IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2019, accepted August 8, 2019, date of publication August 21, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936794

iMTF-GRN: Integrative Matrix Tri-Factorization for Inference of Gene Regulatory Networks

NISAR WANI[1,2] AND KHALID RAZA[2]
[1]Govt. Degree College Baramulla, University of Kashmir, Srinagar 193101, India
[2]Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India

Corresponding author: Khalid Raza (kraza@jmi.ac.in)

The work of N. Wani was supported by the Teacher Fellowship received from University Grants Commission, Ministry of Human Resources Development, Govt. of India vide letter under Grant F.B 27-(TF-45)/2015 through Faculty Development Programme.

ABSTRACT Gene Regulatory Network (GRN) inference using computational approaches has been a highly pursued problem in bioinformatics. Various approaches have been developed to infer GRNs from gene expression data including statistical, machine learning and information theoretic approaches. However, a large number of regulatory relationships remain unpredicted even in the highly studied model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae*. Besides, the regulatory relationships in higher eukaryotes with large genome sizes, such as humans and mice remain mostly unexplored. Majority of the approaches proposed in the literature on GRNs infer molecular interactions from gene expression data alone, despite the fact that gene expression regulation being a product of sequential interactions of multiple biological processes. To capture more regulatory relationships with higher precision, we apply a data fusion and inference model based on Non-negative Matrix Tri-factorization called integrative matrix tri-factorization for GRN inference (iMTF-GRN) that can integrate the diverse type of biological data in a relational learning framework. We, demonstrate that iMTF-GRN model shows improved accuracy in predicting TF-target and miRNA-target gene regulations and performs comparatively better over other state-of-the-art methods.

INDEX TERMS GRN, NMTF, matrix completion, regulatory networks, matrix factorization.

I. INTRODUCTION

Gene regulation is an essential cellular mechanism by which an organism regulates its metabolism and adapts to different environmental conditions. Genes are regulated at multiple levels of regulatory machinery, but the most basic regulation happens at the transcriptional level, involving transcription factor proteins that influence gene expression by binding to regulatory sequences of genes. A GRN, therefore, establishes links between transcription factors (TF) and their target genes, providing a standard representation for transcriptional regulation. Inferring such regulatory networks will help in the elucidation of biological mechanisms that control various cellular processes. Also, mutations in TF coding genes and regulatory sequences that can disrupt standard transcriptional machinery are better understood by GRN inference (GRNI) from high-throughput genome-wide data.

Reverse engineering the GRNs has gained much interest among researchers over the last decade, as the network inference using computational methods is still not a trivial task. This challenge is partly driven by the availability of largescale genome-wide genomic, transcriptomic, proteomic and other omics data and partly because of the increased experimental noise in the data and dimensionality issues of more genes compared to a small number of samples in gene expression analysis. Further, gene expression levels in eukaryotes are influenced by many genomic factors, such as DNA methylation of promoter regions, post-transcriptional silencing by miRNA expression and mutations in TF coding genes or regulatory sequences, making the inference of GRNs difficult from gene expression data alone. A plethora of methods for reconstruction of GRNs has been proposed. Majority of these methods either use compendia of gene expression data from perturbation experiments or they operate on time series data to build GRNs using dynamic models [1], [2].

Among the most popular approaches to infer gene interactions from gene expression data are correlation based methods [1] which compute the pair-wise similarity betweengenes (e.g., Pearson's correlation coefficient, Spearman's correlation coefficient). Besides correlation, Euclidean distances and

The associate editor coordinating the review of this article and approving it for publication was Sotirios Goudos.

information theoretic approaches including mutual information have been used to predict gene interactions [2]. Inference algorithms such as RELNET [3], ARACNE [4], and CLR [5] predict edges between an interacting pair by assigning weighted scores derived from applying mutual information to gene expression data. Also certain mathematical formalisms in the form of Boolean networks [6], [7] and Ordinary Differential Equations (ODEs) have been applied to time-series gene expression data to derive discreet dynamic networks of gene regulation [8], [9]. More recently machine learning [10]–[12] and artificial neural networks (ANNs) based approaches [13], [14] have been applied to learn gene regulatory interactions from genomewide omics data because of their ability to handle variety of nonlinear functions and their robust handling of noisiness in the biological data.

Causal dependencies between genes using expression data have been modeled using probabilistic graphical models such as Bayesian networks (BNs). BNs are very robust in handling randomness and noise inherent in gene expression data and therefore have been found effective to infer causal relationships between genes. The inference problem is modeled as a joint probability distribution function and uses directed acyclic graphs for network reconstruction [15]. Regression is another popular technique that has been rigorously applied to the reconstruction of GRNs from transcriptomic data. For example, GENIE3 [16] is based on variable selection with ensembles of regression trees and was a star performer in the DREAM4 challenge. Among other regression-based methods, least absolute shrinkage and selection operator (LASSO) is most commonly used for GRN inference [17], [18]. However, none of these approaches perform optimally across all the genomes. Marbach *et al.* (2012) have demonstrated that, of the all 35 methods evaluated, the level of precision sharply drops from in silico and *E.coli* datasets to that of *S. cerevisiae* and more complex eukaryotic genomes (e.g., humans, mice, etc.), owing to the increased size of their genomes and multiple levels of control in gene regulation. As a consequence, integration approaches are being developed to construct a more robust and reliable GRN by including heterogeneous datasets from multiple sources along with gene expression data. Data integration techniques that employ network integration approaches [19], probabilistic graphical models [20]–[22], regression models [23], Multiple kernel learning [24], [25] and Non-negative matrix factorization [26], [27] are being used to combine heterogeneous datasets in a biologically relevant manner to infer regulatory networks. For a detailed account of all these methods refer [28].

Here we apply a semi-supervised learning framework using Non-negative matrix tri-factorization (NMTF) based matrix completion approach (iMTF-GRN) for network inference. Proposed primarily for dimensionality reduction problems, NMTF approximates a high dimensional input data matrix from a product of three low-rank non-negative representations. Besides approximating the input matrix, the low dimensional matrices provide the basis and indications for clustering and co-clustering of the objects related via input relational matrix. This clustering property has been established by its proximity to k-means clustering as explained in [29], [30]. Also because of the flexibility that NMTF offers for simultaneous decomposition of multirelation matrices, it becomes easy to integrate data from other sources of biological relevance, making NMTF a favorable tool to fuse multiple biological data. A significant number of research efforts on data integration using NMTF have been reported in the literature [26], [27]. NMTF has been applied to a range of biological inference tasks, such as Drug-target association prediction [31], and gene-function prediction [32].

