CF-GeNe: Fuzzy Framework for Robust Gene Regulatory Network Inference

Muhammad Shoaib B. Sehgal
Faculty of IT, Monash University, Victoria 3842, Australia.
Email: Shoaib.Sehgal@infotech.monash.edu.au

Iqbal Gondal and Laurence S. Dooley
Faculty of IT, Monash University, Victoria 3842, Australia.
Email: {Iqbal.Gondal@infotech.monash.edu.au, lsdaussie@ieee.org}

***Abstract*— Most *Gene Regulatory Network* (GRN) studies ignore the impact of the noisy nature of gene expression data despite its significant influence upon inferred results. This paper presents an innovative *Collateral-Fuzzy Gene Regulatory Network Reconstruction* (CF-GeNe) framework for *Gene Regulatory Network* (GRN) inference. The approach uses the *Collateral Missing Value Estimation* (CMVE) algorithm as its core to estimate missing values in microarray gene expression data. CF-GeNe also mimics the inherent fuzzy nature of gene co-regulation by applying fuzzy clustering principles using the well-established *fuzzy c-means* algorithm, with the model adapting to the data distribution by automatically determining key parameters, like the number of clusters. Empirical results confirm that the CMVE-based CF-GeNe paradigm infers the majority of co-regulated links even in the presence of large numbers of missing values, compared to other data imputation methods including: *Least Square Impute* (LSImpute), *K-Nearest Neighbour Impute* (KNN), *Bayesian Principal Component Analysis Impute* (BPCA) and *ZeroImpute*. The statistical significance of this improved performance has been underscored by *gene selection* and also by applying the *Wilcoxon Ranksum Significance Test*, with results corroborating the ability of CF-GeNe to successfully infer GRN interactions in noisy gene expression data.**

***Index Terms*—Gene Regulatory Networks, Significant Gene Selection, Missing Value Imputation, Collateral Missing Values Estimation (CMVE), Fuzzy Clustering.**

***Software Availability*— The *CF-GeNe* and *CMVE* softwares can be requested by emailing at {Shoaib.Sehgal@gmail.com} or can be downloaded from http://www.gscit.monash.edu.au/~shoaib/.**

I. INTRODUCTION

Gene expression analysis has been widely used for different biological studies. Several statistical and computational intelligence methods have been introduced for cancer class prediction [1-3], differentially expressed gene selection [4, 5] and co-regulated genes clustering under variety of conditions. While these techniques afford biologists a valuable insight into different biological systems, there is still a need for suitable methodologies that will increase the robustness of complex genetic interaction inference that consider many thousands of genes at a time, in order to better understand complex genetic networks.

Despite its wide application, gene expression data frequently contain erroneous values, which are popularly termed as *missing values,* with up to 90% of genes affected [6]. The reasons for missing values are many and varied, ranging from spotting problems, slide scratches, chip blemishes, hybridization errors and image corruption through to simply dust on the slide [7]. Whatever the cause, missing values have the potential to create a significant impact upon subsequent inferences made from the microarray data, in a diversity of applications from gene regulatory network modeling and gene selection to class prediction. To date, most analysis has been based on over/under expressed genes studies and clustering, notwithstanding the fact that differential expression analysis does not fully harness the potential of microarray gene expression data, because genes are independently treated, so any interaction between them is not considered [8]. *Gene Regulatory Networks* (GRN) are able to model how genes regulate different metabolisms that have been shown to lead to diseases, like cancer [9].

Due to the sensitivity of GRN analysis to noisy microarray data, GRN modeling often ignores some locally regulated genetic links resulting in important co-regulation information often being overlooked [10]. Unfortunately no formal analysis has yet been performed to analyze the performance of GRN construction in the presence of these missing values, with the normal strategy adopted being to either simply remove the erroneous data or replaced it with a zero (*ZeroImpute*).

Gene network construction however, is a difficult task due to both the curse of dimensionality i.e., the number of features always being much greater than the number of samples) and multi-colinearity [11]. Additionally, most clustering techniques that are used to cluster co-regulated genes are crisp, despite the inherently fuzzy nature of gene co-regulation where one gene can belong to more than one functional group. These techniques also use

either a fixed number of clusters or determine the number by visual inspection of data which runs contradictory to general clustering principles [12]. There is thus a necessity for a system which can model the inherently fuzzy nature of gene network, while adapting to the underlying data distribution, as well as both managing the noisy nature of the data and minimizing computational complexity.

