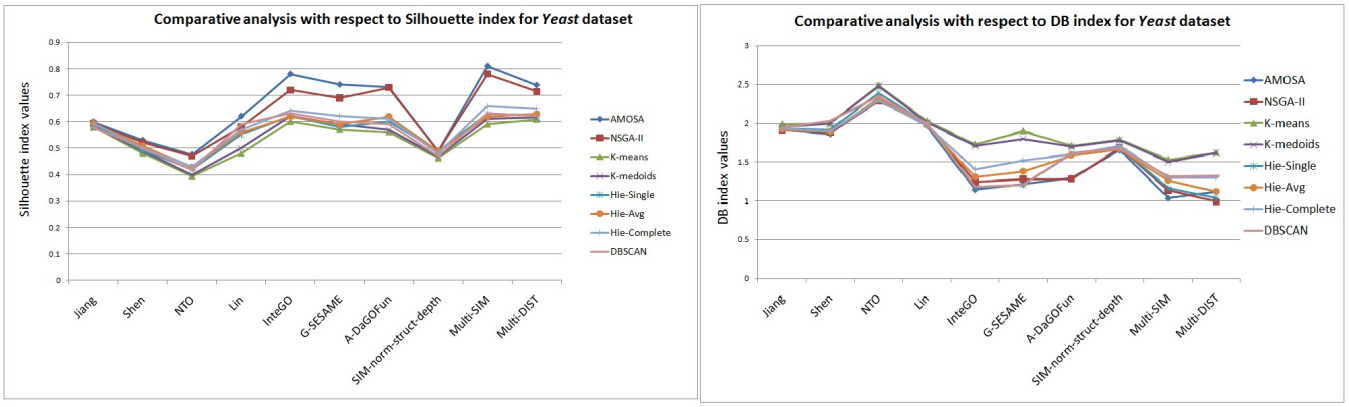


Fig. 4: Comparative results for *Yeast* dataset

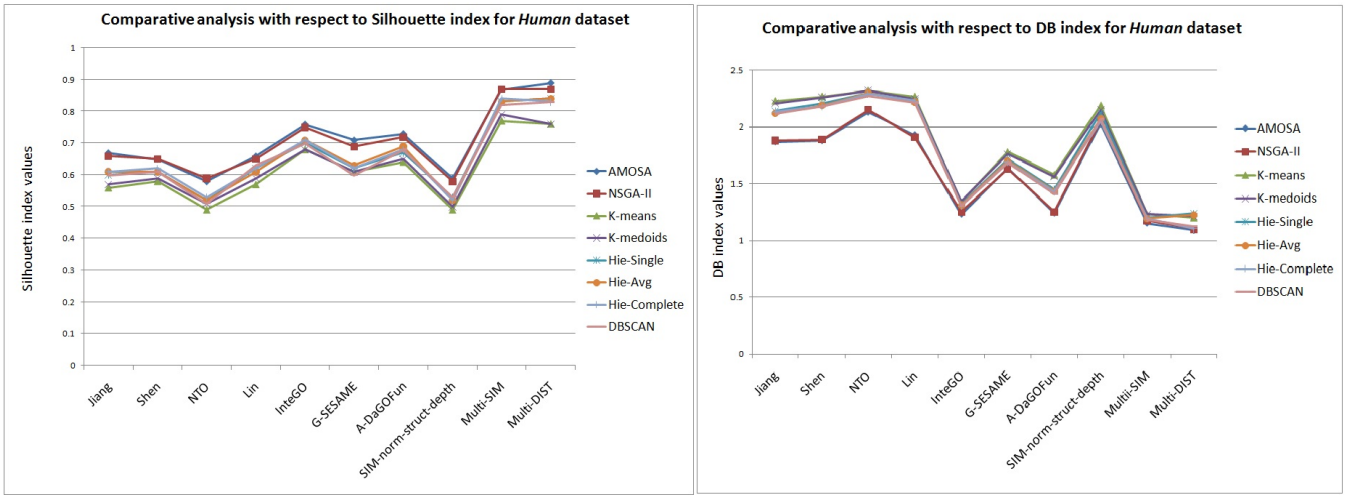


Fig. 5: Comparative results for *Human genome* dataset

genome has been reported. Results clearly signify that genes of the same cluster share the higher percentage of *GO-terms* compared to the whole genome. This indicates that the genes of a particular cluster are more involved in similar biological processes compared to the remaining genes of the genome.