We approach the GRN inference problem from the matrix completion perspective. We use the NMTF data integration framework for fusing heterogeneous datasets and subsequently approximate a partially observed gene regulatory network. Initially, the method is applied to a benchmark dataset of *E.coli* from Faith et al. (2007). To improve the reliability of the inferred GRN, we integrated semantic similarity from GO annotations and known PPI interactions. For eukaryotic genomes, we apply this method to infer a post-transcriptional regulatory network between miRNA and target genes from multiomics datasets.

II. METHODS

The NMTF integration framework adopts a multi-relational approach where in multiple data sources are fused together. Given a collection D of n data sources $\{d_1,d_2,\ldots\ldots,d_n\}$, each data source d_i relates a pair of objects(obj_i, obj_j) such that the matrix $R_{ij} \in \mathbb{R}^{n_i \times m_j}$ for $i \neq j$ represents a relation between n objects of type n_i and m objects of type m_j. For example, a matrix containing regulatory relationships between genes and transcription factors. Also R_{ij} and R_{ji} are asymmetric matrices as they relates heterogeneous objects. On the other hand matrices that relate similar objects (i.e., obj_i), such as gene interaction networks or protein interaction networks are represented by Kernel matrices $K_i \in \mathbb{R}^{n_i \times n_i}$. These matrices are negative graph laplacians ($L = A - D$) transformed into diffusion kernels for introducing regularization in the model [26].

The standard NMTF formulation for obtaining a low rank approximation W of a single relational matrix R_{12} as shown in Figure 1 can be approximated by obtaining low rank factors $G_1 \in \mathbb{R}^{n \times k}$ and $S_{12} \in \mathbb{R}^{k \times k}$ and $G_2^{k \times n}$ such that $W \approx G_1 S_{12} G_2^T$ with rank $k \ll n$. The objective function to obtain such a low rank representation of input matrix R_{12} can be written as:

$$\min_{G \geq 0, S \geq 0} J_1 = \left\| R_{12} - G_1 S_{12} G_2^T \right\|^2 \tag{1}$$

Adding another relation matrix R_{12} to the integration framework would then require an update to equation (1), the updated objective function is given as under:

$$\min_{G \geq 0, S \geq 0} J_2 = \left\| R_{12} - G_1 S_{12} G_2^T \right\|^2 + \left\| R_{13} - G_1 S_{13} G_3^T \right\|^2 \tag{2}$$

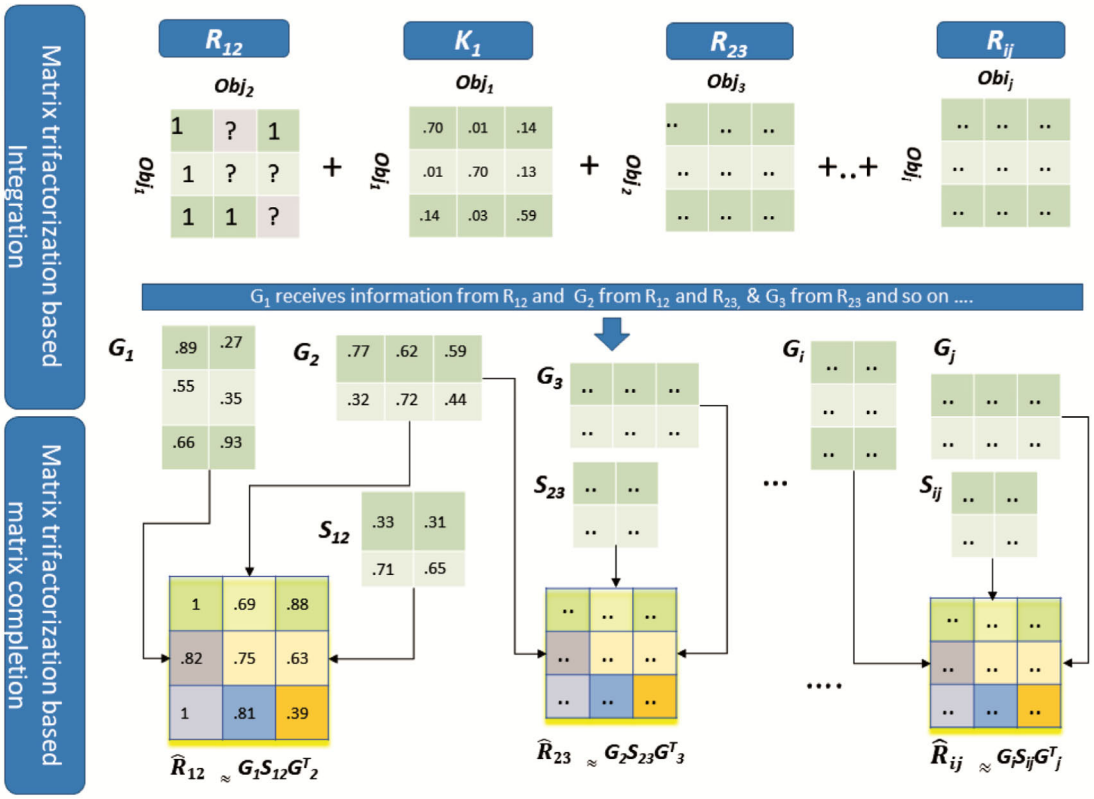

FIGURE 1. Data fusion, tri-factorization and matrix completion, where R_{ij} is a matrix representing relation between objects obj_i and obj_j, G_i, G_j and S_{ij} are low-rank factor matrices, and $\hat{R}_{12}$ is the finally reconstructed of matrix (i.e., matrix completion).