This paper presents a novel *Collateral Fuzzy Gene Regulatory Network Inference* Framework (*CF-GeNe*) to model different GRNs. The proposed model addresses the issue of erroneous expression values by marking them as missing, and then applying the *Collateral Missing Value Estimation* (CMVE) [13] algorithm to impute these values. The model also imitates the fuzzy behaviour of biological systems by using fuzzy logic. The CF-GeNe uses *fuzzy c-means* (FCM) clustering to handle co-regulated genes because this allows genes to belong to multiple pathways, so mimicking biological cellular metabolism [14]. The model searches for co-regulated genes within each luster by computing the *spearman ranked* correlation between clustered genes which makes it much more computationally efficient compared to existing GRN modeling techniques [8, 15] by removing redundant calculations [14]. The CF-GeNe model also adapts to the data distribution by automatically determining the cluster number by using the fuzzy- PBM index which seeks to increase the separation between the cluster centers while concomitantly encouraging the formation of a greater number of clusters [16].

The CF-GeNe model has been rigorously tested for its application to find tumor-specific links in three different breast cancer mutation datasets, for different ranges of randomly introduced missing values. The results corroborated that CF-GeNe can successfully detect locally co-regulated genes across the full range of missing values compared to *ZeroImpute* and other well established imputation algorithms including: *Least Square Impute* (LSImpute), K-*Nearest Neighbour Impute* (KNN) and *Bayesian Principal Component Analysis* (BPCA). The performance of the framework with CMVE as its core estimator, also correctly detected differentially expressed genes in contrast to all the aforementioned imputation strategies. Significant gene selection is especially important in this context because it affords the possibility to analyze gene co-regulations in differentially expressed genes, when in many circumstances complete GRN analysis would be inadmissible due to the high number genetic interactions. The proposed strategy to handle noisy data was also examined with respect to the *Wilcoxon Ranksum Significance* test, with results again revealing that the CF-GeNe model outperformed all the other imputation approaches.

The rest of the paper is organized as follows: Section 2 presents in detail, the new CF-GeNe model. Different imputation strategies and gene selection techniques are then described in Section 3, with a results analysis being provided in Section 4. Finally, some general conclusions are made in Section 5.

II. Collateral-Fuzzy Gene Regulatory Network (CF-GeNe) Framework

The complete CF-GeNe framework is defined in Figure 1. The gene expression data is firstly preprocessed to remove noise and outliers before CMVE-based [13] imputation is applied. The appropriate number of clusters C_{opt} is then computed using fuzzy PBM-index prior to FCM clustering and the gene networks are then finally constructed for different genetic datasets using *spearman ranked* correlation.

CF-GeNe Framework

Pre Condition: Gene expression matrix *Y*

1) Pre-Processing (See Section A).
2) Estimate using CMVE (See Section B).
3) Determine number of clusters C_{opt} using Fuzzy PBM index (See Section C).
4) Cluster using fuzzy c-means (See Section D).
5) Select significant genes ϑ (See Section E)
6) for J $\leftarrow$1:C_{opt}
   Iteratively compute Spearman correlation coefficient ρ using (7) for each selected gene from ϑ with all the genes with in cluster *j*.
   end
7) Group genes with respect to their ρ with other genes to form a network where ρ represents the degree of co-regulation between genes.
8) Compare the network with standard network.
9) Compute conserved and false negative links
10) STOP

Post Condition: Gene Regulatory Networks *N* for normal and cancerous data

Figure 1: GRN Reconstruction Framework

CMVE Algorithm

Pre Condition: Gene expression matrix *Y* with *m* number of genes, *n* samples, *I* missing values, *index*=1

1) Compute absolute covariance *C* using (1).
2) Rank genes (rows) based on *C*.
3) Select the *k* most effective rows R_k
4) Use values of R_k to
   a. Estimate value Φ_1 using Least Square Regression.
   b. Compute Φ_2 and Φ_3 using (2) and (3).
5) Compute missing value of I[*index*] using (4) and reuse in future predictions.
6) Increment *index* and Repeat Steps 1–5 until all missing values of *Y* are estimated.

