# Supplementary: Scalable Non-Linear Graph Fusion for Prioritizing Cancer-Causing Genes

Ekta Shah and Pradipta Maji*

## S1 Extension for Multiple Networks

The network fusion technique, proposed in the main paper for two similarity graphs, can easily be extended for more than two such networks. Let $\mathcal{G}_1, ..., \mathcal{G}_i, ..., \mathcal{G}_m$ be $m$ different initial status matrices, generated using $m$ different similarity metrics and $\mathcal{L}_1, ..., \mathcal{L}_i, ..., \mathcal{L}_m$ represent the corresponding sparse kernels generated from the individual similarity graphs. Using the naive cross-diffusion approach [1], the status matrices are updated in the following manner:

$$\mathcal{G}_i(t+1) = \mathcal{L}_i \left( \frac{1}{m-1} \sum_{j=1, j \neq i}^{m} \mathcal{G}_j(t) \right) \mathcal{L}_i^{\intercal}; \qquad (1)$$

where $i = 1, ..., m$. So, an individual affinity network is updated by diffusing the local neighborhood information into each of the previously updated full kernels. The final unified graph can be obtained as follows:

$$\mathcal{M}(t+1) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{G}_i(t+1). \qquad (2)$$

The decomposition of the iterative procedure shown in (1) leads to the generation of several higher order terms, consisting of multiple sparse kernels. Such terms increase the computational complexity of the fusion algorithm, but tend to have a relatively low impact on the updated similarity matrices. To overcome the limitations associated with the application of (1) to large networks, the analysis presented in Section 3.1 of the main paper is extended for multiple networks. Therefore, the current work presents a low computational cost approach for multiple network fusion. Instead of updating each status matrix separately, the final unified graph is computed using (3). It is the generalization of (14) of the main paper and approximates the unified graph of (2) by ignoring the higher order terms.


- *E. Shah and P. Maji are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {ekta_r,pmaji}@isical.ac.in*


- *This publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation.(Asterisk indicates corresponding author)*


## S2 Computational Complexity

In Algorithm 3, which is reported in Section 4.2 of the main paper, Step 2 and Step 3 compute individual affinity networks $\mathcal{G}_1$ and $\mathcal{G}_2$ for $n$ genes, respectively. The cost of computing the co-expression matrix $\mathcal{G}_1$ is $\mathcal{O}(n^2)$, since it takes a constant time to compute the mutual information between a pair of genes. As reported in [2], the cost of computing the shared neighborhood based similarity between a pair of genes is $\mathcal{O}(n_0^2)$, where $n_0$ is the average number of neighbors to a node in the network under study. Thus, the cost of computing the similarity matrix $\mathcal{G}_2$ is $\mathcal{O}(n^2 n_0^2)$. Step 4 of the algorithm needs the sparse kernels $\mathcal{L}_1$ and $\mathcal{L}_2$, computation of which incurs a cost of $\mathcal{O}(kn^2)$. From Section 4.2 of the main paper, it is known that the cost of computing the fused affinity network in Step 5, using the global and sparse networks, is $\mathcal{O}(2kn^2) + \mathcal{O}(\frac{t}{2}n^3) + \mathcal{O}(6n^3)$. The computation of IoG in Step 7 depends on the relative relevance and relative degree of each gene. The cost of computing the relevance of all the genes in $\mathbb{C}$ is $\mathcal{O}(n)$ and that of degree is $\mathcal{O}(n^2)$. Thus, the overall time complexity of computing IoG for all $n$ genes is $\mathcal{O}(n^2)$. The selection of most important gene from $\mathbb{C}$ has a complexity of $\mathcal{O}(n)$. If $\tilde{n}$ denotes the cardinality of the already-selected set of genes $\mathbb{S}$, then normalization done in Step 15 for each gene $\mathcal{A}_i \in \mathbb{C}$ takes $\mathcal{O}(n)$ time and the computation of $J_{sim}(\mathcal{A}_i, \mathbb{S})$ for each gene in $\mathbb{C}$ takes $\mathcal{O}(\tilde{n})$ time. So, Steps 11-18 have a total computational complexity of $\mathcal{O}(\tilde{n}(n-\tilde{n})) + \mathcal{O}(n(n-\tilde{n}))$. The selection a gene $\mathcal{A}_{max}$, from the set of $(n-\tilde{n})$ genes, incurs a cost of $\mathcal{O}(n-\tilde{n})$. Let $d$ denote the desired number of genes to be selected from $\mathbb{C}$, and its selection incurs a total computational cost of $\mathcal{O}(d\tilde{n}(n-\tilde{n})) + \mathcal{O}(dn(n-\tilde{n}))$. Therefore, the overall computational complexity of the proposed gene selection algorithm is $\mathcal{O}(n^2) + \mathcal{O}(n^2 n_0^2) + \mathcal{O}(kn^2) + \mathcal{O}(\frac{t}{2}n^3) + \mathcal{O}(6n^3) + \mathcal{O}(n^2) + \mathcal{O}(d\tilde{n}(n-\tilde{n})) + \mathcal{O}(dn(n-\tilde{n})) \approx \mathcal{O}(n^2 n_0^2) + \mathcal{O}(\frac{t}{2}n^3)$, since $k, d, \tilde{n} \ll n$.