Similarly, applying NMTF to simultaneously decompose n relation matrices R_{ij} into $G_i \in \mathbb{R}^{n_i \times k_i}$, $G_j \in \mathbb{R}^{n_j \times k_j}$ and $S_{ij} \in \mathbb{R}^{k_i \times k_j}$, the resultant objective function is the summation of objective functions for individual relations. The factor matrices G_i, G_j and S_{ij} with relatively low dimensions ($i.e., k_i \ll n_i, k_j \ll n_j$) can thus be approximated by solving the following optimization problem:

$$\min_{G \geq 0, S \geq 0} J_3 = \sum_{R_{ij} \in \mathbb{R}} \left\| R_{ij} - G_i S_{ij} G_j^T \right\|_2 \quad (3)$$

The non-negative constraints imposed on factor matrices G and S provides easy interpretation of their values in cluster assignment and allows introduction of additional data sources in the form of kernel matrices for regularization. The goal is to make sure that an interacting pair of genes belong to a common group. Any violation of these constraint penalize our objective function. By adding the regularization terms, our final objective function becomes:

$$\min_{G \geq 0, S \geq 0} J_3 = \sum_{R_{ij} \in \mathbb{R}} \left\| R_{ij} - G_i S_{ij} G_j^T \right\|_2 + \sum_{i=1}^{r} tr(G_i^T K_i G_i) \quad (4)$$

where $||.||$ denotes the frobenius norm and $tr(.)$ is the trace of matrix. The key objective in data fusion using NMTF is to share the low-rank matrix factors G_i, $1 \leq i \leq N$ simultaneously among all relation matrices R_{ij} during the factorization process to ensure the contribution of fused datasets in the model. For example, matrix G_2 contributes its share during the simultaneous decomposition of R_{2j} and $R_{i2} \forall 1 \leq i, j \leq N$. Therefore, we can say that the reconstruction of matrix $\hat{R}_{12}$ is influenced by all the datasets related by relation matrices R_{ij}, $i \neq 1$ and $j \neq 1$, whose low-rank factors include either G_1 or G_2. The fusion approach adopted here is in agreement with the principles of intermediate integration, where the basic structure of data remains unchanged during model inference. This approach has been reported to show high predictive accuracy compared to early and late integration approaches [33].

The objective function minimization depends on the lowrank matrices derived using multiplicative update rules. The algorithm starts by initializing matrix factors G, and S using a Random $vcol$ initialization strategy. The factors G_i and S_{ij} are then iteratively updated till the convergence criteria is not met. The convergence criteria in our case is $||R_{ij} - G_i S_{ij} \ G_j^T \ ||_2 < \epsilon$. The term ϵ is a user definedparameter set to 10^{-5}. The details of derivation and p roof of convergence of multiplicative update rules adopted in this study are covered in [27].

A. PARAMETER ESTIMATION

Because the approach is based on matrix factorization, the parameters that need to be estimated are factorization

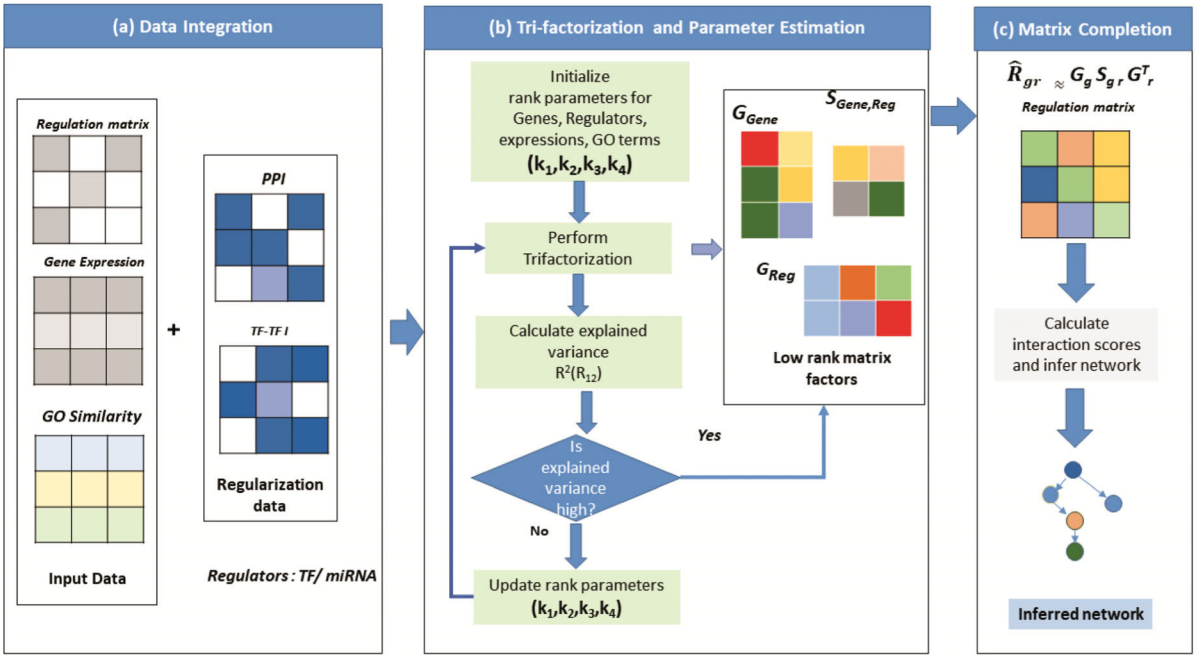

FIGURE 2. Overview of iMTF-GRN method. iMTF-GRN comprises of three important stages: (a) Data integration, where relevant datasets that control gene regulation are fused together. (b) tri-factorization of the fused data and parameter estimation for stable factorization, and (c) reconstruction of regulation matrix from learned latent factors and subsequent network inference.

ranks $k_1,k_{2...}k_r$. The rank parameters are initialized with set of values before factorization is performed. Parameters that maximize the quality of the model are finally chosen for performing factorization. Using these rank parameters, we evaluate the model by calculating explained variance between input relation R_{12} and the estimated relation $\hat{R}_{12}$ as:

$$R^2\left(R_{ij}\right) = 1 - RSS(R_{ij}) / \sum \left[R_{ij}\right]^2 \tag{5}$$

where,

$$RSS\left(R_{ij}\right) = \sum \left[R_{ij} - \hat{R}_{ij}\right]^2 \tag{6}$$

We perform the cross-validation procedure to assess the quality metrics and track changes for different factorization ranks. Rank parameters k_1, $k_{2,...,}k_r$ are chosen when explained variance $R^2\ (R_{ij})$ is very high.