The clustering solutions obtained by performing AMOSA based clustering using *Multi-DIST* proximity measure can also be tested following the above mentioned steps of the biological significance test. For *Mouse genome* and *Human genome* dataset same tests also can be performed.

GO-term	Module %	Genome %
chromatin organization GO:0006325	49.59%	35.35%
response to chemical GO:0042221	58.86%	38.39%
lipid metabolic process GO:0006629	57.38%	25.32%
cell wall organization or biogenesis GO:0071554	46.64%	24.65%
cellular response to DNA damage stimulus GO:0006974	46.64%	25.30%
cytoplasmic translation GO:0002181	55.54%	33.13%

TABLE III: Significant *GO-terms* shared by genes of cluster 1 for *Yeast* dataset

VII. Conclusions and Future Work

In general, expression values available in the microarray datasets are utilized for determining the similarity between two given genes using traditional distance measures like Euclidean, Pearson or symmetry based distance, which in turn aids in identifying functionally similar genes. This process of identification of functionally similar genes does not consider semantic relationships present among these genes. Also, the existing semantic similarity measurement techniques do not consider some important *GO-term* properties in a single measure. In order to overcome these limitations, in the current study, we have adopted an alternative way to represent genes and then defined some new hybrid measures of similarity between them utilizing different mutually exclusive properties of *GO-terms* in gene ontology database. The concept of topological level based gene-gene similarity measure namely $SIM_{norm-struct_{depth}}$ is quite unique and has not explored in any existing technique. Using this measure along with other *GO-term* properties we have also proposed two new proximity measures, named as *Multi-SIM* and *Multi-DIST*. Finally, the proposed semantic similarity and distance mea-

sures have been utilized along with some well-known single and multi-objective based clustering algorithms to determine groups of similar genes. As a case study, three datasets *Mouse genome*, *Yeast* and *Human genome* have been chosen. From the obtained results, we have observed that the best (with respect to Silhouette and DB index) clustering solutions are obtained when AMOSA based clustering is applied with *Multi-SIM* or *Multi-DIST* as the underlying proximity measure.

In future, we would like to develop a multi-view clustering framework considering gene expression data matrix and *gene-GO-term* annotation matrix as two different views. As the underlying proximity measure, our proposed measures can be used. Authors are currently working in this direction.

VIII. Acknowledgement

The authors would like to thank TCS Research Scholar Program for supporting us financially to conduct this research.

References

[1] S. Acharya and S. Saha, "Importance of proximity measures in clustering of cancer and mirna datasets: proposal of an automated framework," *Molecular BioSystems*, vol. 12, no. 11, pp. 3478–3501, 2016.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[3] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," *IEEE transactions on evolutionary computation*, vol. 12, no. 3, pp. 269–283, 2008.

[4] D. P. Bartel, "Micrornas: genomics, biogenesis, mechanism, and function," *cell*, vol. 116, no. 2, pp. 281–297, 2004.

[5] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 301–315, 1998.

[6] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[8] Z. Du, L. Li, C.-F. Chen, P. S. Yu, and J. Z. Wang, "G-sesame: web tools for go-term-based gene similarity analysis and knowledge discovery," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W345–W349, 2009.

[9] H. Fröhlich, N. Speer, A. Poustka, and T. Beißbarth, "Gosim–an r-package for computation of information theoretic go similarities between terms and gene products," *BMC bioinformatics*, vol. 8, no. 1, p. 166, 2007.

[10] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[11] B. Li, J. Z. Wang, F. A. Feltus, J. Zhou, and F. Luo, "Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins," *arXiv preprint arXiv:1001.0958*, 2010.

[12] L. Li, L. Jiao, J. Zhao, R. Shang, and M. Gong, "Quantum-behaved discrete multi-objective particle swarm optimization for complex network clustering," *Pattern Recognition*, vol. 63, pp. 1–14, 2017.

[13] D. Lin *et al.*, "An information-theoretic definition of similarity." in *ICML*, vol. 98, no. 1998. Citeseer, 1998, pp. 296–304.

[14] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, "Gotoolbox: functional analysis of gene datasets based on gene ontology," *Genome biology*, vol. 5, no. 12, p. R101, 2004.