## S3 Generation of Similarity Networks

Generation of similarity networks is one of the most vital stages of the proposed gene selection algorithm. Its importance can be attributed to the fact that they provide a uniform format and scale to represent the information of gene expression data and PPI network. The affinity network constructed using the former data depicts the co-expression

$$\mathcal{M}_m(t+1) = \begin{cases} \frac{1}{m(m-1)} \left[ \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} (\mathcal{L}_i \mathcal{L}_j)^{\frac{t}{2}} \mathcal{G}_i ((\mathcal{L}_i \mathcal{L}_j)^{\frac{t}{2}})^{\intercal} \right], & \text{if } t \text{ is even;} \\ \frac{1}{m(m-1)} \left[ \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} ((\mathcal{L}_i \mathcal{L}_j)^{\frac{t-1}{2}} \mathcal{L}_i) \mathcal{G}_j ((\mathcal{L}_i \mathcal{L}_j)^{\frac{t-1}{2}} \mathcal{L}_i)^{\intercal} \right], & \text{if } t \text{ is odd.} \end{cases} \tag{3}$$

between a pair of genes, while the latter depicts the interactive neighborhood based similarity among them. It must be noted that the network formed using the PPI network is static in nature and independent of disease, while the co-expression network is dependent on the disease. The individual affinity networks can be combined to learn a new functional similarity network that depicts the co-expression and shared connectivity based similarity among the genes.

Given the gene expression data, an affinity network can be generated using various metrics. However, the importance of mutual information, as a measure of co-expression between a pair of genes, has been established in [2]–[6]. The mutual information between a pair of genes $\mathcal{A}_i$ and $\mathcal{A}_j$, represented as $I(\mathcal{A}_i, \mathcal{A}_j)$, is considered as the $(i, j)$-th entry of the symmetric, non-negative weight matrix $W_1(i, j)$. So,

$$W_1(\mathcal{A}_i, \mathcal{A}_j) = I(\mathcal{A}_i, \mathcal{A}_j), \tag{4}$$

where $I(\mathcal{A}_i, \mathcal{A}_j)$ can be computed as reported in [2].

According to the "guilt by association" principle, co-functional genes tend to be closely associated to each other in the PPI network and share physical interactions. Based on this principle, a similarity measure, which depends on the direct and indirect interactions between a pair of genes or proteins, has been proposed in [2] using a weighted PPI network $\mathbb{P}$. The similarity $\mathcal{S}(\mathcal{A}_i, \mathcal{A}_j)$ between genes $\mathcal{A}_i$ and $\mathcal{A}_j$ is defined as [2]:

$$\mathcal{S}(\mathcal{A}_i, \mathcal{A}_j) = \frac{\sum\limits_{\mathcal{A}_k \in \mathcal{N}_{ij}} \min\{\omega_{ik}, \omega_{jk}\}}{\sqrt{\sum\limits_{\mathcal{A}_k \in \mathcal{N}_i} \omega_{ik} * \sum\limits_{\mathcal{A}_k \in \mathcal{N}_j} \omega_{jk}}}; \tag{5}$$

where $\mathcal{N}_i$ and $\mathcal{N}_j$ denote the set of neighbors linked to genes $\mathcal{A}_i$ and $\mathcal{A}_j$, respectively, $\mathcal{N}_{ij}$ denotes the set of common neighbors between genes $\mathcal{A}_i$ and $\mathcal{A}_j$, and $\omega_{ik} \in [0, 1]$ is the weight value for an edge connecting gene $\mathcal{A}_k \in \mathcal{N}_i$ to gene $\mathcal{A}_i$ in the PPI network $\mathbb{P}$. So, the similarity between genes $\mathcal{A}_i$ and $\mathcal{A}_j$ can be computed using (5) from a weighted PPI network, which represents the $(i, j)$-th entry in the symmetric, non-negative matrix $W_2$, as follows:

$$W_2(\mathcal{A}_i, \mathcal{A}_j) = \mathcal{S}(\mathcal{A}_i, \mathcal{A}_j). \tag{6}$$