Post Condition: Y without any missing values.

Figure 2: CMVE Algorithm

The next series of subsections will explain each of these steps in detail, together with the rationale for the choice of each of the aforementioned techniques.

A. Pre-Processing

The data is preprocessed to handle noisy microarray data, with negative values in the gene expression data being marked as missing. The gene expression data is then normalized to $\sigma = 1$ and $\bar{Y} = 0$ before $\log_2$ transformation to minimize the affect of outliers (Step 1- Figure 1). Missing values in this pre-processed data are then imputed using the CMVE algorithm [12].

In describing the CMVE and other imputation techniques, the following nomenclature is adopted. Microarray gene expression matrix $Y \in \mathbb{R}^{m \times n}$ contains m genes and n samples. In Y, every gene i is represented by g_i, so Y in n experiments is arranged as $Y = [g_{1^T} \ldots g_{m^T}]^T \in \mathbb{R}^{m \times n}$. A missing value in gene i for sample j is expressed as $Y(i, j) = g_i(j) = \Xi$.

B. Imputation: Collateral Missing Value Estimation

This imputation strategy (CMVE Method, Figure 2) is based upon generating multiple parallel missing value estimations, which are subsequently combined to construct a final imputation value.

For example, if value $g_i(j)$, of gene i and sample j is missing, then firstly the diagonal covariance of i is computed together with the other gene expressions using (1), where m is the number of genes, i is the gene number with missing value for sample j and G_k is the gene in Y, other than i. Rows are then sorted according to their covariance, with the first k-ranked covariate genes R_k being selected.

$$C = \frac{1}{(n-1)} \sum_{k=1}^{m} (i_k - \bar{i})(G_k - \bar{G}) \tag{1}$$

Instead of applying a distance function, the covariance is used in CMVE because it considers both negative and positive correlation values, in contrast to the Euclidean distance used by KNN, which only considers positive correlations [5]. Another option would have been to use Pearson Correlation, though the net effect is exactly same for z-scored data [17], so the covariance is chosen due to its lower computational complexity. The missing values are then estimated by fusing together multiple estimates Φ_1, Φ_2 and Φ_3 (see (4)), with the various steps involved in the CMVE imputation algorithm now discussed.

Φ_1 is the estimate of $g_i(j)$ (Step 4a, Figure 2, CMVE Method) using the Least Square regression method, while Step 4b estimates two other sets of missing values Φ_2 and Φ_3. The former is estimated using:

$$\Phi_2 = \sum_{i=1}^{k} \phi + \eta - \sum_{i=1}^{k} \xi^2 \tag{2}$$

While the value of Φ_3 is computed using:

$$\Phi_3 = \frac{\sum_{i=1}^{k} (\phi^T \times I)}{k} + \eta \tag{3}$$

where η and ϕ in (2) and (3) are obtained using the *Non Negative Least Square* (NNLS) method [4]. Finally, value χ for $g_i(j)$ is computed (Step 5) using:

$$\chi = \alpha.\Phi_1 + \beta.\Phi_2 + \gamma.\Phi_3 \tag{4}$$

where α, β and γ are set to 0.33 to ensure an equal weighting to the respective estimates Φ_1, Φ_2 and Φ_3. The rationale for this choice is that as each estimate is highly data dependent, it avoids any bias being introduced towards one particular estimate.

Missing value imputation is then followed by computing the appropriate number of clusters C_{opt} (Step 3, Figure 1), with the next section explaining how this is achieved.

C. Computation of the Number of Clusters

To compute C_{opt}, a grid search is used to maximize the fuzzy PBM-index for a given expression data Y. For each individual cluster, the PBM index is calculated as follows:

$$PBM(C_{opt}) = \left(\frac{1}{C_{opt}} \times \frac{E_1}{E_{C_{opt}}} \times D_{C_{opt}} \right) \tag{5}$$

where

$$E_{C_{opt}} = \sum_{k=1}^{C_{opt}} \sum_{j=1}^{N} u_{kj} \| Y_j - C_k \|, \; D_{C_{opt}} = \max_{i,j=1}^{C_{opt}} \| C_i - C_j \|,$$

N is the total number of genes, u_{kj} is a partition matrix for cluster k and gene j and C_k is the cluster centre for the k^{th} cluster. The rationale for using the PBM-index, apart from its better performance over other validity measures, is that by maximizing the PBM-index distance it concomitantly reduces the distance between a cluster centre and its elements, which results in coarser clustering [16].

