Scalable Non-Linear Graph Fusion for Prioritizing Cancer-Causing Genes

Ekta Shah and Pradipta Maji∗

Abstract—In the past few decades, both gene expression data and protein-protein interaction (PPI) networks have been extensively studied, due to their ability to depict important characteristics of disease-associated genes. In this regard, the paper presents a new gene prioritization algorithm to identify and prioritize cancer-causing genes, integrating judiciously the complementary information obtained from two data sources. The proposed algorithm selects disease-causing genes by maximizing the importance of selected genes and functional similarity among them. A new quantitative index is introduced to evaluate the importance of a gene. It considers whether a gene exhibits a differential expression pattern across sick and healthy individuals, and has a strong connectivity in the PPI network, which are the important characteristics of a potential biomarker. As disease-associated genes are expected to have similar expression profiles and topological structures, a scalable non-linear graph fusion technique, termed as ScaNGraF, is proposed to learn a disease-dependent functional similarity network from the co-expression and common neighbor based similarity networks. The proposed ScaNGraF, which is based on message passing algorithm, efficiently combines the shared and complementary information provided by different data sources with significantly lower computational cost. A new measure, termed as DiColIN, is introduced to evaluate the quality of a learned affinity network. The performance of the proposed graph fusion technique and gene selection algorithm is extensively compared with that of some existing methods, using several cancer data sets.

Index Terms—Disease gene identification, microarray data analysis, protein-protein interaction network, network fusion.

1 INTRODUCTION

Functional genomics has been an area of active research in the past few years. It aims to study the genetic factors that regulate a particular disorder. Currently, the OMIM database lists more than 2000 genetic disorders, which are influenced by mutations in the genetic structure [1]. Identification of such mutated genes, which play an active role in the progression and onset of a disease, has been an area of intense research. This acts as the first step towards gaining a better insight about the functions regulated by them, the pathways in which they are involved and the other genes with which they interact. Their early detection can aid in improving the clinical procedure and thus, increases the survival chances for a patient [2].

Technological advancement in the field of biology has led to the production of a huge volume of diverse data forms, which includes gene expression, protein expression, DNA methylation, protein-protein interaction (PPI) network, and so on. Each of the data types depicts different characteristics of disease related genes. For example, the expression profiles of genes show a difference in behavior over diseased and non-diseased individuals, if the transcription level of the genes has been affected due to the disease [3], [4]. The advent of microarray technology facilitates capturing and monitoring the expression profiles of thousands of genes simultaneously. Several feature selection approaches have been proposed for identification of differentially expressed genes from microarray gene expression data [5], [6]. The major drawback of such approaches lies in the inability to capture functional characteristics of disease-associated genes. The PPI networks possess the ability to depict the confidence in an interaction between two proteins. In [7], [8], it has been shown that the genes, which co-regulate a particular disease, are closely coupled with each other in the PPI network. In this regard, several gene prioritization techniques, based on PPI networks, have been proposed [9]–[11]. The existing methods of gene selection, reported in [5], [6], [9]–[11], utilize either gene expression data or PPI network to identify candidate disease genes. The microarray data based approaches for gene selection fail to harness the interaction affinity among the genes or proteins, while the PPI network based approaches are unable to capture differential expressibility of candidate disease genes.

In order to exploit the complementary information provided by each data source, data integration based methods for gene selection have been evolved [12], [13]. Such techniques are based on the hypothesis that the disease-associated genes or proteins are closely linked to the protein products of differentially expressed and functionally similar genes in the PPI network. Therefore, integrated methods of gene selection utilize both the data types to extract genes related to the disease. In [14], [15], a two-step approach for gene selection has been proposed, where a subset of genes has been extracted from gene expression data, and then given as input to the PPI network to identify genes which act as connectors between each pair of candidate genes. Wu et al. [12] proposed an algorithm which is based on the assumption that disease-causing genes express differential
behavior by network rewiring or networked differential expression. The approach prioritizes genes that are extracted from subnets of differentially expressed genes. Recently, Choi et al. [16] proposed an algorithm, which uses gene expression profiles to weigh the interaction graphs and then applies a modified pageRank algorithm to prioritize disease-associated genes. Santoni et al. [17] proposed a method to integrate PPI network with co-expression data, obtained through microarray analysis. The method is based on separate clustering of two data forms. The clusters so obtained are merged into a special graph and its cliques are considered as the subset of strongly connected proteins.

A regularized logistic regression model has been proposed in [18], where information about the degree of a node is incorporated using an adaptive elastic net to reduce the dependency of developed model on gene expression profiles. In [19], a gene selection approach has been described, which aims to extract genes that are co-expressed in the gene expression data and densely connected to each other in the PPI network. To select disease-related genes, Chen et al. [20] proposed a method that utilizes random walk with restart and shortest path algorithm. However, majority of the data integration based approaches reported above tend to use two data types in a sequential manner. The gene selection algorithms that utilize gene expression profiles and interaction networks simultaneously have been proposed in [21], [22]. These approaches assume that genes associated with the same disorder tend to share similar functional characteristics and are closely connected to one another. Several similarity measures, based on the shared neighborhood between a pair of genes, have been proposed in [21], [22]. The algorithms, reported in [21], [22], curate a set of genes that are differentially expressed and topologically similar, in terms of shared neighborhood.

In existing gene selection approaches, the importance of a gene is quantified using either gene expression data or PPI network. For instance, the gene expression data can be used to measure the relevance of a gene with respect to the class labels, while the PPI networks can be used to compute different topological measures, like degree, betweenness and closeness, that depict the importance of a gene in the network. However, the complementary information provided by the two data sources remains unused. Hence, some new measures need to be defined that can quantify the importance of a gene with respect to both the data sources. Moreover, the functional similarity measures, introduced in [21], [22], are based on PPI networks that are static in nature and independent of the disease under consideration. On the contrary, if dynamic similarity networks are generated by integrating the disease-related information present in the gene expression data sets with the PPI networks, they will definitely be a useful source of information. However, these two data forms have a conflicting scale of representation and an incompatible representation format. For example, gene expression data has a tabular format, while PPI network has a graphical representation. So, approaches that are free from both scale and format of representation are needed, to achieve an unbiased integration of various data types.

Recently, similarity networks have become popular to deal with the heterogeneity associated with different data sources [23]. The gene similarity networks can be computed using each of the data types, thereby, providing a uniform format and scale to represent them [23], [24]. The affinity networks have been used in diverse fields for data integration [23]–[27]. In [25], [27], individual networks have been linearly combined to form a new network, assuming that the complementary information between different data sources is linearly provided, which may not always be true. In [23], a non-linear graph fusion framework, termed as cross-diffusion, has been proposed, where individual affinity networks are combined in a non-linear manner to generate a single unified graph. In [24], another non-linear graph fusion method, termed as similarity network fusion (SNF), has been proposed, where cross-diffusion approach of [23] has been modified to make the updated affinity networks free from the scale of self-similarity. In [28], the SNF has been used to generate a unified graph, on which random walk with restart algorithm is used to prioritize genes on the basis of the association between genes and phenotypes. A robust version of the non-linear graph fusion approach has been proposed in [29], which uses decision tree based clustering random forest on each view of the data to create a robust affinity graph. The final unified graph is derived using SNF on the affinity graphs obtained. Another variant of the graph fusion approach has been presented in [30], where anchor graph representations have been used to depict the similarity between data and anchor points. The diffusion process is performed on the anchor graphs in a later stage of each iteration to generate the updated similarity network. The non-linear graph fusion approaches, reported in [23], [24], [26], have mostly been used on low-dimensional similarity networks, although the gene similarity networks obtained from both gene expression data and PPI network are known to have a high dimension. Hence, the scalability of existing network fusion approaches to such large networks is a major challenge. Some efforts in this regard have been made in [30] through anchor graph representation. However, the iterative procedure hinders its application to large-scale networks.