In the proposed gene selection method, the individual affinity networks, namely, $W_1$ and $W_2$, are combined using ScaNGraF to learn a new affinity network $\mathcal{M}_2$ using (14) of main paper. The final unified network retains the strong similarities supported by atleast one network and the weak affinities favored by both the networks. The learned network possesses the ability to depict both co-expression and shared neighborhood based similarity among the genes. Also, it is dynamic in nature as affinity between a pair of genes depends on the disease under study.

## S4 Clustering on Similarity Network

The spectral clustering algorithm [7] can be used to partition an affinity network, by an optimal graph cut. Given an affinity network $[\mathcal{M}_2]_{n\times n}$, which is to be partitioned into $c$ clusters, an indicator vector $Y_j = [y_{1j}, ..., y_{ij}, ..., y_{nj}]$, corresponding to the $j$-th cluster, can be defined as follows:

$$y_{ij} = \begin{cases} 1, & \text{if the } i\text{-th node belongs to the } j\text{-th cluster;} \\ 0, & \text{otherwise;} \end{cases} \tag{7}$$

where $i = 1, ..., n$ and $j = 1, ..., c$. Thus, a matrix $[Y]_{n\times c}$ can be defined using the label indicator vectors. In [8], it has been shown that the label indicator vectors can be used to minimize the normalized ratio cut. However, it has been shown that normalized ratio cut minimization problem can be reduced to the standard trace minimization problem under certain relaxations, which can be expressed as

$$\min_{H \in \Re^{n\times c}} trace(H^{\intercal} \mathfrak{L} H) \quad \text{subject to } H^{\intercal} H = I; \tag{8}$$

where $H = \mathbb{M}^{1/2} Y$, $\mathbb{M}$ is a diagonal matrix, such that $\mathbb{M}(i, i) = \sum\limits_{j=1}^{n} \mathcal{M}_2(i, j)$, $\mathfrak{L}$ is the normalized laplacian matrix, defined as $\mathfrak{L} = \mathbb{M}^{-1/2} \mathrm{L} \mathbb{M}^{-1/2}$, and $\mathrm{L} = \mathbb{M} - \mathcal{M}_2$, is the laplacian matrix for the similarity matrix $\mathcal{M}_2$. Based on Rayleigh-Ritz theorem, the minimization problem of (8) can be solved by the matrix $H$, which contains the first $c$ eigenvectors of $\mathfrak{L}$ [8]. The eigengap heuristic, defined as the maximum possible difference between two consecutive eigenvalues, is used to determine the optimal value of $c$. The set of genes is partitioned into $c$ clusters, using the spectral clustering algorithm reported in [7].

## S5 Description of Data Sets Used

Four colon cancer gene expression data sets, namely, GSE25070, GSE44861, GSE10950 and GSE24514, obtained from the NCBI GEO repository [9] are used in the present study, along with the human protein interactions derived from the STRING 3.0 database [10].

1) **GSE25070:** It is the gene expression data retrieved from study of Hinoue et al. [11]. The data set contains the gene expression profiles of 26 colorectal tumors matched histological to normal adjacent colonic tissue samples. Illumina Ref-8 whole-genome expression BeadChip with 24526 probes corresponding to 18491 genes was used to obtain the gene expression profiles.
2) **GSE44861:** It is a gene expression data from the tissues of colon cancer patients, which is obtained using the Affymetrix HT Human Genome U133A Array. It contains the expression profiles of 111 colonic samples that were obtained from 56 tumors and 55 adjacent noncancerous tissues. It contains the

expression profiles for 22277 probes that correspond to 13496 genes [12].

3) **GSE10950:** It is the gene expression data put forward by Jiang et al. [13]. It compares the expression profiles of 24 colon tumor tissues to histological matched normal tissue samples. The Illumina humanRef-8 v2.0 expression beadchip, which comprises of 22185 probes corresponding to 18197 genes, was used to extract the expression profiles.
4) **GSE24514:** It is the gene expression data of human MSI colorectal cancer and normal colonic mucosa, prepared by Alhopuro et al. [14] using the Affymetrix Human Genome U133A Array. It contains the expression profiles of 34 MSI colorectal cancers and 15 normal colonic mucosas. The Affymetrix array contains the expression data for 22283 probes, which map to 12985 genes.
5) **STRING 3.0:** It is a large online database resource providing both experimental and predicted protein interactions, along with a confidence score. The nodes in the graph correspond to proteins, while edge denotes the association between two proteins. Experimental repositories, computational prediction methods, etc., have been used to derive direct and indirect interactions. The confidence in any interaction represents the probability with which a pair of nodes may be present together in a metabolic process. The confidence in the interaction between two proteins is used as the weightage of each interaction in the current study.