[15] G. K. Mazandu, E. R. Chimusa, M. Mbiyavanga, and N. J. Mulder, "A-dago-fun: an adaptable gene ontology semantic similarity-based functional analysis tool," *Bioinformatics*, vol. 32, no. 3, pp. 477–479, 2015.

[16] G. K. Mazandu and N. J. Mulder, "Dago-fun: tool for gene ontology-based functional analysis using term information content measures," *BMC bioinformatics*, vol. 14, no. 1, p. 284, 2013.

[17] ——, "Information content-based gene ontology semantic similarity approaches: toward a unified framework theory," *BioMed research international*, vol. 2013, 2013.

[18] M. Mistry and P. Pavlidis, "Gene ontology term overlap as a measure of gene functional similarity," *BMC bioinformatics*, vol. 9, no. 1, p. 327, 2008.

[19] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "On biclustering of gene expression data," *Current Bioinformatics*, vol. 5, no. 3, pp. 204–216, 2010.

[20] J. Peng, S. Uygun, T. Kim, Y. Wang, S. Y. Rhee, and J. Chen, "Measuring semantic similarities by combining gene ontology annotations and gene co-function networks," *BMC bioinformatics*, vol. 16, no. 1, p. 44, 2015.

[21] J. Peng, Y. Wang, and J. Chen, "Towards integrative gene functional similarity measurement," in *BMC bioinformatics*, vol. 15, no. 2. BioMed Central, 2014, p. S5.

[22] A. Pesaranghader, S. Matwin, M. Sokolova, and R. G. Beiko, "simdef: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes," *Bioinformatics*, vol. 32, no. 9, pp. 1380–1387, 2015.

[23] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[25] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[26] ——, "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[27] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio, "Correlation between gene expression and go semantic similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 2, no. 4, pp. 330–338, 2005.

[28] Y. Shen, S. Zhang, and H.-S. Wong, "A new method for measuring the semantic similarity on gene ontology," in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*. IEEE, 2010, pp. 533–538.

[29] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[30] 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.

[31] C. Wolting, C. J. McGlade, and D. Tritchler, "Cluster analysis of protein array results via similarity of gene ontology annotation," *BMC bioinformatics*, vol. 7, no. 1, p. 338, 2006.

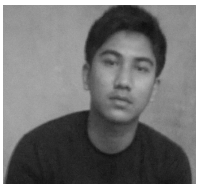
[32] K. Y. Yeung and R. E. Bumgarner, "Multiclass classification of microarray data with repeated measurements: application to cancer," *Genome biology*, vol. 4, no. 12, p. R83, 2003.

[33] H. Yu, R. Jansen, G. Stolovitzky, and M. Gerstein, "Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications," *Bioinformatics*, vol. 23, no. 16, pp. 2163–2173, 2007.

[34] P. Zhang, J. Zhang, H. Sheng, J. J. Russo, B. Osborne, and K. Buetow, "Gene functional similarity search tool (gfsst)," *BMC bioinformatics*, vol. 7, no. 1, p. 135, 2006.

Sudipta Acharya Sudipta Acharya received the B.Tech. degree in Information Technology from West Bengal University of Technology, India, in 2011 and the M.Tech. degree in Information technology from National Institute of Technology Durgapur, India in 2013. She is currently working towards the Ph.D. degree at the Indian Institute of Technology Patna, India. She is a coauthor of 20 papers. Her research interests include multi-objective optimization, clustering, pattern recognition, data mining and agent and multi-agent system. She is the recipient of TCS fellowship and IIT Patna high value fellowship award for her research achievement during her Ph.D.

Sriparna Saha Sriparna Saha is an Assistant professor of Department of Computer Science and Engineering of Indian Institute of Technology Patna, India. She obtained her PhD degree in computer science from Indian Statistical Institute Kolkata, India in the year 2011. Her research interests include pattern recognition, multiobjective optimization and biomedical information extraction. She is the author of a book published by Springer-Verlag. She is the recipient of the Google India Women in Engineering Award, 2008, NASI Young Scientist Platinum Jubilee Award 2016, BIRD Award 2016 and IEI Young Engineers' Award 2016. Her h-index is 17 and total citation count of her papers is 1768 (according to Google scholar).



Prasanna Pradhan Prasanna Pradhan received both BCA and MCA degree from Sikkim Manipal Institute of Technology, Sikkim, India in 2017. Currently he is working as associate application developer in Accenture.