B. PREDICTION OF INTERACTING PAIRS

In order to identify new interacting pairs between genes and their regulators, we compute the mean association scores of all the known regulators of the given gene. Candidate pairs $(g,\ r*)$ whose estimated association score from $\hat{R}_{12}$ is above the mean association for all known regulators of the given gene g are the new predicted pairs:

$$\hat{R}_{12}\left(g, r^*\right) > \frac{1}{|S(g)|} \sum_{r \in S(g)} \hat{R}_{12}\left(g, r\right) \tag{7}$$

where $S\ (g)$ is the set of known regulators of g. The above equation uses a row centric rule to identify transcription factors which might regulate the given gene. In case our input does not contain any known regulation for the given gene, we can identify new gene-regulator pairs by applying column centric rules. An overview of the proposed model from data integration to network inference is depicted in Figure 2.

C. DATASETS

1) E. COLI DATA

We downloaded a compendium of gene expression profiles compiled by Fait et al. (2007) containing 445 conditions for 4345 genes publicly available at Many Microbe Microarrays Database (M3D) website *(http://m3d.mssm.edu/)*. These expression profiles have been collected under different conditions including heat shock, pH changes, antibiotics, genetic perturbations and varying oxygen concentrations. Besides, the expression data, the *e.coli* benchmark dataset contains regulatory relationships between 1211 genes and 154 TFs, validated from RegulonDB [34]. For regularization, proteinprotein interaction data of *E.coli* was downloaded from BioGRID (version 3.4) [35] and semantic similarity of GO MF annotations for 4345 genes was calculated using the GOSemSim [36] package in R.

2) TCGA GLIOBLASTOMA DATA

For eukaryotic GRN inference, we downloaded multi-omics data for glioblastoma multiforme (GBM) from *(https:/gdac.broadinstitute.org/)*, a public data sharing portal from the Broad Institute which hosts The Cancer Genome Atlas (TCGA) data [37]. Datasets such as mRNA expression, miRNA expression, DNA methylation that influence the transcriptional and post-transcriptional regulation were selected

for data fusion. A partially observed miRNA-gene relational network is constructed by selecting genes from Microcosm [38] and mirTarBase [39] databases. For regularization, we downloaded a gene-gene interaction network from BioGRID (version 3.4).

D. GRN INFERENCE IN E.COLI

For GRN inference in *e.coli* the relationship between n_1 Genes and n_2 TFs is represented using a relational matrix $R_{12}^{n_1 \times n_2}$. The elements of this high dimensional matrix are binary values $R_{12}[g][t] = 1$, if a transcription factor t regulates a gene g, and 0 otherwise. R_{12} is a sparse matrix as it contains only partially observed gene-TF interactions downloaded from *RegulonDB*. The other data sources considered for the inference task are gene expression profiles connecting genes and experimental conditions R_{13} the elements of R_{13} are real-valued, representing the expression of genes across different experimental conditions. Besides these two datasets, biological knowledge in the form of protein-protein interactions and gene ontology (GO) semantic similarity between genes is also incorporated to serve as constraint matrices for regularization. The target matrix $\hat{R}_{12}$, reconstructed from the low-rank factor matrices $\hat{R}_{12} \approx G_1 S_{12} G_2^T$ is more complete than the original R_{12}. This relational matrix can now be used to extract new links between unobserved gene-TF pairs based on the interaction scores generated by the matrix completion approach.

E. INFERENCE OF MIRNA-TARGET GENE NETWORK

To infer a post-transcriptional gene regulatory network between miRNAs and their targets, we construct relational matrices from multi-omics datasets, such as miRNA expressions, mRNA expression, and methylation expression data. In order to predict new miRNA-gene relations, we build R_{12}, a partially observed binary matrix of regulatory relations between miRNA and their target genes from Microcosm and mirTarBase databases. A regulatory relationship within $R_{12}[m][g]$ is set to 1 if miRNA regulates gene g and 0 otherwise. Other Omics data sources that serve as support data and provide complementary information for gene regulation are integrated into the fusion frameworks as R_{13} (miRNA expression in tissues samples), R_{23}, (genes expression and tissue samples), R_{24} (genes and methylation expression data). Besides, we also supply a similarity matrix in the form of a gene-gene interaction diffusion kernel as K_2 relation to serve as a constraint matrix for regularization.

III. RESULTS

We illustrate the application of iMTF-GRN to infer TFtarget/miRNA-target gene relations from a benchmark *E. coli* dataset and multi-omics TCGA glioblastoma data. The proposed method is implemented using scikit-fusion package [27] and the implementation is available at github *(https://github.com/waninisar/iMTF-GRN)*. We evaluate the effectiveness of our inference approach first by comparing it to methods using standalone gene expression data for the inference tasks as well as methods that integrate multiple omics datasets for improved inference of TF-gene target relations. We also fuse multiple omics data for inference of a post-transcriptional regulatory network between miRNAs and their targets from a TCGA glioblastoma dataset.

A. E. COLI REGULATORY NETWORK

After applying iMTF-GRN to benchmark *e.coli* data, we performed a 5-fold cross-validation to make sure that model is trained on all the known regulations. At each crossvalidation step, important performance metrics such as precision, recall, and F1 score are recorded and then averaged for the whole cross-validation procedure. We then compared our network inference method with inference methods that predict TF-target gene regulations, such as CLR [5] and SIRENE [10] because both of these methods have been developed using the same benchmark data.

We also compared our method with other state-of-theart data integration and inference methods, such as multiple kernel learning (MKL) and integrated random forest (iRafNet) on the benchmark dataset. For MKL we generated an RBF kernel from R_{13} (gene expression for 445 experimental conditions), a linear kernel from R_{14} (Go semantic similarity) and a diffusion kernel from proteinprotein interactions. The parameters for these kernels are estimated by performing an internal 5-fold cross-validation. The prediction task is defined as a binary classification learning whereby the model outputs the probability scores of TF-target gene associations. Scores are selected based on different thresholds at various precision levels by calculating the precision and recall score between test set labels and classifier outputs.