Each microarray data set is pre-processed by standardizing each sample to zero mean and unit variance. The genes that are common to both the PPI network and gene expression data only are further considered. It has been observed that GSE25070, GSE44861, GSE10950 and GSE24514 have 12475, 10364, 13008 and 10413 genes, respectively, in common with the PPI network. The present study is based on the assumption that disease-associated genes are differentially expressed and closely connected to each other. So, the genes with a low variance, low differential expressibility and weak connectivity are discarded from the gene sets obtained above. The genes that have a variance below 0.5% of the maximum variance in the set are discarded from further consideration. In order to remove genes that exhibit poor differential expressibility and low connectivity in the PPI network, their average relevance and average degree of connectivity are considered as thresholds. Hence, the genes that have both their relevance and degree below the thresholds are discarded from further processing. Thus, it leads to the generation of a reduced set of 5485, 5859, 7901 and 5706 genes for GSE25070, GSE44861, GSE10950 and GSE24514 data, respectively. The gene sets obtained after pre-processing the data sets are used for the generation of individual networks and learning the final unified network.

## S6 Experimental Results and Discussion

This section presents the performance of different algorithms, in terms of cellular components (CC) and molecular functions (MF) of gene ontology (GO). The biological significance for the generated set of genes is analyzed using the ClueGO v1.8 [15]. ClueGO computes enrichment score for the GO terms identified, based on hypergeometric distribution. The significance of any GO term to a group of genes is represented using the corrected p-value. A lower p-value denotes a higher significance of annotated term.

The proposed gene prioritization algorithm 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 [16]. 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.

Table S1 reports the comparative study between the gene sets curated using the individual affinity networks, SNF and ScaNGraF based learned affinity networks. Extensive analysis of Table S1 shows that the proposed ScaNGraF annotates to disease causing CC and MF terms with the lowest p-values using the four cancer data sets. The individual networks and the SNF based network also annotate to CC and MF terms with significantly low p-values. Careful analysis of the results reported in Table S1 establishes the superiority of the two non-linear graph fusion approaches over the individual affinity networks. The improvement can be attributed to the fact that the complementary information, provided by the individual affinity networks, is integrated into the learned functional similarity networks. The analysis of the reported results establishes the improvement obtained using the ScaNGraF based network over the individual and SNF based affinity networks.

Table S1 compares the performance of the genes curated using the full and reduced set of genes. Extensive analysis of Table S1 shows that the proposed reduced set based approach annotates to disease-associated MF terms with the lowest p-values using each of the four data sets. Table S1 also reports the comparative study between the proposed criterion (IoG+FS) and the two related criteria, namely, relevance and IoG. Careful analysis of the results show that the proposed criterion annotates to disease associated CC and MF terms with the lowest p-values. The table also establishes the importance of the IoG criterion by annotating to disease-associated CC and MF terms with significantly low p-values.

Finally, Table S2, compares the performance of the proposed algorithm with that MR+PPIN [17], mRMR+PPIN [5], MRMS+PPIN [18], RelSim [6], SiFS [2], CPR [19], CLAIM [20], PeC [21], and two different models of NGP [22], namely, NGP-ND and NGP-NR with respect to CC and MF terms. Careful analysis of the results shows that the CC and MF terms annotated by the proposed algorithm bear a much relevant association to the disease. Moreover, the proposed algorithm annotates to these terms with a significantly lower p-value.

TABLE S1
Gene Ontology Based Biological Analysis of Top 200 Genes Identified Using Proposed and Other Gene Selection Approaches