In this regard, the paper presents a novel gene selection algorithm, which aims to identify and prioritize genes that are relevant to the disease under study. The proposed algorithm is based on the hypothesis that disease-associated genes are differentially expressed and strongly connected to other genes in the PPI network. In order to quantify the importance of a gene, a new quantitative index is introduced, which considers both gene expression data and PPI network to compute the differential expressibility and degree of connectivity of a gene. The disease-associated genes are also known to have similar expression patterns and share a large number of common neighbors in PPI network. A Scalable Non-linear Graph Fusion technique, termed as ScaNGraF, is proposed to learn a functional similarity network, which captures the shared and complementary information provided by the co-expression and common neighborhood based affinity networks. The learned functional similarity network is dynamic in nature, in the sense that it depends on the disease under study. The proposed ScaNGraF is based on message passing algorithm and has a significantly lower computational cost than existing similar graph fusion approaches. A new measure, termed as DicCoIN, is proposed to evaluate the quality of a learned network. The proposed gene selection algorithm prioritizes
potential disease genes, by maximizing the importance of selected genes and functional similarity among them. The performance of the proposed graph fusion approach and gene selection algorithm is demonstrated on four colon cancer data sets, along with a comparison with other related methods of network fusion and gene selection. All the results indicate that the proposed graph fusion technique is an efficient data integration approach and scalable for large networks. Also, the proposed gene selection method is quite promising and may be a useful tool for identification of potential biomarkers.

2 Basics of Cross-Diffusion Process

Given a data set \(X = \{x_i | i \in \{1, ..., n\}\}\), consisting of \(n\) objects, a finite weighted graph \(G = (V, E, W)\) can be constructed, where \(V\) is the set of \(n\) nodes, \(E \subseteq V \times V\) is the set of edges connecting the nodes, and \(W\) is an \(n \times n\) similarity matrix, where \(W(i, j) > 0\) if the nodes \(x_i\) and \(x_j\) are connected to each other in \(G\). The weight value \(W(i, j)\) indicates the level of similarity between two nodes. The weight matrix \(W\) is normalized as follows:

\[
  \mathcal{G}(i, j) = W(i, j) \times \left( \sum_{k=1}^{n} W(i, k) \right)^{-1}, \tag{1}
\]

to compute a full kernel \(\mathcal{G}\) on the graph \(G\) such that \(\sum_{j=1}^{n} \mathcal{G}(i, j) = 1\). A sparse kernel \(\mathcal{L}\) can be defined using \(G\), such that it depicts the local affinity of a particular node. In other words, the \(k\)-nearest neighbors to a node are considered to define its local neighborhood, which is used to construct the \(k\)-nearest neighbor (\(k\)-NN) graph. The pairwise similarity between non-neighboring points is set to zero. Let, \(\mathcal{N}\) be the set of nearest neighbors to node \(x_i\) in \(G\). Thus, the sparse kernel \(\mathcal{L}\) may be defined as follows:

\[
  \mathcal{L}(i, j) = \begin{cases} 
    \frac{W(i, j)}{\sum_{k \in \mathcal{N}_i} W(i, k)}, & \text{if } x_j \in \mathcal{N}_i \\
    0, & \text{otherwise.} 
  \end{cases} \tag{2}
\]

Here, it has been assumed that local similarities are more reliable than non-local ones. Moreover, the local affinity information can be propagated to non-neighbors using a diffusion process, through reliable local connections [23].

Let \(m\) be the number of affinity networks. For \(m = 2\), two similarity networks can be computed, which correspond to two graphs, namely, \(G_1(V, E_1, W_1)\) and \(G_2(V, E_2, W_2)\), respectively, such that \(|V| = n\). Let \(G_1\) and \(G_2\) be the full kernels or status matrices corresponding to graphs \(G_1\) and \(G_2\), respectively, obtained using (1). Similarly, \(\mathcal{L}_1\) and \(\mathcal{L}_2\) are defined as the sparse kernels obtained using (2). Then, the cross-diffusion process is defined as [23]:

\[
  G_1(t + 1) = \mathcal{L}_1 G_2(t) \mathcal{L}_1^T; \quad G_2(t + 1) = \mathcal{L}_2 G_1(t) \mathcal{L}_2^T; \tag{3}
\]

where \(G_1(t)\) and \(G_2(t)\) are the updated status matrices obtained after the \((t - 1)\)-th iteration, \(G_1(1) = G_1\) and \(G_2(1) = G_2\) are two initial status matrices, and \(A^T\) represents the transpose of a matrix \(A\). The process is based on the message passing algorithm [31], which passes the edge information from one network to another via an iterative cross-diffusion process. The procedure is repeated till the final unified graph converges. After \(t\) steps, the final status matrix \(M(t + 1)\) is computed as

\[
  M(t + 1) = \left[ G_1(t + 1) + G_2(t + 1) \right] / 2. \tag{4}
\]

The cross-diffusion approach uses the common neighborhood between the \(k\)-NN graph \(\mathcal{L}\) and full kernel \(G\) to propagate the affinity between nodes. So, the approach is robust to noise associated with similarity measures. It relies more on the local connections, than the remote ones, to assign similarities to non-neighbors. A strong affinity between a pair of nodes may be reflected using a single data type or similarity measure, but may not be supported by other data sources or similarity measures. In such cases, the cross-diffusion approach preserves strong linkages favored by one or more graphs and weak connections are omitted, thereby, reducing noise in the unified graph [23].

3 Scalable Non-Linear Graph Fusion

This section presents a novel non-linear graph fusion approach, termed as ScaNGrAF, to combine the shared and complementary information supported by two or more data sources. The proposed approach is based on the cross-diffusion technique due to Wang et al. [23].

3.1 Integration of Two Networks

In this section, the proposed graph fusion technique is described for two views of the data, that is, for \(m = 2\). Given a finite iteration \(t\), the proposed approach directly computes the updated similarity matrix for that particular \(t\), in a non-iterative manner, with a significantly lower computational cost. Let us assume that \(G_1(1) = G_1\) and \(G_2(1) = G_2\) are two initial full kernels or status matrices, which are computed by normalizing the weight matrix \(W_i\) as follows:

\[
  G_i(j, k) = \begin{cases} 
    \frac{W_i(j, k)}{\sum_{j \neq k} W_i(j, k)}, & \text{if } j \neq k \\
    1/2, & \text{if } j = k. 
  \end{cases} \tag{5}
\]

The initial status matrices are used, along with the sparse kernels \(\mathcal{L}_1\) and \(\mathcal{L}_2\) computed using (2), to get the final unified graph. Based on (3), we have

\[
  G_1(t + 1) = \mathcal{L}_1 G_2(t) \mathcal{L}_1^T = (\mathcal{L}_1 \mathcal{L}_2) G_1(t - 1)(\mathcal{L}_1 \mathcal{L}_2)^T \\
  = (\mathcal{L}_1 \mathcal{L}_2)^2 G_1(t - 3)((\mathcal{L}_1 \mathcal{L}_2)^2)^T \\
  \cdots \\
  = (\mathcal{L}_1 \mathcal{L}_2)^k G_1(t - (2k - 1))((\mathcal{L}_1 \mathcal{L}_2)^k)^T. \tag{6}
\]

Since the initial state of the diffusion process is represented by \(t = 1\), the minimum value for \([t - (2k - 1)]\) must be 1. Therefore, we have

\[
  t - (2k - 1) = 1 \Rightarrow k = \frac{t}{2}. \tag{7}
\]

For \(k\) to be an integer value, \(t\) must be an even number. So, using \(k = t/2\) in (6), \(G_1(t + 1)\) can be rewritten as:

\[
  G_1(t + 1) = (\mathcal{L}_1 \mathcal{L}_2)^{t/2} G_1((\mathcal{L}_1 \mathcal{L}_2)^{t/2})^T. \tag{8}
\]

Similarly, \(G_2(t + 1)\) can be written as follows:

\[
  G_2(t + 1) = (\mathcal{L}_2 \mathcal{L}_1)^{t/2} G_2((\mathcal{L}_2 \mathcal{L}_1)^{t/2})^T. \tag{9}
\]
Now, if \( t \) is an even number, then \( (t + 1) \) must be odd. Let \( t' = t + 1 \), be an odd number. Thus, solving for \( G_1(t' + 1) \) using (3), we have

\[
G_1(t' + 1) = (L_1L_2)^kG_1(t' - (2k - 1))(L_1L_2)^k\top. \tag{10}
\]

As \( t' \) is an odd number and \( (2k - 1) \) must be odd, so \( |t' - (2k - 1)| = 0 \) is not a feasible solution in the proposed approach. Hence, the minimum even value of \( |t' - (2k - 1)| \) is 2. Thus,

\[
t' - (2k - 1) = 2 \Rightarrow k = \frac{t' - 1}{2}. \tag{11}
\]

Here, \( t' \) represents an odd value of \( t \). Thus, we can say that \( k = \frac{t' - 1}{2} \), for odd values of \( t \). So, if \( t \) is an odd number, the updated similarity matrices \( G_1(t + 1) \) and \( G_2(t + 1) \) can be expressed as:

\[
G_1(t + 1) = (L_1L_2)^{\frac{t - 1}{2}}G_1(2)((L_1L_2)^{\frac{1}{2}L_2})\top;
\tag{12}
\]

\[
G_2(t + 1) = (L_2L_1)^{\frac{t - 1}{2}}G_2(2)((L_2L_1)^{\frac{1}{2}L_1})\top;
\tag{13}
\]

Hence, for \( m = 2 \), the final unified similarity graph or network \( M_2(t + 1) \) can be expressed as in (14), by combining (8), (9), (12) and (13). The proposed approach can also be extended for integrating multiple similarity graphs, which is presented in Section S1 of the supplementary material.