For iRafNet we consider the gene expression compendium as the main dataset and the protein-protein interactions (PPI) between *E.coli* genes as the supporting data. Sampling weights are then derived from PPI data by building a diffusion kernel as $K = e^H$ where H is a graph laplacian for PPI data. Sampling weights from K are derived as $W^{PPI}{}_{i,j} = K(i,j)$ i, e. the element $K(i, j)$. The sampling weights thus obtained are then integrated with main dataset (i.e., gene expression data). Putative regulatory links are then predicted using importance scores generated using the iRafNet R package [40]. After the execution of iMTF-GRN inferencing procedure, it was observed that the known regulations at 60% and 80% precision levels that are correctly predicted by iMTF-GRN is higher than those predicted by other comparable methods (Table 1). The recall and F1 scores obtained for all the methods at different precision levels is summarized in Table 1. The improvement with iMTF-GRN can be attributed to two factors, first, the additional biological knowledge in the form of proteinprotein interactions and GO similarity scores between *E.coli* genes to aid in the inference process and the second factor is the ability of the Non-negative matrix factorization based methods to detect context-dependent hidden patterns of gene expression and little sensitivity to initial conditions and a priori selection of genes.

TABLE 1. Recall rate of TF-gene target prediction algorithms at 60% and 80% precision levels. The values for CLR and SIRENE were directly taken from [10].

METHOD	RECALL AT 60%	RECALL AT 80%	F1 SCORE AT 60% PRECISION
CLR	0.075	0.055	0.13
SIRENE	0.445	0.176	0.51
MKL	0.65	0.44	0.62
iRafNet	0.72	0.58	0.65
MTF-GRN	**0.820**	**0.68**	**0.69**

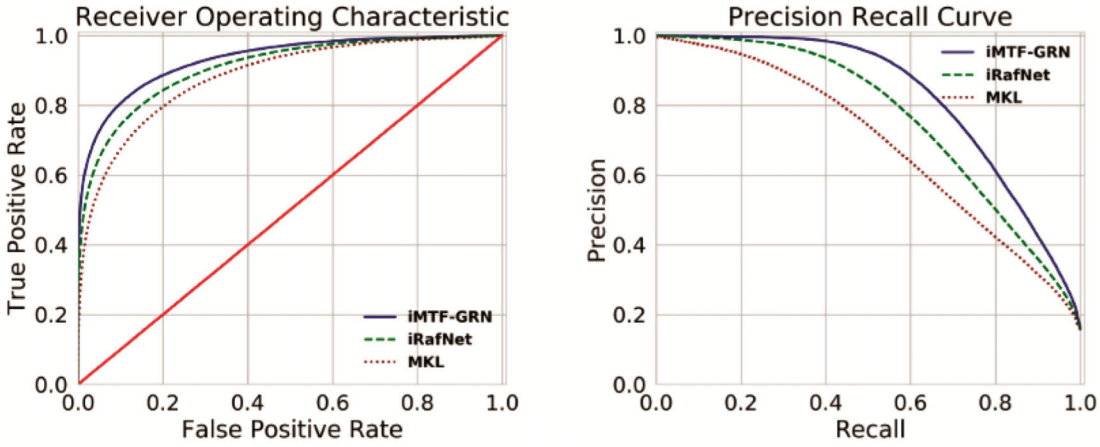

FIGURE 3. ROC and precision recall curves for iMTF-GRN, MKL and iRafNet.

The robustness with which the matrix factorization captures the biological correlations in gene expression data by describing tens of thousands of genes using a far less number of metagenes provides for a general method for pattern and class discovery from biological datasets [41]. The ROC and precision-recall curves for iRafNet, MKL and iMTF-GRN are plotted in Figure 3.

To evaluate the network inference potential of the iMTTFGRN on *E. Coli* benchmark, we perform the prediction procedure on the whole *E. coli* network at 60% precision level. For each of the 154 TFs in R_{12}, we select all the gene-TF pairs with a score above an estimated threshold calibrated from the cross-validation procedures. We complete the partially observed R_{12} by constructing $\hat{R}_{12}$ from the latent factors G_1, S_{12} and G_2^T and search for the gene-TF pairs with an interaction score above an estimated threshold (>0.50). In addition to the 3293 known regulations in our data, we predict 1266 new regulations (attached as supplement), out of which subset interactions are listed in Table 2. These interactions were not part of our input data, and their validity has been ascertained by searching for relevant literature and databases.

A graphical depiction of a subnetwork comprising more than 200 interactions both known and predicted is captured in Figure 4. The reason for choosing a small subset from the entire network stems from the fact that, almost all the predicted regulations from this set were thoroughly investigated from literature and many *E.coli* databases (e.g., *RegulonDB*, TEC, etc.). Also, the F1 scores obtained for the different combination of data using iMTF-GRN across multiple cross-validation runs is plotted in Figure 5, the inclusion of additional data consistently improves the F1 scores, thereby improving the model performance.

B. MIRNA-GENE POST-TRANSCRIPTIONAL REGULATORY NETWORK

To demonstrate the network inference potential of iMTFGRN on eukaryotic genomes, we apply the process to a TCGA multi-omics dataset of glioblastoma in order to predict unobserved miRNA-gene interactions. We reconstruct our target approximation $\hat{R}_{12}$ from its low rank factors G_1, S_{12} and G_2^T. We identify new miRNA-gene target pairs $\hat{R}_{12}[r, g^*]$ by using equation (7).

A list of 20 validated miRNA-gene targets along with their association scores is listed in Table. 3. For validation of predictions, we searched the Human miRNA Disease Database (HMDD) [42] for miRNAs that are associated

TABLE 2. List of TFs and their target genes predicted by iMTF-GRN not present in R_{12}.