| Data | Methods | | Cellular Components: Term and P-Value | | Molecular Functions: Term and P-Value | |
|---|---|---|---|---|---|---|
| GSE25070 | Network | CE | cytoplasmic side of plasma membrane | 1.05E-02 | metalloendopeptidase activity | 2.29E-05 |
| | | CN | secretory granule | 2.74E-05 | hormone activity | 1.88E-06 |
| | | SNF | secretory granule | 7.44E-05 | hormone activity | 1.51E-06 |
| | Full Gene Set | | serine/threonine protein kinase complex | 4.87E-02 | glycosaminoglycan binding | 2.28E-04 |
| | Criteria | Relevance | lamellipodium membrane | 1.35E-03 | glycosaminoglycan binding | 1.04E-05 |
| | | IoG | nuclear euchromatin | 6.66E-03 | glycosaminoglycan binding | 1.74E-06 |
| | Proposed | | extracellular space | 4.76E-23 | cytokine receptor binding | 1.99E-10 |
| GSE44861 | Network | CE | chromosome | 2.18E-11 | unfolded protein binding | 5.88E-06 |
| | | CN | cell surface | 6.25E-11 | receptor binding | 7.07E-26 |
| | | SNF | extracellular matrix | 4.96E-06 | transcription factor binding | 4.18E-07 |
| | Full Gene Set | | extracellular matrix | 8.62E-09 | transcription factor binding | 1.75E-06 |
| | Criteria | Relevance | nuclear euchromatin | 2.93E-02 | carbonate dehydratase activity | 1.45E-05 |
| | | IoG | cytoplasmic vesicle lumen | 2.45E-07 | transcription factor binding | 9.54E-06 |
| | Proposed | | cytoplasmic vesicle lumen | 2.18E-08 | transcription factor binding | 4.52E-07 |
| GSE10950 | Network | CE | myelin sheath | 1.33E-04 | protein binding involved in cell adhesion | 4.65E-04 |
| | | CN | cell-substrate junction | 7.28E-06 | transcription factor binding | 3.59E-06 |
| | | SNF | cell-substrate junction | 2.42E-05 | transcription factor binding | 3.12E-06 |
| | Full Gene Set | | membrane region | 5.21E-06 | transcription factor binding | 3.20E-06 |
| | Criteria | Relevance | RNA polymerase II transcription factor complex | 2.87E-02 | pyridoxal phosphate binding | 3.05E-02 |
| | | IoG | cell-substrate junction | 5.57E-06 | transcription factor binding | 3.67E-06 |
| | Proposed | | cell surface | 2.24E-11 | RNA polymerase II transcription factor activity, sequence-specific DNA binding | 2.08E-08 |
| GSE24514 | Network | CE | contractile fiber part | 9.01E-10 | actin binding | 1.75E-04 |
| | | CN | chromosome | 1.87E-35 | tubulin binding | 2.37E-07 |
| | | SNF | membrane raft | 1.98E-13 | transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding | 2.03E-06 |
| | Full Gene Set | | chromosomal part | 7.03E-11 | cyclin-dependent protein serine/ threonine kinase regulator activity | 2.88E-05 |
| | Criteria | Relevance | condensed chromosome, centromeric region | 1.18E-08 | chemokine receptor binding | 1.95E-04 |
| | | IoG | chromosomal part | 6.73E-13 | cytokine activity | 1.24E-07 |
| | Proposed | | membrane microdomain | 4.64E-08 | cytokine receptor binding | 2.91E-11 |

CE: Co-Expression based affinity network; CN: Common Neighbor based affinity network

## S7 Biological Significance Analysis of Annotated Terms

The present section reports a detailed biological significance based analysis of different BP, MF and KEGG pathway terms annotated by the proposed algorithm.

### S7.1 Annotated BP Terms

The biological analysis of the BP terms annotated by the proposed algorithm is discussed below:

- **Response to oxygen-containing compound:** The presence of oxygen containing compounds has been shown to cause protein dysfunction and DNA damage, leading to gene mutations and cell death. It also activates signaling pathways, such as NF-$\kappa$B and p38 MAPK, to affect cell proliferation, differentiation, and apoptosis [23].
- **Regulation of cell motility:** It indicates that the selected genes play an active role in promoting cancer cell metastasis through MMPs that regulate cell invasion and motility. They also activate the Wnt-signaling pathway, which leads to proliferation in the intestinal stem cells [24].
- **Regulation of cellular component movement:** This denotes that the set of selected genes acts as key regulators in early phases of tumor growth by controlling the Wnt pathway, which is an important regulator of cell proliferation. Movement of the cellular components is mostly caused due to cell division, which triggers malignancy in tumors, followed by metastases [25].
- **Regulation of cell proliferation:** The genes selected using the proposed approach annotate to the term, thereby indicating their role in the growth of cell population or tumor growth. Genes, like SOX2 and TRIM52 are known to be actively involved in the colorectal cancer cell proliferation through the STAT3 signaling pathway [26].