**Algorithm 1: Computation of Final Unified Graph**

**Input:** Full kernels \( G_1, G_2 \); sparse or local kernels \( L_1, L_2 \); number of iterations \( t \).

**Output:** Final unified graph \( M_2 \).

1. **Function:** ScaNGraF \((G_1, G_2, L_1, L_2, t)\)
   2. if \( t \) is even then
   3. \( t' \leftarrow t/2 \)
   4. else
   5. \( t' \leftarrow (t - 1)/2 \)
   6. end
   7. \( L_{12} \leftarrow L_1 \times L_2 \)
   8. \( L_{12}' \leftarrow \text{ComputePowers}(L_{12}, t') \)
   9. \( L_{21} \leftarrow L_2 \times L_1 \)
   10. \( L_{21}' \leftarrow \text{ComputePowers}(L_{21}, t') \)
   11. if \( t \) is even then
   12. \( G_1(t + 1) \leftarrow L_{12}' \times G_1 \times L_{12}'\top \)
   13. \( G_2(t + 1) \leftarrow L_{21}' \times G_2 \times L_{21}'\top \)
   14. else
   15. \( G_1(t + 1) \leftarrow (L_{12}'L_1) \times G_2 \times (L_{12}'L_1)\top \)
   16. \( G_2(t + 1) \leftarrow (L_{21}'L_2) \times G_1 \times (L_{21}'L_2)\top \)
   17. end
   18. Normalize \( G_1(t + 1) \) and \( G_2(t + 1) \) using (5)
   19. \( M_2(t + 1) \leftarrow \frac{G_1(t + 1) + G_2(t + 1)}{2} \)
   20. end

Careful analysis of (14) shows that the final unified graph can be expressed in terms of two initial status matrices \( G_1 \) and \( G_2 \). The proposed network integration technique is outlined in Algorithm 1. It can be seen that the fused network can be directly computed using two initial full kernels \( G_1 \) and \( G_2 \) and two local affinity matrices \( L_1 \) and \( L_2 \). Algorithm 2 outlines a cost-effective technique to raise a given matrix to a particular power. Instead of the brute-force approach for computing the powers of a matrix, the algorithm makes use of logarithmic scale to serve the purpose. If the power of a similarity matrix is computed in a brute-force manner, the proposed algorithm needs to perform at most \( \left(\frac{k}{4}\right) \) matrix multiplication, which is still better than the \( t \) iterations needed by the existing cross-diffusion approach.

**Algorithm 2: Computation of Powers of a Matrix**

**Input:** Matrix \( B \); power \( t \).

**Output:** Matrix \( B^t \).

1. **Function:** ComputePowers \((B, t)\)
   2. \( p \leftarrow 2 \)
   3. repeat
   4. \( B^p \leftarrow B^{p/2} \times B^{p/2} \)
   5. \( p \leftarrow 2 \times p \)
   6. until \( p < t \)
   7. \( p \leftarrow p/2 \)
   8. \( d \leftarrow t - p \)
   9. repeat
   10. \( B^d \leftarrow \text{ComputePowers}(B, d) \)
   11. \( B^{p+d} \leftarrow B^p \times B^d \)
   12. \( p \leftarrow p + d \)
   13. \( d \leftarrow t - p \)
   14. until \( d > 0 \)
   15. end

**3.2 Complexity Analysis**

In this section, the performance of the proposed graph fusion technique is compared with that of cross-diffusion approach proposed in [23], for \( m = 2 \). Careful analysis of (3) shows that the updation of each similarity network requires two matrix multiplication, where every such operation has a worst case computational complexity of \( O(n^3) \). So, updation of two status matrices, in every iteration, involves a computational cost of \( O(4n^3) \). Now, if the graph fusion technique continues for \( t \) steps, the total computational cost for the cross-diffusion approach is \( O(4tn^3) \).

On the other hand, the proposed network fusion technique performs the product of two sparse matrices raised to some power with the initial status matrices. In Section 2, it is described that \( L_1 \) and \( L_2 \) are two sparse matrices, where the maximum number of non-zero elements is \( (kn) \). Thus, the multiplication of two sparse matrices has a computational complexity of \( O(kn^2) \) [32]. In the proposed approach, \( (L_1L_2) \) and \( (L_2L_1) \) are needed to update \( G_1 \) and \( G_2 \), respectively. The total cost of computing the product of two sparse matrices is \( O(2kn^2) \). Henceforth, the matrices \( (L_1L_2) \) and \( (L_2L_1) \) are treated as dense matrices, since the product of two sparse matrices may lead to the generation of a dense matrix. In the next step, \( (L_1L_2) \) and \( (L_2L_1) \) are raised to a power of almost \( \frac{1}{2} \). This involves a cost of \( O(\frac{1}{2}n^3) \). Analysis of (14) shows that for odd values of \( t \), the number of matrix multiplication performed is higher than that for even values of \( t \). Thus, in the worst case, six matrix multiplication need to be performed, having a
computational complexity of $O(6n^3)$. Therefore, the proposed network fusion technique has an overall computational complexity of $[O(2knp^2) + O((\frac{k}{n})^3) + O(6n^3)]$. Given that the overall computational complexity of the proposed graph fusion approach is less than that of the existing cross-diffusion algorithm [23, 24], we have

$$2kn^2 + (t/2)n^3 + 6n^3 < 4nt^3 \Rightarrow 2k/n + t/2 + 6 < 4t. \quad (15)$$

As $k \ll n$, $\frac{k}{n} \approx 0$. Hence, (15) can be rewritten as

$$t/2 + 6 < 4t \Rightarrow t \geq 2. \quad (16)$$

Therefore, it can be seen that the proposed network fusion approach computes the final unified graph with relatively lower computational cost if $t \geq 2$. Careful analysis of Section 3.1 also shows that the updated status matrices that can be obtained for $t = 2$ are $G_1(3)$ and $G_2(3)$. It is evident from (14) that the updated similarity matrices can be expressed in terms of the initial status matrices using (14), only if $t \geq 2$.

### 3.3 Evaluation of Learned Network Quality

The cross-diffusion algorithm, explained in Section 2, as well as proposed ScanNGraf, introduced in Section 3.1, consider two sparse kernels for network fusion. Each sparse kernel can be constructed from an affinity matrix using $k$-nearest neighbor rule. So, the fused network depends on the value of $k$. Another important parameter of the graph fusion technique is the number of iterations $t$, over which the affinity matrices are updated. So, the fused network, learned from the individual affinity graphs, depends on both $k$ and $t$. However, none of the existing network fusion approaches have reported how to determine the optimal values of these two parameters.

In this regard, a new quantitative index, termed as DiColIN, is introduced to evaluate the quality of a learned network. The measure is based on the Difference in Contribution of Individual Networks (DiColIN) in the learned affinity network and is defined as follows:

$$\text{DiColIN} = \left\| \frac{T_1 - T_2}{T_M} \right\|_T,$$  \quad (17)

where $T_1$ and $T_2$ represent the number of edges contributed by the individual affinity networks $W_1$ and $W_2$, respectively, in the final fused network $M_2$, $T_M$ denotes the total number of edges in $M_2$, and $\|T_1 + T_2\| \leq T_M$. Some basic properties of DiColIN are described as follows:

1. $\text{DiColIN} = 0$ iff $T_1 = T_2$, when $T_1, T_2 > 0$;
2. $\text{DiColIN} = 1$ if either $T_1 = T_M$ or $T_2 = T_M$; and
3. $\text{DiColIN} \in [0, 1]$.