After selecting C_{opt} the data is clustered using FCM (See Step 4, Figure 1). There are many reasons for employing a clustering strategy prior to GRN modeling, including:

i. It removes redundant searches for co-regulated genes and makes GRN modeling computationally more efficient.
ii. Correlations in microarray data can occur serendipitously due to the *curse of dimensionality*, and lead to false GRN links [18].
iii. Genes tend to exhibit local rather than global relations between themselves, so fuzzy clustering can assist in identifying local patterns within the gene expression data [10].
iv. Genetic behavior is inherently fuzzy in nature because a gene can be associated with more than one functional group. FCM allows a gene to belong to more than one cluster with a certain degree of fuzziness [14].

D. Clustering: Fuzzy C Means

The objective function of FCM which has to be minimized is given by:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{Copt} u_{ij}^m \| Y_i - C_j \| \quad (6)$$

Where u_{ij} is a partitioning matrix and represents the degree of membership of gene Y_i in functional group j, C_j is the cluster centre of the functional group j and m determines the degree of fuzziness and is normally set to be ≥2 [19]. The partitioning matrix u_{ij} is iteratively computed from:

$$u_{ij} = \frac{1}{\sum_{k=1}^{Copt}\left[\frac{\| Y_i - C_j \|}{\| Y_i - C_k \|}\right]^{\frac{2}{m-1}}} \quad (7)$$

where the centroid C_j is given by :

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m . Y_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (8)$$

Complete GRN analysis is often infeasible due to the high number of genetic interactions, so a minimum of three times the total number of genes in a GRN is invoked, even when the Markov blanket property [20] is used to select the genetic links. One possible solution to this problem is to select a set of target genes and then analyze their co-regulations with all the other genes present in the dataset. To pick this set of target genes, CF-GeNe selected the significant genes using the *Between Group to within Group Sum of Squares* (BSS/WSS) method which is now explained.

E. Gene Selection: Between Group to within Group Sum of Squares

This gene selection method identifies those genes which concomitantly have large inter-class variations and small intra-class variations. For any gene i in $Y \in \mathbb{R}^{m \times n}$, BSS/WSS is calculated as follows:

$$BSS(i)/WSS(i) = \frac{\sum_{t=1}^{T}\sum_{q=1}^{Q} F(L_t = q)(\bar{Y}_{qi} - \bar{Y}_i)^2}{\sum_{t=1}^{T}\sum_{q=1}^{Q} F(L_t = q)(Y_{it} - \bar{Y}_{qi})^2} \quad (9)$$

where T is the training sample size, Q is the number of classes and $F(\bullet)$ is a Boolean function which is 1 if the condition is true, and zero otherwise. $\bar{Y}_i$ denotes the average expression level of gene i across all samples and $\bar{Y}_{qi}$ is the average expression level of gene i across all samples belonging to class q. Genes are then ranked accordingly to their BSS/WSS ratios, from the highest to the lowest to form a significant gene expression matrix ϑ, with the first p genes selected for subsequent GRN construction (Step 5, Figure 1). It should be noted however, that CF-GeNe is able to model the complete GRN even if gene selection is not performed which shows the scalability of the framework.

F. Gene Regulatory Network Construction

Spearman rank correlation is computed between each gene G_i in the selected genes set ϑ and all genes G_j, within the cluster (Figure 1, Step 6) using:

$$\rho = \frac{6\sum D_g^{\,2}}{N_g(N_g^{\,2} - 1)} \quad (10)$$

where D_g is the distance between the ordered gene pair G_I and G_J, while N_g is the number of pairs. There are two principal reasons for choosing this particular correlation metric [21]:

i. Correlation coefficients are highly sensitive to outliers and spearman rank correlation convert distances into a sequence of ranks, which is also useful in the handling of heterogeneous data.

ii. It correlates sequences that are either co-vary or contra-vary without imposing the restriction they should be linearly related.