### S7.2 Annotated MF Terms

The biological analysis of different MF terms annotated by the proposed algorithm is described as follows:

- **Cytokine receptor binding:** The role of cytokines in growth and progression of breast cancer cells has been established in [27]. The study has shown that cytokines on breast cancer cell lines are actively involved in decreased cell-cell association with increased cellular motility, inhibition of cellular proliferation and several morphological changes like cell elongation and decrease in inter-cellular adhesion.
- **Transcription factor binding:** The genes selected using the GSE44861 data set play active role in the dysregulation of transcription factors, which are known to be actively involved in the growth and progression of colorectal cancer cells. Binding of the transcription factors to their receptors is important in order to suppress the proliferative and pro-apoptotic growth in the tumors. Thus, dysregulation of the

TABLE S2
Gene Ontology Based Biological Analysis of Top 200 Genes Identified Using Proposed and Existing Gene Selection Algorithms

| Data | Methods | Cellular Components: Term and P-value | | Molecular Functions: Term and P-value | |
|---|---|---|---|---|---|
| GSE25070 | MR+PPIN | nuclear chromatin | 4.62E-13 | nuclear hormone receptor binding | 7.33E-10 |
| | mRMR+PPIN | nuclear chromatin | 9.75E-18 | RNA polymerase II transcription factor binding | 7.65E-14 |
| | MRMS+PPIN | nuclear chromatin | 8.52E-10 | RNA polymerase II transcription factor binding | 3.63E-08 |
| | RelSim | extracellular space | 3.69E-25 | receptor binding | 5.72E-26 |
| | SiFS | extracellular space | 1.57E-21 | G-protein coupled receptor binding | 9.41E-09 |
| | CPR | catalytic complex | 5.43E-32 | enzyme binding | 2.94E-30 |
| | CLAIM | ruffle | 1.35E-02 | protein kinase C binding | 2.55E-03 |
| | PeC | Cul3-RING ubiquitin ligase complex | 2.31E-02 | * | * |
| | NGP-ND | chromosomal part | 7.35E-19 | enzyme binding | 2.78E-31 |
| | NGP-NR | chromosomal part | 2.70E-11 | enzyme binding | 1.86E-34 |
| | Proposed | extracellular space | 4.76E-23 | cytokine receptor binding | 1.99E-10 |
| GSE44861 | MR+PPIN | * | * | ATP-dependent RNA helicase activity | 2.86E-06 |
| | mRMR+PPIN | * | * | * | * |
| | MRMS+PPIN | * | * | * | * |
| | RelSim | membrane region | 5.59E-06 | transcription factor binding | 1.13E-12 |
| | SiFS | kinetochore | 3.09E-14 | transcription factor binding | 1.18E-10 |
| | CPR | catalytic complex | 1.88E-26 | macromolecular complex binding | 1.59E-21 |
| | CLAIM | microvillus | 2.42E-02 | ligase activity, forming carbon-nitrogen bonds | 1.57E-03 |
| | PeC | Cul3-RING ubiquitin ligase complex | 3.44E-03 | NAD+ ADP-ribosyltransferase activity | 3.06E-04 |
| | NGP-ND | membrane region | 3.19E-16 | enzyme binding | 5.30E-37 |
| | NGP-NR | chromosomal part | 4.49E-18 | enzyme binding | 1.11E-32 |
| | Proposed | cytoplasmic vesicle lumen | 2.18E-08 | transcription factor binding | 4.52E-07 |
| GSE10950 | MR+PPIN | * | * | * | * |
| | mRMR+PPIN | * | * | monosaccharide binding | 5.51E-04 |
| | MRMS+PPIN | * | * | ubiquitin conjugating enzyme activity | 1.99E-04 |
| | RelSim | chromosome | 2.11E-26 | chromatin binding | 1.52E-16 |
| | SiFS | side of membrane | 2.46E-05 | cell adhesion molecule binding | 1.50E-06 |
| | CPR | catalytic complex | 6.30E-36 | enzyme binding | 1.99E-27 |
| | CLAIM | presynaptic membrane | 5.02E-03 | tropomyosin binding | 4.99E-03 |
| | PeC | * | * | * | * |
| | NGP-ND | cell junction | 1.88E-17 | enzyme binding | 2.41E-48 |
| | NGP-NR | cell junction | 1.08E-20 | enzyme binding | 1.85E-30 |
| | Proposed | cell surface | 2.24E-11 | RNA polymerase II transcription factor activity, sequence-specific DNA binding | 2.08E-08 |
| GSE24514 | MR+PPIN | * | * | * | * |
| | mRMR+PPIN | SNARE complex | 6.53E-05 | SNAP receptor activity | 2.67E-05 |
| | MRMS+PPIN | * | * | * | * |
| | RelSim | chromosome | 3.24E-26 | poly(A) RNA binding | 1.23E-17 |
| | SiFS | chromosome | 2.59E-13 | Ran GTPase binding | 8.68E-04 |
| | CPR | catalytic complex | 3.81E-28 | macromolecular complex binding | 1.08E-25 |
| | CLAIM | chromosome, centromeric region | 1.23E-05 | chemokine receptor binding | 6.07E-03 |
| | PeC | Cul3-RING ubiquitin ligase complex | 3.37E-03 | NAD+ ADP-ribosyltransferase activity | 2.77E-04 |
| | NGP-ND | chromosomal part | 1.02E-20 | enzyme binding | 1.34E-34 |
| | NGP-NR | chromosomal part | 4.28E-18 | transcription factor binding | 1.30E-31 |
| | Proposed | membrane microdomain | 4.64E-08 | cytokine receptor binding | 2.91E-11 |