Property 1 says that DiColIN achieves the lowest value 0 only if two individual networks have an equal contribution. So, low values of DiColIN specify that the learned affinity network is unbiased and has an almost equal contribution from both the affinity networks. Moreover, Property 2 indicates that DiColIN attains its highest value of 1, when the learned similarity network is formed from the edges contributed by a single affinity network. Thus, a high value for DiColIN indicates that the learned network is heavily biased towards a particular affinity graph.

The contribution of each network, in terms of $T_p$, in the final affinity network, is computed as [24]:

$$T_p = \begin{cases} T_p + 1, & \text{if } \Delta(i, j) > \Delta \text{ and } \delta_p(i, j) < \delta_k(i, j); \\ T_p, & \text{otherwise}; \end{cases} \quad (18)$$

where $\delta_p(i, j)$ denotes the change in affinity observed for an edge $(i, j)$ in the network $M_2$, with respect to the $p$th network $W_p$, $\Delta(i, j)$ depicts the difference in $\delta_p(i, j)$ between two individual networks, $\Delta = (0.1, \text{ for example})$ is the threshold, and $p \in \{1, 2\}$. In the current study, $\delta_p(i, j)$ and $\Delta(i, j)$ are defined as follows:

$$\delta_p(i, j) = \frac{|M_2(i, j) - W_p(i, j)|}{M_2(i, j)}, \quad (19)$$

and

$$\Delta(i, j) = |\delta_1(i, j) - \delta_2(i, j)|. \quad (20)$$

As the learned affinity network depends on the parameters $k$ and $t$, the contribution of individual networks in the final affinity network varies with $k$ and $t$. So, the DiColIN measure can be used to evaluate the quality of a fused network as well as to find the optimum values of network parameters. The optimum value of $(k, t)$, denoted as $(k^*, t^*)$, is obtained using the following relation:

$$(k^*, t^*) = \arg\min_{(k,t)} \{\text{DiColIN}\}, \quad (21)$$

where $k \in [k_{\min}, k_{\max}]$ and $t \in [t_{\min}, t_{\max}]$. Here, $k_{\min}$ and $k_{\max}$ denote the minimum and maximum number of nearest neighbors, considered in (2), while $t_{\min}$ and $t_{\max}$ denote the minimum and maximum number of iterations, over which the affinity networks are updated.

### 4 Prioritization of Disease-causing Genes

This section presents a novel gene selection algorithm, which utilizes the complementary information provided by the two data sources. A new quantitative index is introduced to evaluate the importance of a gene. The ScanNGraf, introduced in Section 3, is used to learn a disease-dependent functional similarity network by integrating co-expression and common neighbor based similarity networks. The spectral clustering algorithm is used to partition the functional similarity network into multiple clusters and a reduced set $R$ of effective genes is selected. Finally, the proposed algorithm curates a subset of potential biomarkers from $R$, by maximizing the importance of the selected genes and functional similarity among them.

#### 4.1 Importance of Gene

Both gene expression data and PPI network have an important role in identifying potential disease genes. The
relevance of a gene, which can be computed using gene expression data, is one of the most frequently used measures to depict the differential expressibility of a gene. On the other hand, degree of a gene is extensively used to capture its topological structure in the PPI network. Although these two measures are important in identifying disease genes, they are not always correlated with each other. In other words, differentially expressed genes may have a low degree in the PPI network, while genes with a high degree may have a uniform expression pattern.

Thus, a new measure, termed as importance of gene (IoG), is introduced here. It utilizes the information supported by the two data forms, to prioritize genes that are differentially expressed and have a strong connectivity in the PPI network. For a given gene expression data set $E$ and PPI network $\mathbb{P}$, the importance of a gene $A_i$ is defined as

$$\text{IoG}(A_i) = \alpha \hat{\gamma}(A_i, \mathbb{D}) + (1 - \alpha) \overline{D}(A_i),$$  

(22)

where $\hat{\gamma}(A_i, \mathbb{D})$ denotes the relative relevance of a gene $A_i$ with respect to the set of class labels $\mathbb{D}$, $\overline{D}(A_i)$ denotes the relative degree of the gene $A_i$ in the given PPI network, and $\alpha \in [0, 1]$ is a weight parameter, controlling the relative importance of relevance and degree of a gene.

A gene expression data can be represented as $E = \{e_{ij}\}_{n \times p}$ where $n$ denotes the number of genes, $p$ denotes the number of samples, and $e_{ij}$ is the expression level of a gene $A_i$ in the $j$-th sample. Let $\mathbb{C} = \{A_1, ..., A_n\}$ denote the set of $n$ genes, while $\mathbb{D}$ denote the set of class labels, corresponding to $p$ samples. The relevance of a gene $A_i$ with respect to the class labels $\mathbb{D}$ can be computed as

$$\gamma(A_i, \mathbb{D}) = I(A_i, \mathbb{D}),$$  

(23)

where $I(A_i, \mathbb{D})$ represents the mutual information between gene $A_i$ and set of class labels $\mathbb{D}$. It represents the change in entropy associated with $A_i$ given the information about $\mathbb{D}$. It can be seen that the mutual information between $A_i$ and $\mathbb{D}$ is maximum if the conditional entropy of $A_i$ given $\mathbb{D}$ is minimum. The relative relevance of a gene $A_i$, with respect to a set $\mathcal{R} \subseteq \mathbb{C}$, is defined as:

$$\hat{\gamma}(A_i, \mathbb{D}) = \frac{\gamma(A_i, \mathbb{D}) - \min_{A_j \in \mathcal{R}} \{\gamma(A_j, \mathbb{D})\}}{\max_{A_j \in \mathcal{R}} \{\gamma(A_j, \mathbb{D})\} - \min_{A_j \in \mathcal{R}} \{\gamma(A_j, \mathbb{D})\}}.$$  

(24)

So, the value of $\hat{\gamma}(A_i, \mathbb{D})$ lies in the interval $[0, 1]$ and depends on the reduced set $\mathcal{R}$.

On the other hand, a weighted PPI network is represented as $\mathbb{P} = \{\omega_{ij}\}_{n \times n}$, where $\omega_{ij} \in [0, 1]$ represents the weight value for an edge between genes $A_i$ and $A_j$. It denotes the confidence in the interaction between two genes. The degree of a node or gene in PPI network is defined as

$$D(A_i) = \|\omega_i\|_0; \quad \|\omega_i\|_0 = \lim_{p \to 0} \sum_{j=1}^{n} |\omega_{ij}|^p$$  

(25)

is called the $L_0$-norm. It is the number of non-zero elements in the vector $\omega_i$. The relative degree $\overline{D}(A_i)$ of a gene $A_i$, with respect to the set $\mathcal{R} \subseteq \mathbb{C}$, is defined as follows:

$$\overline{D}(A_i) = \frac{D(A_i) - \min_{A_j \in \mathcal{R}} \{D(A_j)\}}{\max_{A_j \in \mathcal{R}} \{D(A_j)\} - \min_{A_j \in \mathcal{R}} \{D(A_j)\}}.$$  

(26)

It can easily be inferred that $\overline{D}(A_i)$ also depends on the reduced set $\mathcal{R}$ and lies in the range $[0, 1]$. On the basis of (24) and (26), some properties of the proposed IoG measure, defined in (22), can be stated as follows:

1) $0 \leq \text{IoG}(A_i) \leq 1$
2) $\text{IoG}(A_i) = 1$, if and only if $\hat{\gamma}(A_i, \mathbb{D}) = 1$ and $\overline{D}(A_i) = 1$, for $\alpha \in (0, 1)$.
3) $\text{IoG}(A_i) = 0$, if and only if $\hat{\gamma}(A_i, \mathbb{D}) = 0$ and $\overline{D}(A_i) = 0$, for $\alpha \in (0, 1)$.