III. Missing Value Estimation Algorithms

This Section outlines various missing value imputation algorithms used in this paper to both assess the effect of missing values on GRN inference and to provide comparative results to evaluate the performance of the CMVE technique, which is the preferred imputation strategy for the new CF-GeNe framework.

KNN [22] estimates missing values by searching for the k nearest genes using the Euclidean distance function and then taking the weighted average. The method however, does not consider negative correlations [6] and has the disadvantage of using a predetermined value of k regardless of the dataset being used. LSImpute which was developed by Bø et al, [23] estimates the missing values using LS regression, while BPCA [7] uses Bayesian Principal Component Analysis to impute the missing values, though this technique only exploits global correlations within the data structure, which can lead to erroneous estimates if data possesses a strong latent local correlation structure [7].

IV. Analysis of Results

To quantitatively evaluate the affect of missing values on GRN inference, the well accepted breast cancer microarray dataset by Hedenfalk et al, [24] was used in all the experiments. The dataset contains 7, 7 and 8 samples of BRCA1, BRCA2 and Sporadic mutations (neither BRCA1 nor BRCA2) respectively, with each data sample containing microarray data of 3226 genes. To evaluate the performance of different imputation methodologies on gene co-regulation inference, the first

step was to construct a GRN $\tilde{N}$ from the data which had no missing values. As the complete GRN is not feasible to analyze where minimum genetic links are three times the number of genes, even if the markov blanket property is considered; a set of genes Ъ was selected using the BSS/WSS from the original data with no missing values, to serve as the target gene set for analysis. Between 1% and 5% of known expression values were then randomly removed and subsequently imputed using *ZeroImpute*, BPCA, LSImpute, KNN and CMVE. A GRN was constructed for each imputed dataset and compared with the original network $\tilde{N}$ to determine both the *Conserved Link* (CL) and *False Negative* (FN) links. These are respectively defined as follows:

DEFINITION 1: *Conserved Link* is the co-regulation present in both the original network $\tilde{N}$ and the network constructed after imputation.

DEFINITION 2: *False Negative* is the regulatory link not present in $\tilde{N}$ and falsely detected in the network constructed after imputation.

Figures 3-7 graphically demonstrate the high number of CL achieved by CMVE compared to other imputation techniques for all the datasets, due to its better estimation ability for the significant genes for a statistical significant level of 0.05 [25]. For higher number of missing values the performance of *ZeroImpute*, BPCA, LSImpute and KNN all drop significantly compared to CMVE which demonstrates its capacity to detect conserved links (See Figures 5-7) and shows its estimation accuracy, especially for higher numbers of missing values. It is also clear from the results that simply ignoring missing values (*ZeroImpute*), decreases the CL rate for gene co-regulation, with the ensuing likelihood that important links which may be of interest for subsequent biological analysis, will be overlooked (See Figures 4-8).

Similarly, Figures 8-12 clearly show the lower *false negative* rate of CMVE for all the three datasets across the range of missing values, compared to other estimation strategies. Once again it upholds its performance for higher missing values especially in the range of 4% and 5% missing data (Figures 11 and 12) compared to other imputation strategies whose performance degrades noticeably at higher missing values. This confirms that CMVE, as a core estimator in CF-GeNe, not only has ability to conserve the true links, but also has a lower false negative rate. Not surprisingly, simply ignoring missing values in the gene expression data can introduce high FN rate as clearly witnessed in Figures 10-12.

To further investigate the affect of missing values on genetic inference, significant genes were iteratively selected from these imputed matrices and compared with the Ъ to computed *True Positive* (TP) rate.

DEFINITION 3: A *True Positive* gene is concomitantly present in the selected gene set, when genes are selected both with and without randomly introduced missing values.

Figure 13 demonstrates the consistently higher TP rate achieved by CMVE, compared to other imputation techniques for a set of 50 genes which again shows its better estimation capability. A similar improved performance was also observed for various ranges of selected genes.

Finally, to quantifiably corroborate the superior imputation performance analysis for GRN inference with CMVE as its core missing value estimator, the *two-sided Wilcoxon Rank sum statistical significance* test was applied. This is formally defined as follows:

DEFINITION 4: The *P*-value of significance for the hypothesis H_0, $Y = Y_{est}$ where Y and Y_{est} are the actual and estimated matrices respectively, is calculated using:

$$H_0,\ P\text{-}Value = 2P_r(R \le y_r) \tag{11}$$

where y_r is the sum of the ranks of observations for Y and R is the corresponding random variable.