binding mechanism has been shown to adversely affect normal cell functioning [28].

- **RNA polymerase II transcription factor activity, sequence-specific DNA binding:** In [29], it has been shown that transcription factors play an important role in the regulation of genes, like MACC1 and BTF3. RNA polymerase II binds to the promoter of MACC1, which regulates the transcriptional regulation of the receptor tyrosine kinase gene. The transcription process plays an important role in colorectal cancer progression and metastasis.

### S7.3 Annotated KEGG Pathways

A detailed analysis of the KEGG pathway terms annotated by the proposed algorithm is described next.

- **TNF signaling pathway:** The genes selected using the proposed algorithm play an active role in regulating the TNF levels in colon cancer cells. The different variants of TNF have been shown to play an active role in inducing tumor cell apoptosis and accelerating tumor invasion and metastasis. They are also known to regulate colon cancer cell migration and invasion by upregulating the levels of TROP-2 protein [30].
- **Pathways in cancer**: It denotes that the genes selected using the proposed algorithm are involved in various pathways that regulate the growth and progression of cancer.

## REFERENCES

[1] B. Wang, J. Jiang, W. Wang, Z. Zhou, and Z. Tu, "Unsupervised Metric Fusion by Cross Diffusion," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2997–3004.

[2] P. Maji and E. Shah, "Significance and Functional Similarity for Identification of Disease Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 6, pp. 1419–1433, 2017.

[3] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[4] P. Maji, "$f$-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1063–1069, 2009.

[5] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network," *PLoS ONE*, vol. 7, no. 4, p. e33393, 2012.

[6] P. Maji, E. Shah, and S. Paul, "RelSim: An Integrated Method to Identify Disease Genes Using Gene Expression Profiles and PPIN Based Similarity Measure," *Information Sciences*, vol. 384, pp. 110–125, 2017.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 2001, pp. 849–856.

[8] U. Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[9] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: Archive for Functional Genomics Data Sets-Update," *Nucleic Acids Research*, vol. 41, no. 1, pp. 991–995, 2013.

[10] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering, "The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.

[11] T. Hinoue, D. J. Weisenberger, C. P. E. Lange, H. Shen, H.-M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, R. A. E. M. Tollenaar, and P. W. Laird, "Genome-Scale Analysis of Aberrant DNA Methylation in Colorectal Cancer," *Genome Research*, vol. 22, no. 2, pp. 271–282, 2012.

[12] B. M. Ryan, K. A. Zanetti, A. I. Robles, A. J. Schetter, J. Goodman, R. B. Hayes, W.-Y. Huang, M. J. Gunter, M. Yeager, L. Burdette, S. I. Berndt, and C. C. Harris, "Germline Variation in NCF4, an Innate Immunity Gene, is Associated With an Increased Risk of Colorectal Cancer," *International Journal of Cancer*, vol. 134, no. 6, pp. 1399–1407, 2014.

[13] X. Jiang, J. Tan, J. Li, S. Kivimäe, X. Yang, L. Zhuang, P. L. Lee, M. T. W. Chan, L. W. Stanton, E. T. Liu, B. N. R. Cheyette, and Q. Yu, "DACT3 is an Epigenetic Regulator of Wnt/$\beta$-Catenin Signaling in Colorectal Cancer and Is a Therapeutic Target of Histone Modifications," *Cancer Cell*, vol. 13, no. 6, pp. 529–541, 2008.