The analysis of (22) shows that the parameter $\alpha$ regulates the importance of relative relevance and relative degree of the gene under study. If $\alpha = 1$, then the proposed measure depends only on the relative relevance of the gene to the class labels. In such cases, the IoG prioritizes genes based on their relevance values and does not take their topological connectivity into consideration. On the other hand, if $\alpha = 0$, then the relative degree of a gene in the PPI network is considered, thereby, failing to capture the differential expressibility of a gene. Thus, $0 < \alpha < 1$ must hold to select genes that are differentially expressed and have a high degree of connectivity.

### 4.2 Proposed Gene Selection Algorithm

In this section, a novel gene selection algorithm is proposed, which prioritizes potential biomarkers by maximizing the importance of the selected genes and functional similarity among them. In order to quantify the importance of a gene, a new measure, termed as IoG, is introduced in Section 4.1, while a scalable graph fusion approach, termed as ScaNGrAF, is proposed in Section 3 to compute the functional similarity among the genes. The proposed ScaNGrAF is used to learn a new similarity network from two individual affinity networks $W_1$ and $W_2$. The proposed gene selection algorithm selects genes by maximizing the following objective function:

$$J = J_{\text{IoG}} + J_{\text{sim}},$$  

(27)

where

$$J_{\text{IoG}} = \sum_{A_j \in S} \text{IoG}(A_j)$$

$$J_{\text{sim}} = \alpha \sum_{A_j \in S} \hat{\gamma}(A_i, \mathbb{D}) + (1 - \alpha) \sum_{A_j \in S} \overline{D}(A_j)$$

(28)

$$J_{\text{deg}} = \alpha J_{\text{rel}} + (1 - \alpha) J_{\text{deg}}.$$

Here, $J_{\text{rel}}$ and $J_{\text{deg}}$ denote the total relative relevance and total relative degree of the selected set of genes $S$, respectively, while $J_{\text{sim}}$ depicts the total functional similarity within the set $S$, which is defined as follows:

$$J_{\text{sim}} = \sum_{A_i \neq A_j \in S} \mathcal{M}_2(A_i, A_j).$$  

(29)

In order to solve the optimization problem reported in (27), a greedy search approach is used here, which is presented in Algorithm 3. The proposed algorithm can be broadly divided into three main stages. Firstly, the gene expression data and PPI network are used to compute the co-expression and common neighbors based similarity network, $W_1$ and $W_2$, respectively. In the present study, mutual information is used as a measure of co-expression between a pair of genes $A_i$ and $A_j$, represented as $I(A_i, A_j)$. So,

$$W_1(A_i, A_j) = I(A_i, A_j),$$  

(30)
where \( I(A_i, A_j) \) can be computed as reported in [22]. The affinity network \( W_2 \) can be computed using the weighted PPI network based similarity measure, represented as \( S(A_i, A_j) \), defined in [22]. Thus,

\[
W_2(A_i, A_j) = S(A_i, A_j). \tag{31}
\]

The detailed procedure for generation of the functional similarity networks is reported in Section S3 of the supplementary material. The ScaNGraF is used to compute the unified graph from the individual networks. In the next stage, the proposed IoG measure is used to quantify the importance of the genes and select the gene \( A_j \) of highest importance. Finally, the algorithm evaluates the IoG of every gene \( A_i \in C \) and its functional similarity with respect to the set of selected genes \( S \), to iteratively prioritize potential disease genes.

**Algorithm 3: Proposed Gene Selection Algorithm**

**Input**: Set of genes \( C = \{A_1, ..., A_n\} \); gene expression data \( E \); weighted PPI network \( \mathbb{P} \).

**Output**: Set of \( d \) potential disease related genes \( S \).

1. **Function**: PrioritizeGenes\((C, E, \mathbb{P})\)
2. Compute \( G_1 \) using (30) and (5) based on \( E \)
3. Compute \( G_2 \) using (31) and (5) based on \( \mathbb{P} \)
4. Compute \( L_1 \) and \( L_2 \) from \( G_1 \) and \( G_2 \), using (2)
5. \( M_2 = \text{ScaNGraF}(G_1, G_2, L_1, L_2, t) \)
6. Set \( S \leftarrow \emptyset \)
7. Compute \( \text{IoG}(A_i) \), \( \forall A_i \in C \), using (22)
8. \( A_j \leftarrow \arg \max_{A_i \in C} \{ \text{IoG}(A_i) \} \)
9. \( S \leftarrow S \cup A_j \) and \( C \leftarrow C \setminus A_j \)
10. repeat
11. for \( A_i \in C \) do
12. \( J_{sim}(A_i, S) \leftarrow 0 \)
13. for \( A_k \in S \) do
14. if \( M_2(A_i, A_k) > \epsilon \), then
15. \( J_{sim}(A_i, S) = J_{sim}(A_i, S) + \frac{M_2(A_i, A_k)}{\max_{A_j \neq A_i} M_2(A_i, A_j)} \)
16. end
17. end
18. \( J_t \leftarrow \text{IoG}(A_i) + \frac{J_{sim}(A_i, S)}{|S|} \)
19. \( A_{max} \leftarrow \arg \max_{A_i \in C} \{ J_t \} \)
20. \( S \leftarrow S \cup A_{max} \)
21. \( C \leftarrow C \setminus A_{max} \)
22. until \(|S| < d| \)
23. end.

Step 14 of Algorithm 3 uses a threshold \( \epsilon \) while updating \( J_{sim}(A_i, S) \) that denotes the total similarity of gene \( A_i \) with respect to the set \( S \). The threshold \( \epsilon \) is used to increase the robustness of the algorithm by ignoring weak similarities among the genes. In the present work, the average similarity of the learned affinity network is used as the threshold \( \epsilon \) for discarding the weak affinities. In Step 15, Algorithm 3 normalizes the affinity between genes \( A_i \in C \) and \( A_k \in S \) to ensure that the functional similarity between them is comparable with the IoG measure of a gene under consideration. The computational complexity of Algorithm 3 is presented in Section S2 of the supplementary material.

### 4.3 Clustering on Functional Similarity Network

Among the large number of genes present in both gene expression data and PPI network, only a small fraction of them contain disease related information. Moreover, genes which exhibit similar functional behavior tend to be co-expressed and possess similar topological connectivity. So, disease related genes, which are co-expressed and have co-functional behavior, must be closely connected to each other in the final unified network \( M_2 \), learned using ScaNGraF. Hence, a reduced set of effective genes \( \mathcal{R} \) can be extracted from the complete set \( C \), based on \( M_2 \), without compromising the disease related information. The reduced set \( \mathcal{R} \) is then used from Step 7 of Algorithm 3, instead of the complete set \( C \), to identify disease related genes.

The spectral clustering algorithm [33] is used to partition the set of genes \( C \) into \( c \) clusters, based on eigen decomposition of unified network \( M_2 \). The eigengap heuristic, defined as the maximum possible difference between two consecutive eigenvalues, is used to determine the optimal value of \( c \). The present study is based on the hypothesis that disease-associated genes are either directly connected to each other or indirectly connected through some hub nodes in the PPI network. The hub nodes are known to play an active role in disease onset and progression. So, clusters having a higher number of hub genes are expected to contain disease related information. In the current study, the cluster that has highest number of hub genes is considered to identify disease-associated genes. It reduces the search space and time complexity of the gene selection algorithm.

### 5 Experimental Results and Discussion

This section presents the performance of the proposed gene prioritization algorithm, which is based on a non-linear graph fusion approach, termed as ScaNGraF, and a new measure, called IoG. So, the performance of the proposed ScaNGraF is compared with that of the individual affinity networks as well as SNF [24]. A comparative study of the potential disease genes, predicted using the complete and reduced set of genes, is also undertaken to establish the importance of extracting a reduced set of effective genes from the learned affinity network. The proposed gene selection criterion and the IoG measure are extensively studied to establish their efficiency in curating potential disease genes. Finally, the performance of the proposed gene prioritization algorithm is compared with that of some existing data integration based approaches for gene selection, namely, MR+PPIN [2], mMR+PPIN [14], MRMS+PPIN [19], RelSim [21], CPR [16], CLAIM [17], PeC [19], NGP [12] and SiFS [22]. The software version of the proposed algorithm is available at www.isical.ac.in/~bibil/results/scangraf/scangraf.html.