TRANSCRIPTIONAL FACTOR	TARGET GENES	REFERENCE PMID
ArcA	aceA, acnA, acnB, fadA, fadB	9421904,26843427
ArgR	argB, argC, hisj, hisM, hisP, hisQ, nusA	26843427,9421904
CpxR	cpxA, csgD, csgE, recJ, potl	25735747,26527724
Crp	dsdA, envZ, fucA, sodB, yeiP	26843427,15743952,26527724
Fur	cyoA, cyoB, exbB, exbD,fecI, fecR, zwf	9485415,26527724, 12644513
H-NS	aceA, aceZB, osmY, serA, serC	26527724,21673794
Lrp	cadB, csiE,dadX, dppA, gabD, osmC, osmY	15340867,25735747,19052235
MarA	acnA, rimK, ybjC, ybjN	11395452,24860636,26843427
Rob	acrA, acrB, fumC, rimK, ybjC, ybjN	10850996,24860636,11395452
RpoH	dapA, fabZ,, rfbB, rpoE	16818608,26527724,8244018
SoxS	aceB, acrB, oxyR, sodA, ybaO	11395452,24860636,26843427

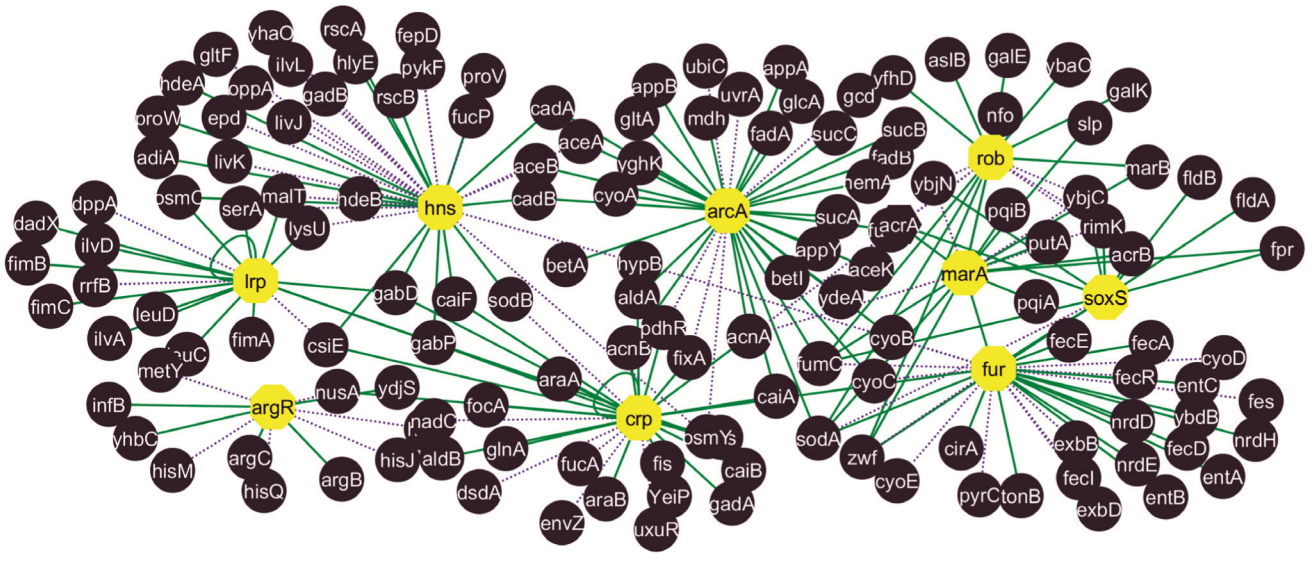

FIGURE 4. A snapshot of predicted network between TFs (yellow) and their target genes. Green lines indicate known regulations and blue dashed lines show predicted interactions.

with GBM and compared these with the newly identified miRNA-gene interaction. Additionally we performed a 5-fold cross validation procedure and generated ROC and precision-recall scores averaged across all the cross validation runs. The proportion of known miRNAgene pairs from the input relation R_{12} correctly predicted by the algorithm are true positives, on the other hand false positives are the predicted miRNA-gene pairs not present in the input relation R_{12}. We generate AUROC (average Reciever operating characteristics) and AUPR (average Precision-Recall) of the cross validation runs to evaluate the inference process for each combination of the omics datasets being integrated as shown in Table 4. We assess our prediction accuracy for miRNA-gene pairs against HMDD and mir2Disease [43]

TABLE 3. List of predicted top scoring mirNA-gene interactions and their validation from HMDD, mir2Disease(M2D), mirTarBase databases and PubMed.

MiRNA	Target Gene	Interaction Score	Database	PMID
hsa-miR-137	EZH2	0.921310025	HMDD	25939439
hsa-mir-30a-5p	BDNF	0.896386098	HMDD	21178384
hsa-miR-299-5p	SOX4	0.847136949	mirTarBase	-
hsa-miR-326	EGFR	0.836344306	HMDD	23302469
hsa-miR-137	CDK6	0.831422545	M2D	18577219
hsa-miR-142-3p	KLF4	0.831239295	TarBase (V.8)	-
hsa-miR-133b	IDH1	0.814940133	TargetScan	28804724
hsa-mir-329	RNF165	0.813841425	HMDD	23302469
hsa-miR-128-1	SOX11	0.811149147	mirTarBase	-
hsa-miR-429	SOX2	0.802881482	HMDD	28749077
hsa-miR-206	CCND1	0.778097953	mirTarBase	-
hsa-miR-198	PROX1	0.736641973	TargetScan	28035380
hsa-miR-17-5p	PTEN	0.731914078	HMDD	21483847
has-mir-520b	CCND1	0.712909496	TargetScan	26700671
hsa-miR-1	Mef2a	0.710329621	M2D	18759060
hsa-miR-181d	MALT1	0.709433307	HMDD	28286260
hsa-miR-210	BDNF	0.599359912	HMDD	25586423
hsa-miR-504-5p	TMEM8A	0.586117283	TargetScan	-
hsa-miR-184	BCL2	0.584226655	HMDD	22844109
hsa-miR-142-5p	CCND1	0.573554602	TargetScan	-

TABLE 4. AUROC and AUPR for different combinations of omics data. MGI - miRNA-gene interaction, GE – gene expression, ME- miRNA expression, DNA methylation expression.

Dataset	AUROC	AUPR
MGI	0.8354	0.5538
MGI+ GE	0.8451	0.5749
MGI+GE+ME	0.8767	0.6561
MGI+GE+ME+DME	0.8985	0.6881

databases and also make searches of the experimental literature on the PubMed central. A list of 108 top scoring miRNA-gene pairs (score $>= 0.50$) validated from multiple databases that store miRNA-target gene information is provided as supplementary data. A subset of these validated interactions is shown in Table 3 along with the PMID of the reference literature. The biological relevance of these miRNA-gene interaction pairs to Glioblastoma (GBM) is ascertained by analyzing the pertinent scientific literature.

For example, miR-206,miR-142-5p, and miR-520b predicted to interact with CCND1 (Cyclin D1) have been reported to activelyfunctionastumorsuppressors by targeting Cyclin-dependent kinases(CDKs) via CyclinD1.