[14] P. Alhopuro, H. Sammalkorpi, I. Niittymäki, M. Biström, A. Raitila, J. Saharinen, K. Nousiainen, H. J. Lehtonen, E. Heliövaara, J. Puhakka, S. Tuupanen, S. Sousa, R. Seruca, A. M. Ferreira, R. M. W. Hofstra, J.-P. Mecklin, H. Järvinen, A. Ristimäki, T. F. Orntoft, S. Hautaniemi, D. Arango, A. Karhu, and L. A. Aaltonen, "Candidate Driver Genes in Microsatellite-Unstable Colorectal Cancer," *International Journal of Cancer*, vol. 130, no. 7, pp. 1558–1566, 2012.

[15] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, "ClueGO: a Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 2009.

[16] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale," *Nature Methods*, vol. 11, 2014.

[17] P. Maji and S. Paul, *Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics*. Springer-Verlag, London, 2014, p. 304.

[18] S. Paul and P. Maji, "Gene Expression and Protein-Protein Interaction Data for Identification of Colon Cancer Related Genes Using $f$-Information Measures," *Natural Computing*, vol. 15, no. 3, 2016.

[19] J. Choi, S. Park, Y. Yoon, and J. Ahn, "Improved Prediction of Breast Cancer Outcome by Identifying Heterogeneous Biomarkers," *Bioinformatics*, vol. 33, no. 22, pp. 3619–3626, 2017.

[20] D. Santoni, A. Swiercz, A. Żmieko, M. Kasprzak, M. Blazewicz, P. Bertolazzi, and G. Felici, "An Integrated Approach (CLuster Analysis Integration Method) to Combine Expression Data and Protein-Protein Interaction Networks in Agrigenomics: Application on Arabidopsis thaliana," *OMICS: A Journal of Integrative Biology*, vol. 18, no. 2, pp. 155–165, 2014.

[21] M. Li, H. Zhang, J.-X. Wang, and Y. Pan, "A New Essential Protein Discovery Method Based on the Integration of Protein-Protein Interaction and Gene Expression Data," *BMC Systems Biology*, vol. 6, no. 1, pp. 15–24, 2012.

[22] C. Wu, J. Zhu, and X. Zhang, "Integrating Gene Expression and Protein-Protein Interaction Network to Prioritize Cancer-Associated Genes," *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–10, 2012.

[23] Z. Wang, S. Li, Y. Cao, X. Tian, R. Zeng, D.-F. Liao, and D. Cao, "Oxidative Stress and Carbonyl Lesions in Ulcerative Colitis and Associated Colorectal Cancer," *Oxidative Medicine and Cellular Longevity*, vol. 2016, 2016.

[24] K. A. Frewer, A. J. Sanders, S. Owen, N. C. Frewer, R. Hargest, and W. G. Jiang, "A Role for WISP2 in Colorectal Cancer Cell Invasion and Motility," *Cancer Genomics Proteomics*, vol. 10, no. 4, pp. 187–196, 2013.

[25] G. Ciasca, M. Papi, E. Minelli, V. Palmieri, and M. D. Spirito, "Changes in Cellular Mechanical Properties During Onset or Progression of Colorectal Cancer," *World Journal of Gastroenterology*, vol. 22, no. 32, pp. 7203–7214, 2016.

[26] S. Pan, Y. Deng, J. Fu, Y. Zhang, Z. Zhang, X. Ru, and X. Qin, "TRIM52 Promotes Colorectal Cancer Cell Proliferation Through the STAT3 Signaling," *Cancer Cell International volume*, vol. 19, no. 57, 2019.

[27] A. M. Douglas, G. A. Goss, R. L. Sutherland, D. J. Hilton, M. C. Berndt, N. A. Nicola, and C. G. Begley, "Expression and Function of Members of the Cytokine Receptor Superfamily on Breast Cancer Cells," *Oncogene*, vol. 14, p. 661669, 1997.

[28] P. Laissue, "The Forkhead-Box Family of Transcription Factors: Key Molecular Players in Colorectal Cancer Pathogenesis," *Molecular Cancer*, vol. 15, 2019.

[29] M. Juneja, K. Ilm, P. M. Schlag, and U. Stein, "Promoter Identification and Transcriptional Regulation of the Metastasis Gene MACC1 in Colorectal Cancer," *Molecular Oncology*, vol. 7, no. 5, pp. 929 – 943, 2013.

[30] P. Zhao and Z. Zhang, "TNF-$\alpha$ promotes colon cancer cell migration and invasion by upregulating TROP-2," *Onclology Letters*, vol. 15, pp. 3820–3827, 2018.


**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.

**Pradipta 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.