Four colon cancer gene expression data sets, namely, GSE25070, GSE44861, GSE10950 and GSE24514, obtained from the NCBI GEO repository [34] are used in the present study, along with the human protein interactions derived from the STRING 3.0 database [35]. Each microarray data set is pre-processed by standardizing each sample to zero mean and unit variance. The values of \( k_{\min} \), \( k_{\max} \), \( t_{\min} \) and \( t_{\max} \), reported in Section 3.3, are set to 50, 200, 5 and 50, respectively. The number of nearest neighbors \( k \) is varied in the interval of 50, while the number of iterations \( t \) is
incremented in gap of 5. The quality of the learned network is evaluated for different combinations of \( k \) and \( t \), using the DiCoIN measure introduced in Section 3.3, and the optimal parameter values \((k^*, t^*)\) are \((100, 5)\), \((50, 10)\), \((50, 5)\) and \((50, 10)\) for GSE25070, GSE44861, GSE10950 and GSE24514, respectively.

Three known colon cancer gene lists, namely, List A, List B and List C, are used to analyze the performance of the curated gene sets. Here, List A is a collection of 438 genes prepared by Sabates-Bellver et al. [36], while List B comprises 134 genes, prepared by Nagaraj and Reverter [37]. List C consists of colon cancer related 264 genes that are obtained from the disease ontology (DO) database [38]. Four more gene lists are also considered, which are formed by merging the above three lists as follows: List AB = List A \( \cup \) List B, List AC = List A \( \cup \) List C, List BC = List B \( \cup \) List C and List ABC = List A \( \cup \) List B \( \cup \) List C. Moreover, the biological process (BP) of gene ontology (GO) and KEGG pathway based enrichment analysis are performed using the ClueGO v1.8 [39], while the DO based analysis is performed using the DOSE package [38].

### 5.1 Effectiveness of Proposed ScaNGraF Approach

In the present study, the proposed ScaNGraF is used to learn a disease-dependent functional similarity network, from two individual affinity networks corresponding to gene expression data and PPI network. In order to establish the effectiveness of ScaNGraF, the performance of the fused similarity network obtained using ScaNGraF is compared with that of SNF and two individual affinity networks, namely, co-expression (CE) and common neighbor (CN) networks. Spectral clustering is performed on each of the affinity networks to extract a reduced set of genes, which is further used to curate the set of 200 top-ranked genes. The comparative performance analysis is reported in Fig. 1 and Table 1 for four colorectal cancer data sets. Fig. 1 compares the performance of four affinity networks, in terms of degree of overlapping with the seven disease related gene lists. Careful analysis of Fig. 1 shows that the functional similarity network obtained using ScaNGraF performs better than the individual affinity networks in 23 cases out of total 28 cases. Moreover, it performs better than the SNF in 19 cases out of total 28 cases, while in 7 out of the remaining 9 cases, the performance of SNF is comparable with that of ScaNGraF. Overall, the ScaNGraF exhibits a significantly better performance in 19 out of 28 cases.

Table 1 reports the biological analysis of the gene sets curated using the four networks. Careful analysis of the results shows that the genes selected by the ScaNGraF based network annotate to disease-associated BP terms and KEGG pathways with significantly lower p-values. Extensive analysis of these terms in the supplementary document establishes their role in the growth and progression of the disease. From Table 1, it can also be inferred that the ScaNGraF annotates to disease-associated DO term with p-values, lower than that of the other networks. A detailed analysis of the annotated terms to colorectal cancer is reported in the supplementary document.

### 5.2 Computational Efficiency of ScaNGraF

The present section aims to establish the computational efficiency of the proposed ScaNGraF approach. The perfor-
Fig. 2. Execution time of ScaNGraF, CrDP and SNF.

Fig. 3. Variation in speed-up with respect to computational nodes.

Table 1: GO, KEGG and DO Based Analysis of Top 200 Genes Identified Using Proposed and Other Gene Selection Approaches

<table>
<thead>
<tr>
<th>Data</th>
<th>Methods</th>
<th>Biological Process: Term and P-Value</th>
<th>KEGG Pathway: Term and P-Value</th>
<th>DO: Term and P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE25070</td>
<td>CE</td>
<td>regulation of growth</td>
<td>1.85E-07</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>response to oxygen-containing compound</td>
<td>3.75E-24</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>SNF</td>
<td>response to lipid</td>
<td>1.75E-14</td>
<td>TNF signaling pathway</td>
</tr>
<tr>
<td></td>
<td>Full Gene Set</td>
<td>response to lipid</td>
<td>1.68E-14</td>
<td>HIF-1 signaling pathway</td>
</tr>
<tr>
<td>Criteria</td>
<td>Relevance</td>
<td>one-carbon metabolic process</td>
<td>1.52E-04</td>
<td>Nitrogen metabolism</td>
</tr>
<tr>
<td></td>
<td>IoG</td>
<td>response to nitrogen compound</td>
<td>2.05E-15</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>response to oxygen-containing compound</td>
<td>5.87E-25</td>
<td>TNF signaling pathway</td>
</tr>
<tr>
<td>GSE44861</td>
<td>CE</td>
<td>DNA metabolic process</td>
<td>7.16E-19</td>
<td>Mismatch repair</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>cellular response to chemical stimulus</td>
<td>5.75E-42</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>SNF</td>
<td>regulation of cell migration</td>
<td>2.01E-22</td>
<td>AGE-RAGE signaling pathway</td>
</tr>
<tr>
<td></td>
<td>Full Gene Set</td>
<td>response to oxygen-containing compound</td>
<td>1.24E-21</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Criteria</td>
<td>Relevance</td>
<td>flavonoid metabolic process</td>
<td>3.21E-08</td>
<td>Nitrogen metabolism</td>
</tr>
<tr>
<td></td>
<td>IoG</td>
<td>response to endogenous stimulus</td>
<td>1.24E-21</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>regulation of cell motility</td>
<td>7.47E-24</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>GSE10950</td>
<td>CE</td>
<td>gland development</td>
<td>5.35E-08</td>
<td>Hepatitis B</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>enzyme linked receptor protein signaling pathway</td>
<td>6.01E-12</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>SNF</td>
<td>enzyme linked receptor protein signaling pathway</td>
<td>4.73E-12</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>Full Gene Set</td>
<td>enzyme linked receptor protein signaling pathway</td>
<td>4.97E-12</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>Criteria</td>
<td>Relevance</td>
<td>cytokinesis</td>
<td>4.97E-03</td>
<td>Starch and sucrose metabolism</td>
</tr>
<tr>
<td></td>
<td>IoG</td>
<td>enzyme linked receptor protein signaling pathway</td>
<td>2.56E-11</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>regulation of cellular component movement</td>
<td>6.62E-21</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>GSE24514</td>
<td>CE</td>
<td>muscle contraction</td>
<td>1.17E-11</td>
<td>cGMP-PKG signaling pathway</td>
</tr>
<tr>
<td></td>
<td>CN</td>
<td>cell cycle process</td>
<td>3.89E-60</td>
<td>Cell cycle</td>
</tr>
<tr>
<td></td>
<td>SNF</td>
<td>regulation of programmed cell death</td>
<td>1.09E-20</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td></td>
<td>Full Gene Set</td>
<td>mitotic cell cycle process</td>
<td>2.62E-18</td>
<td>Cell cycle</td>
</tr>
<tr>
<td>Criteria</td>
<td>Relevance</td>
<td>mitotic cell cycle process</td>
<td>1.39E-11</td>
<td>Cell cycle</td>
</tr>
<tr>
<td></td>
<td>IoG</td>
<td>cell proliferation</td>
<td>5.05E-24</td>
<td>Cell cycle</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>regulation of cell proliferation</td>
<td>2.20E-01</td>
<td>Pathways in cancer</td>
</tr>
</tbody>
</table>

Performance of the ScaNGraF is compared with that of SNF and cross-diffusion process (CrDP), in terms of execution time and speed-up. The application of non-linear graph fusion approaches to large-scale affinity networks generally takes a large amount of time. In order to address this shortcoming, the openMP C programming interface is used, which enables to develop a multi-threaded implementation of the proposed ScaNGraF approach. In the present study, an Intel(R) Xeon(R) CPU E7-4890 v2 @ 2.80GHz, x86_64 server computational facility is used. It consists of 40 CPU cores and each core contains 40 threads. Thus, a multi-threaded implementation of the proposed ScaNGraF is developed, which can effectively utilize the computational potential of the system.
proposed algorithm over its single-thread counterpart. Careful analysis of Fig. 3 shows that increasing the number of computational nodes or execution threads aids in achieving a significant speed-up in learning the fused affinity network using ScaNGraF. A steep increase in speed-up is attained by gradually increasing the number of execution threads from 5 to 50. Additionally, Table 2 reports the execution time, speed-up and efficiency of the proposed ScaNGraF algorithm. In the present study, the multi-thread based parallel execution is performed using 25 computational threads. Analysis of Table 2 shows that the multi-thread implementation speeds up the algorithm by almost 10 times and has an efficiency of around 40 percent.