CDKs are an essential family of multifunctional enzymes that can modify various proteins substrates involved in cell cycle progression. CCND1 forms a complex with CDK4/CDK6, a regulatory subunit required for cell cycle G1/S1 transition, any overexpression of this gene alters cell cycle progress and may contribute to tumorigenesis. Using RT-PCR and western blot analysis, authors in [44] report overexpression of CCND1 in U251, U87 cell lines and GBM tissues. They also show that CCND1 is negatively correlated with miR-520b expression. To demonstrate the role of miR-520b, they transfected U87, U251 cell line with miR-520b mimics, overexpression of miR-520b lead to significant decrease in cell proliferation and colony formation in these cells, suggesting miR-520b induced growth inhibition and apoptosis promotion in GBM cells. Since both miR-206 and miR-142-5p also interact with CCND1 as shown in Table 3, a quick survey of the relevant literature confirms the presence of binding sites in the 3' UTR of CCND1 for miR-206 as reported in [45]. Also, CCND1-miR-142-5p interaction has been verified from TargetScan [46]; therefore, we can assume a tumor suppressor role similar to miR-520b and miR-206 when this miRNA is upregulated.

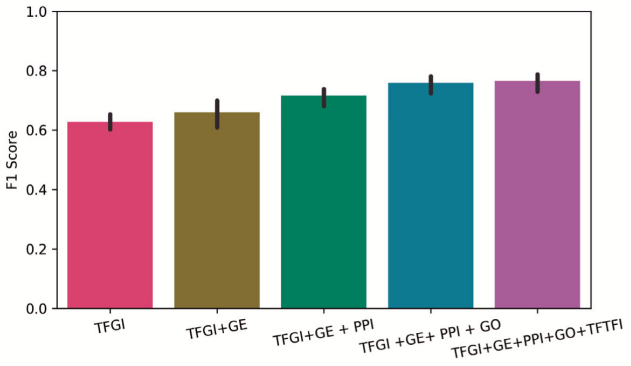

FIGURE 5. F1 scores averaged across different cross validation runs for different combinations of data. TFGI- TF-gene interaction, GE - gene expression, PPI- protein-protein interaction, GO - Gene Ontology semantic similarity, TFTFI- TF-TF interaction.

Another predicted interaction not present in either HMDD or mir2Disease is the miR-299-5p, SOX4 interacting pair. After searching the relevant literature, we could not find a direct study explaining the regulatory role of this pair. However, it has been reported in [47] that SOX4 inhibits the cell growth in GBM and induces *G0/G1* cell cycle arrest through the p53-p21 signaling pathway. In this study, SOX4 overexpression has been reported to inhibit the growth in LN229, A172G, and U87 GBM cell lines. On the other hand, miR-299-5p has been reported to be overexpressed in various GBM cell lines (e.g., A172, T98G) and its knockdown significantly inhibiting cell proliferation and promoting apoptosis. Since, SOX4 acts as a tumor suppressor, overexpression of miR-299-5p in GBM cells will silence SOX4 expression, disrupting its tumor-suppressing ability, thereby promoting tumorigenesis by acting as an oncomiR. On the other hand, miR-504 and miR-128-1 are downregulated in GBM cell lines as reported in [48] and [49] and also have similar expression profiles as explained in [50], up-regulation of these microRNAs in human GBM cells suggests a tumor suppressor role for both of them. Similarly, for the miR-142-3p, KLF4 (Krüppel-Like Factor 4) interacting pair, an upregulated miR-142-3p will suppress the expression of KLF4 whose overexpression in GBM cells has been reported in [51].

IV. CONCLUSION AND DISCUSSIONS

iMTF-GRN is a matrix completion based approach to infer gene regulatory networks from a combination of genomic and other biological datasets. When a network inferencing task is formulated as a classification problem, the choice of selecting negative examples for training becomes a non-trivial task. Matrix completion offers a more straightforward approach to complete a partially observed gene-TF interaction matrix by reconstructing the target matrix from its latent factors, thereby obtaining essential threshold scores for suggesting new regulatory relationships.

This paper demonstrates the application of iMTF-GRN on both prokaryotic and eukaryotic genomes. For prokaryotes, we inferred a transcriptional regulatory network from a compendium of gene expression data of E.coli, a benchmark data used by Martin *et al.* [7] and Mordelet and Vert [10] for CLR and SIRENE algorithms. We, fuse additional biological datasets that provide complementary information for gene regulation such as PPI, GO similarities and TF-TF interaction and Gene-TF interaction data and evaluate the performance of the method with CLR, SIRENE, MKL and iRafNet using precision, recall, and F1 score metrics. We validated a set of 50 TF-gene interactions fromTranscription factor profiling of *E.coli* (TEC) [52] and latest *RegulonDB* databases that were predicted by iMTF-GRN method at 60% precision level.

To evaluate the effectiveness of the method on the eukaryotic genome, we fused multiple TCGA omics datasets of glioblastoma downloaded from TCGA Broad Institute data portal for miRNA-gene post-transcriptional regulatory network inference [37]. We rank the miRNA-gene pairs using interaction scores and assess the biological relevance of the predicted interactions from HMDD and mir2Disease databases. Besides, we also identify the miRNA-gene pairs whose role in glioblastoma progression and suppression has been explained through careful scanning of the relevant scientific literature. Their direct role as tumor suppressors or the promoters of tumorigenesis in glioblastoma awaits further experimental validation.

Although iMTF-GRNF presents an efficient computational framework for data fusion and offers much potential in identifying and understanding essential patterns central to mechanisms of gene regulation. However, there are certain limitations as well. For example, iMTF-GRN needs some prior known TF-target/miRNA-target gene relations and is biased towards interaction pairs where the known regulations are higher. This bias makes it unable to predict new targets where no known regulations are present in the input data. Despite the process of data fusion improving the performance of the approach significantly, iMTF-GRN does not implement a weighing mechanism for the integrated datasets in order to measure the contribution of each dataset for robust selection of relevant data.

REFERENCES

[1] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.
[2] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. suppl_2, pp. S231–S240, 2002.
[3] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," in *Biocomputing*. Singapore: World Scientific, 1999, pp. 418–429.
[4] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinf.*, vol. 7, no. 1, p. S7, 2006.
[5] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, no. 1, p. e8, 2007.
[6] T. Akutsu, S. Miyano, and S. Kuhara, "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function," *J. Comput. Biol.*, vol. 7, nos. 3–4, pp. 331–343, 2000.