The motivation for using this particular test is that compared to some other parametric significance tests [26], it does not mandate data of equal variance, which is vital given that the data variance can become severely disturbed due to erroneous estimation, especially for instance, when applying *ZeroImpute* (See Figure 14).

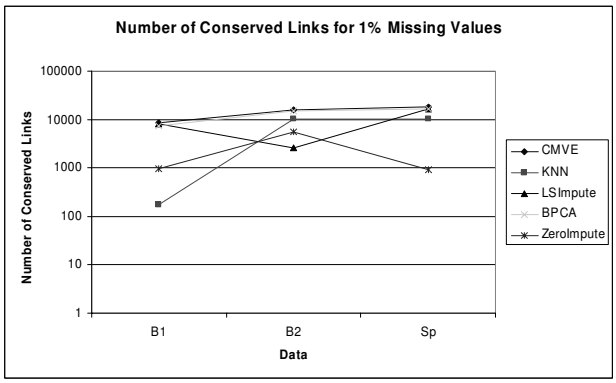

Figure 3: Number of Conserved Gene Regulatory Links for 1% Missing Values

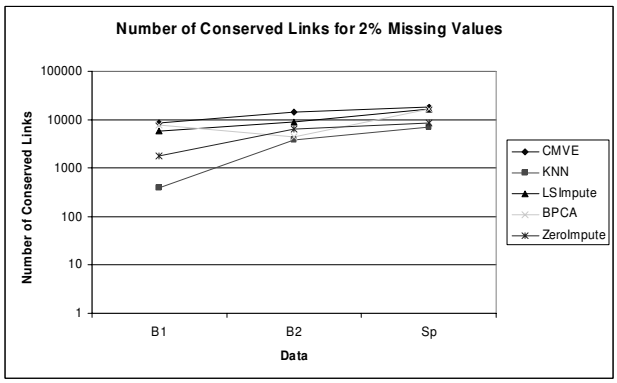

Figure 4: Number of Conserved Gene Regulatory Links for 2% Missing Values

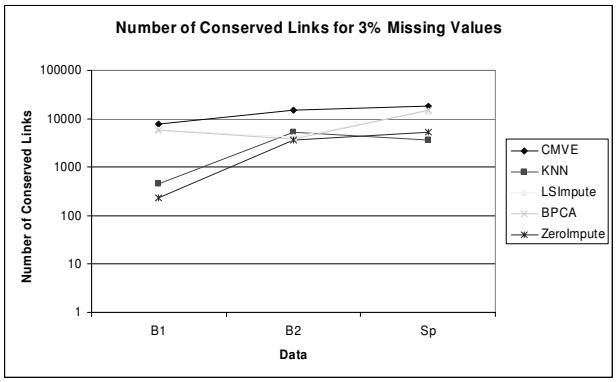

Figure 5: Number of Conserved Gene Regulatory Links for 3% Missing Values

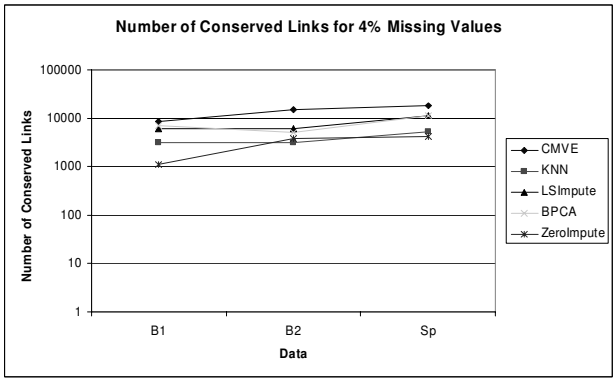