5.3 Importance of Network Clustering
The proposed gene selection method considers a reduced set of genes, instead of the complete gene set, to identify disease-causing genes. The reduced set is obtained by performing spectral clustering on the functional similarity network, learned using ScaNGraF. In order to demonstrate the effectiveness of the proposed approach, extensive experiments are performed on four colon cancer data sets. Fig. 4 compares the performance of the proposed gene selection algorithm on the reduced set of effective genes with that of the complete gene set, in terms of overlap with the known gene lists. The results of Fig. 4 show that the set of potential biomarkers, extracted using the reduced set, performs significantly better than that of the complete set in all the cases.

Table 1 also reports the comparative study between the reduced and full gene set, using GO, KEGG and DO. The analysis of Table 1 shows that the former gene set outperforms the latter by curating potential biomarkers that annotate to disease-associated BP, KEGG pathways and DO terms with the lowest p-value in all 12 cases. The association of the annotated terms, obtained using the reduced set, to colorectal cancer is reported in the supplementary document. The results reported in Fig. 4 and Table 1 establish the fact that the set of genes curated from the reduced set is highly enriched with the colorectal cancer related genes.

5.4 Significance of Proposed Criterion
The proposed method selects genes from the reduced set by maximizing the importance of genes (IoG) and functional similarity (FS) among them. To establish the importance of the proposed criterion (IoG+FS) over two related criteria, namely, relevance and IoG, extensive experimentation is carried out on four data sets, and corresponding results are reported in Fig. 5 and Table 1. All the results reported in Fig. 5 show that the proposed IoG measure performs better than the relevance criterion in 22 out of total 28 cases, while the proposed criterion (IoG+FS) outperforms both IoG and relevance in 26 out of total 28 cases.
Table 1 reports the biological analysis of the genes curated using three gene selection criteria. The analysis shows that the IoG criterion annotates to disease-relevant BP and KEGG pathway terms with a significantly low p-value. However, the proposed criterion (IoG+FS) outperforms both IoG and relevance using BP, KEGG pathway and DO based analysis by annotating disease-associated terms, as explained in supplementary document, with significantly lower p-values.

### 5.5 Comparative Performance Analysis

Finally, the performance of the proposed algorithm is compared with that of MR+PPIN [2], mMR+PPIN [14], MRMS+PPIN [15], RelSim [21], SiFS [22], CPR [16], CLAIM [17], PeC [19], and two different models of NGP [12], namely, NGP-ND and NGP-NR. Fig. 6 compares the performance of the proposed and existing gene selection algorithms, in terms of overlap with the known disease related gene lists. The graphs reported in Fig. 6 show that the proposed algorithm outperforms all the existing approaches of gene selection in 24 cases out of total 28 cases. Among the remaining 4 cases, the RelSim performs better than the proposed algorithm with respect to List C and List BC, for GSE25070 data and the proposed algorithm exhibits a performance comparable to that of the SiFS for GSE10950 with respect to List A and List AB.

Table 3 compares the performance of different gene selection algorithms using GO, KEGG and DO based analysis. Careful analysis of Table 3 shows that some existing algorithms, like CPR and NGP-ND, annotate to different BP terms with p-values lower than the proposed algorithm in majority of the cases. However, the terms do not bear any relevant association to the disease. On the contrary, the proposed algorithm annotates to BP terms, which bear a much stronger association to the disease, as explained in supplementary document, with a significantly low p-value. Careful study of Table 3 using the KEGG pathway enrichment analysis demonstrates the efficiency of the proposed algorithm in annotating disease-associated KEGG terms with significantly low p-values. The DO based comparative study reported in Table 3 shows that the proposed algorithm outperforms the existing gene selection algorithms in 3 out of 4 cases. From the results reported in Fig. 6 and Table 3, it can be inferred that the genes selected using the proposed algorithm have a significantly higher overlap with the known disease related genes and annotate to more relevant disease-associated processes and pathways, in comparison to the existing algorithms.

In order to establish the effectiveness of different gene selection algorithms in curating disease-causing genes, a new measure, based on the GoD package in R [40], is introduced next. The GoD package computes a score corresponding to a gene, based on its annotation to the disease under study. Higher score denotes a better association of a gene with the disease under consideration. So, the GoD score can be used to rank the selected genes based on their association to the disease ontology. The generated GoD scores, which range from 0 to 1, are grouped into 11 bins in intervals of 0.1. In other words, the zeroth bin contains genes with GoD scores equal to 0, the first bin contains genes with GoD scores ranging in the interval [0, 0.1], the second bin contains genes with scores lying in the interval (0.1, 0.2] and so on. Finally, the tenth bin contains genes whose GoD scores range in the interval (0.9, 1.0]. In general, an algorithm that maximizes
the gene frequency in the higher indexed bins must be considered to be more effective. Thus, the bin frequency is computed corresponding to each of the gene selection algorithms. The bin frequencies are then normalized such that their sum equals to 1. Let \( \mathcal{F}_i \) denote the normalized frequency of the \( i \)-th GoD bin. In order to judge the quality of a candidate gene set, a weighted score called \( WS_{\text{GoD}} \) based on the normalized bin frequency \( \mathcal{F}_i \) of GoD, is defined as follows:

\[
WS_{\text{GoD}} = 2 \times \sum_{i=0}^{10} Q_i \mathcal{F}_i;
\]

where \( Q_i \) is the weight assigned to the \( i \)-th bin and

\[
Q_i = \begin{cases} 
1/2^{(11-i)}, & \text{if } i = 2, 3, \ldots, 10 \\
1/2^9, & \text{if } i = 1 \\
0, & \text{if } i = 0.
\end{cases}
\]

It must be noted that \( WS_{\text{GoD}} \in [0, 1] \) and its higher value denotes a better association of the gene set with the disease under study. In the present study, the genes selected using RelSim, SiFS, CPR, CLAIM, PeC, NGP-ND, NGP-NR and the proposed algorithm are used to compute the \( WS_{\text{GoD}} \), and corresponding results are reported in Table 4. Analysis of the results shows that the proposed algorithm attains the highest value of \( WS_{\text{GoD}} \) in 3 out of 4 cases, which establishes the effectiveness of the proposed algorithm in curating a set of potential biomarkers.

6 CONCLUSION

The paper presents a novel gene selection algorithm for prioritizing disease related genes. The algorithm judiciously integrates the shared and complementary information provided by the gene expression data and PPI network to extract the diseased genes. The genes are selected by maximizing their importance and functional similarity among them. A new quantitative index is introduced to quantify the importance of a gene. The proposed measure is based on the differential expressibility of a gene as well as its connectivity in the PPI network. A scalable non-linear graph fusion approach, termed as ScaNGraf, is proposed to learn a disease-dependent functional similarity network, with a significantly low computational cost. The co-expression and common neighbor based similarity networks are used to capture the complementary information provided by the two data sources and learn the dynamic affinity network. A new quantitative index is introduced to evaluate the quality of a learned network.