[7] S. Martin, Z. Zhang, A. Martino, and J.-L. Faulon, "Boolean dynamics of genetic regulatory networks inferred from microarray time series data," *Bioinformatics*, vol. 23, no. 7, pp. 866–874, 2007.

[8] K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak, and J. J. Tyson, "Integrative analysis of cell cycle control in budding yeast," *Mol. Biol. Cell*, vol. 15, no. 8, pp. 3841–3862, 2004.

[9] A. Climescu-Haulica and M. D. Quirk, "A stochastic differential equation model for transcriptional regulatory networks," *BMC Bioinf.*, vol. 8, no. 5, p. S4, 2007.

[10] F. Mordelet and J.-P. Vert, "SIRENE: Supervised inference of regulatory networks," *Bioinformatics*, vol. 24, no. 16, pp. i76–i82, 2008.

[11] Z. Gillani, M. S. H. Akash, M. M. Rahaman, and M. Chen, "CompareSVM: Supervised, support vector machine (SVM) inference of gene regularity networks," *BMC Bioinf.*, vol. 15, no. 1, p. 395, 2014.

[12] Y. Ni, D. Aghamirzaie, H. Elmarakeby, E. Collakova, S. Li, R. Grene, and L. S. Heath, "A machine learning approach to predict gene regulatory networks in seed development in arabidopsis," *Frontiers Plant Sci.*, vol. 7, p. 1936, Dec. 2016.

[13] N. Noman, L. Palafox, and H. Iba, "Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model," in *Natural Computing and Beyond*. Tokyo, Japan: Springer, 2013, pp. 93–103.

[14] K. Raza and M. Alam, "Recurrent neural network based hybrid model for reconstructing gene regulatory network," *Comput. Biol. Chem.*, vol. 64, pp. 322–334, Oct. 2016.

[15] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comput. Biol.*, vol. 7, nos. 3–4, pp. 601–620, 2000.

[16] A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, 2010, Art. no. e12776.

[17] E. P. van Someren, B. L. Vaes, W. T. Steegenga, A. M. Sijbers, K. J. Dechering, and M. J. Reinders, "Least absolute regression network analysis of the murine osteoblast differentiation network," *Bioinformatics*, vol. 22, no. 4, pp. 477–484, 2005.

[18] M. Gustafsson, M. Hörnquist, J. Lundström, J. Björkegren, and J. Tegnér, "Reverse engineering of gene networks with LASSO and nonlinear basis functions," *Ann. New York Acad. Sci.*, vol. 1158, no. 1, pp. 265–275, 2009.

[19] 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, no. 3, p. 333, 2014.

[20] X. Zhou, X. Wang, R. Pal, I. Ivanov, M. Bittner, and E. R. Dougherty, "A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks," *Bioinformatics*, vol. 20, no. 17, pp. 2918–2927, 2004.

[21] T. Santra, W. Kolch, and N. B. Kholodenko, "Integrating Bayesian variable selection with modular response analysis to infer biochemical network topology," *BMC Syst. Biol.*, vol. 7, no. 1, p. 57, 2013.

[22] M. Banf and S. Y. Rhee, "Enhancing gene regulatory network inference through data integration with Markov random fields," *Sci. Rep.*, vol. 7, Feb. 2017, Art. no. 41174.

[23] N. Omranian, J. M. O. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused LASSO on multiple data sets," *Sci. Rep.*, vol. 6, Feb. 2016, Art. no. 20533.

[24] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," in *Biocomputing*. Singapore: World Scientific, 2003, pp. 300–311.

[25] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein–protein interactions," *Bioinformatics*, vol. 21, no. suppl_1, pp. i38–i46, 2005.

[26] H. Wang, H. Huang, C. Ding, and F. Nie, "Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization," *J. Comput. Biol.*, vol. 20, no. 4, pp. 344–358, 2013.

[27] M. Žitnik and B. Zupan, "Data fusion by matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 41–53, Jan. 2015.

[28] N. Wani and K. Raza, "Integrative approaches to reconstruct regulatory networks from multi-omics data: A review of state-of-the-art methods," *Preprints*, 2018. doi: 10.20944/preprints201804.0352.v1.

[29] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. SIAM Int. Conf. Data Mining*, 2005, pp. 606–610.

[30] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 1–12.

[31] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, "Patient-specific data fusion for cancer stratification and personalised treatment," in *Biocomputing*. Singapore: World Scientific, 2016, pp. 321–332.

[32] M. Žitnik and B. Zupan, "Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold," in *Biocomputing*. Singapore: World Scientific, 2014, pp. 400–411.

[33] P. Pavlidis, J. Weston, J. Cai, and W. S. Noble, "Learning gene functional classifications from multiple data types," *J. Comput. Biol.*, vol. 9, no. 2, pp. 401–411, 2002.

[34] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sánchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jiménez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, and A. Martinez-Antonio, "RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions," *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D394–D397, 2006.

[35] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D535–D539, 2006.

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

[37] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[38] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[39] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W. T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang, "miRTarBase: A database curates experimentally validated microRNA–target interactions," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D163–D169, 2010.

[40] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, 2015.

[41] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–4169, 2004.

[42] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "HMDD v2.0: A database for experimentally supported human microrna and disease associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1070–D1074, 2013.

[43] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, no. suppl_1, pp. D98–D104, 200.

NISAR WANI is currently an Assistant Professor with the Department of Higher Education, Govt. Degree College Baramulla, University of Kashmir, Srinagar, India. He is also registered as a Research Scholar with the Department of Computer Science, Jamia Millia Islamia, New Delhi, India. His research interests include biological network analysis, systems biology, and computational biology.

KHALID RAZA received the Ph.D. degree in computational biology and soft computing from Jamia Millia Islamia, New Delhi, India, where he is currently an Assistant Professor with the Department of Computer Science. He has contributed over 40 research articles in refereed international journals, conference proceedings, and as book chapters. His research interests include systems biology, soft computing techniques, microarray, and NGS analysis. He is also an Active Reviewer for leading Bioinformatics Journals. He is also a supervising/co-supervising six doctoral students.