Extensive experiment is performed to establish the effectiveness of the proposed algorithm, using some colorectal cancer data sets. The proposed and existing algorithms are evaluated based on their ability to capture known disease-related genes. In order to gain an insight about the different disease-related processes and pathways in which the selected gene set is involved, an extensive biological significance analysis is performed using gene ontology, KEGG pathway enrichment analysis and disease ontology. A new quantitative index, based on the GoD package, has been introduced to study the association of the curated gene set with the disease under study. Careful analysis of the reported results establishes the effectiveness of the proposed gene selection algorithm in curating potential biomarkers. The importance of the learned affinity network and IoG measure in prioritizing potential disease genes has also
been established. The scalability of the proposed ScaNGraF algorithm to high dimensional data sets and its computational efficiency have been demonstrated. The proposed gene selection algorithm has been shown to be a useful technique for large-scale network fusion and disease gene identification.

### Table 3

Go, KEGG and DO Based Analysis of Top 200 Genes Identified Using Proposed and Existing Gene Selection Algorithms

<table>
<thead>
<tr>
<th>Data</th>
<th>Methods</th>
<th>Biological Process: Term and P-Value</th>
<th>KEGG Pathway: Term and P-Value</th>
<th>DO: Term and P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE25070</td>
<td>MR+PPIN</td>
<td>intracellular steroid hormone receptor signaling pathway</td>
<td>7.92E-13</td>
<td>Thyroid hormone signaling pathway 2.01E-05</td>
</tr>
<tr>
<td>GSE25070</td>
<td>mRMR+PPIN</td>
<td>cellular response to steroid hormone stimulus</td>
<td>4.26E-11</td>
<td>Thyroid hormone signaling pathway 1.44E-13</td>
</tr>
<tr>
<td>GSE25070</td>
<td>MRMS+PPIN</td>
<td>regulation of protein phosphorylation</td>
<td>1.64E-21</td>
<td>Thyroid hormone signaling pathway 1.12E-08</td>
</tr>
<tr>
<td>GSE4861</td>
<td>RelSim</td>
<td>response to oxygen-containing compound</td>
<td>1.56E-35</td>
<td>Pathways in cancer 2.86E-13</td>
</tr>
<tr>
<td>GSE4861</td>
<td>SIFS</td>
<td>positive regulation of macromolecular metabolic process</td>
<td>3.94E-27</td>
<td>Pathways in cancer 3.93E-08</td>
</tr>
<tr>
<td>GSE4861</td>
<td>CPR</td>
<td>regulation of fibroblast proliferation</td>
<td>2.04E-53</td>
<td>Pathways in cancer 3.09E-04</td>
</tr>
<tr>
<td>GSE4861</td>
<td>CLAIM</td>
<td>pigment metabolic process</td>
<td>1.07E-02</td>
<td>Chemokine signaling pathway 3.35E-02</td>
</tr>
<tr>
<td>GSE4861</td>
<td>PeC</td>
<td>positive regulation of response to stimulus</td>
<td>5.81E-26</td>
<td>Epstein-Barr virus infection 1.91E-18</td>
</tr>
<tr>
<td>GSE4861</td>
<td>NGP-NR</td>
<td>protein phosphorylation</td>
<td>8.79E-25</td>
<td>Pathways in cancer 1.05E-19</td>
</tr>
<tr>
<td>GSE4861</td>
<td>Proposed</td>
<td>response to oxygen-containing compound</td>
<td>6.81E-25</td>
<td>TNF signaling pathway 2.29E-10</td>
</tr>
</tbody>
</table>

| GSE25070 | MR+PPIN | RNA secondary structure unwinding | 1.95E-07 | * | colorectal cancer 1.26E-01 |
| GSE25070 | mRMR+PPIN | regulation of cell proliferation | 2.08E-08 | * | colorectal cancer 1.85E-01 |
| GSE25070 | MRMS+PPIN | regulation of cell death | 6.89E-21 | * | colorectal cancer 1.38E-05 |
| GSE10950 | SIFS | regulation of cell death | 2.18E-46 | * | colorectal cancer 1.18E-14 |
| GSE10950 | CLAIM | positive regulation of response to stimulus | 1.49E-39 | * | colorectal cancer 1.34E-13 |
| GSE10950 | CPR | pigment metabolic process | 3.18E-04 | Mineral absorption 2.57E-02 | colorin neoplasms 7.57E-02 |
| GSE10950 | PeC | regulation of cell death | 8.66E-03 | * | optic atrophy 4.73E-01 |
| GSE10950 | NGP-NR | positive regulation of cellular metabolic process | 2.12E-44 | Epstein-Barr virus infection 1.67E-27 | colorectal cancer 1.10E-07 |
| GSE10950 | NGP-NR | regulation of cell motility | 5.81E-40 | Pathways in cancer 1.43E-24 | colorectal cancer 7.98E-13 |
| GSE24514 | MR+PPIN | regulation of cell death | 1.57E-09 | * | colorectal cancer 1.73E-01 |
| GSE24514 | mRMR+PPIN | regulation of cellular component movement | 5.27E-09 | * | colorectal cancer 1.77E-01 |
| GSE24514 | MRMS+PPIN | regulation of cellular component movement | 8.66E-13 | * | colorectal cancer 1.80E-01 |
| GSE24514 | RelSim | regulation of cellular component movement | 8.19E-09 | * | colorectal cancer 1.80E-01 |
| GSE24514 | SIFS | regulation of cellular component movement | 4.73E-16 | * | colorectal cancer 1.59E-06 |
| GSE24514 | CLAIM | regulation of cellular component movement | 4.86E-13 | * | colorectal cancer 1.59E-06 |
| GSE24514 | CPR | regulation of cellular component movement | 5.80E-24 | * | colorectal cancer 1.59E-06 |
| GSE24514 | PeC | regulation of cellular component movement | 5.17E-21 | * | colorectal cancer 1.59E-06 |
| GSE24514 | NGP-NR | regulation of cellular component movement | 5.80E-24 | * | colorectal cancer 1.59E-06 |

### Table 4

GoD Based Analysis of Top 200 Selected Genes

<table>
<thead>
<tr>
<th>Different Algorithms</th>
<th>GSE25070</th>
<th>GSE4861</th>
<th>GSE10950</th>
<th>GSE24514</th>
</tr>
</thead>
<tbody>
<tr>
<td>RelSim</td>
<td>0.1832</td>
<td>0.1070</td>
<td>0.0171</td>
<td>0.1243</td>
</tr>
<tr>
<td>SIFS</td>
<td>0.1173</td>
<td>0.1206</td>
<td>0.0013</td>
<td>0.0824</td>
</tr>
<tr>
<td>CPR</td>
<td>0.1488</td>
<td>0.1481</td>
<td>0.1480</td>
<td>0.1487</td>
</tr>
<tr>
<td>CLAIM</td>
<td>0.0000</td>
<td>0.0438</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>PeC</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NGP-ND</td>
<td>0.1247</td>
<td>0.1286</td>
<td>0.1409</td>
<td>0.1369</td>
</tr>
<tr>
<td>NGP-NR</td>
<td>0.1492</td>
<td>0.1462</td>
<td>0.1172</td>
<td>0.1243</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.1470</td>
<td>0.1842</td>
<td>0.1716</td>
<td>0.1881</td>
</tr>
</tbody>
</table>

### References

6. F. Barrenas, S. Chavali, P. Holme, R. Mobini, and M. Benson, “Network Properties of Complex Human Disease Genes Identified...”


Ekta Shah received the BTech and MTech degree in Computer Science from West Bengal University of Technology, India in 2012 and Indian Statistical Institute, Kolkata in 2015, respectively. Currently, she is a senior research fellow in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her research interests include pattern recognition, machine learning, computational biology and bioinformatics, and so forth. She has published a few papers in international journals and conferences. She has received the 2nd Prize in the Eighth IDRBT Doctoral Colloquium in 2018.

Pradipita Maji (Senior Member, IEEE) received the BSc degree in Physics, the MSc degree in Electronics Science, and the PhD degree from Jadavpur University, India, in 1998, 2000, and 2005, respectively. Currently, he is a Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include machine learning, pattern recognition, medical imaging, computational biology and bioinformatics, and so forth. He has published around 150 papers in international journals and conferences. He is a Fellow of the National Academy of Sciences, India. He received the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory, India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, the 2011 Young Scientist Award from the Indian National Science Academy, India, and the 2015 Young Faculty Research Fellowship from the Ministry of Electronics and Information Technology, Government of India. He has been selected as the 2009 Young Associate of the Indian Academy of Sciences, India.