Figure 6: Number of Conserved Gene Regulatory Links for 4% Missing Values

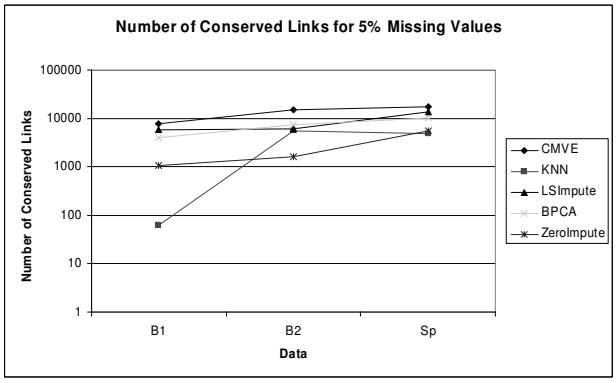

Figure 7: Number of Conserved Gene Regulatory Links for 5% Missing Values

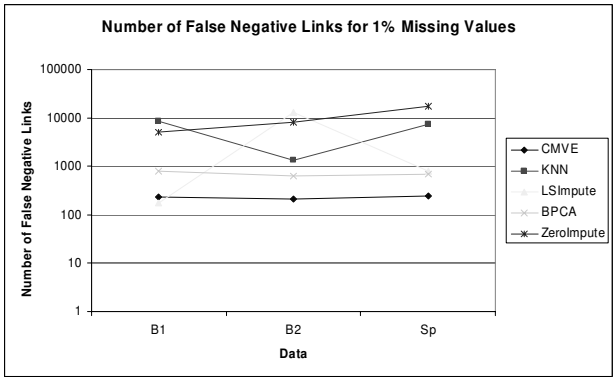

Figure 8: Number of False Negative Gene Regulatory Links for 1% Missing Values

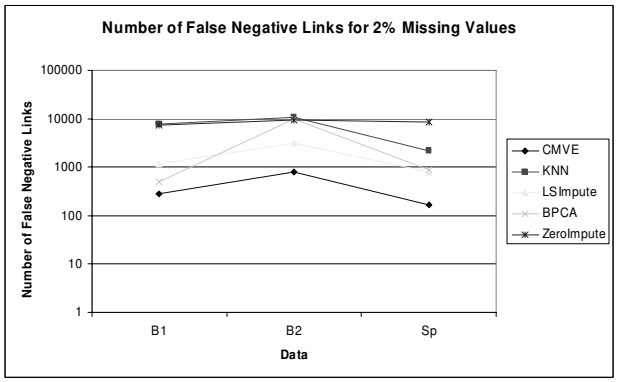

Figure 9: Number of False Negative Gene Regulatory Links for 2% Missing Values

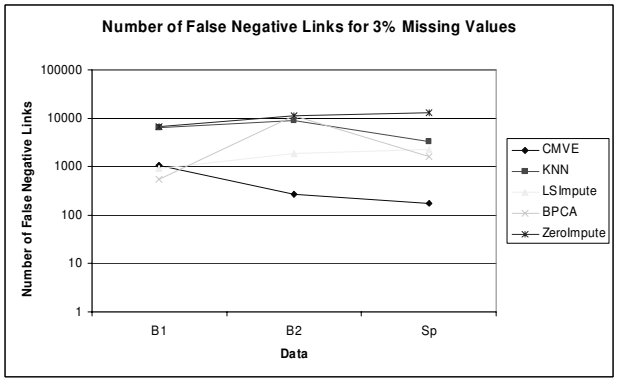

Figure 10: Number of False Negative Gene Regulatory Links for 3% Missing Values

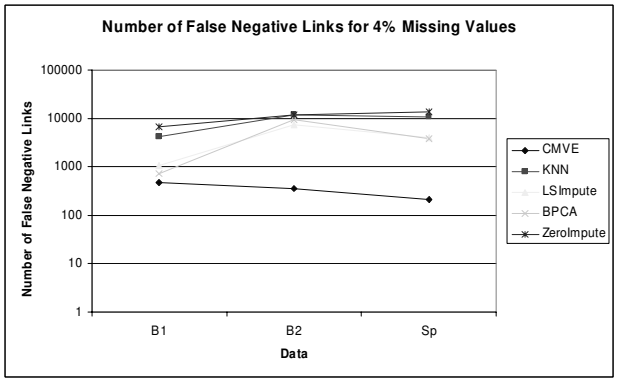

Figure 11: Number of False Negative Gene Regulatory Links for 4% Missing Values

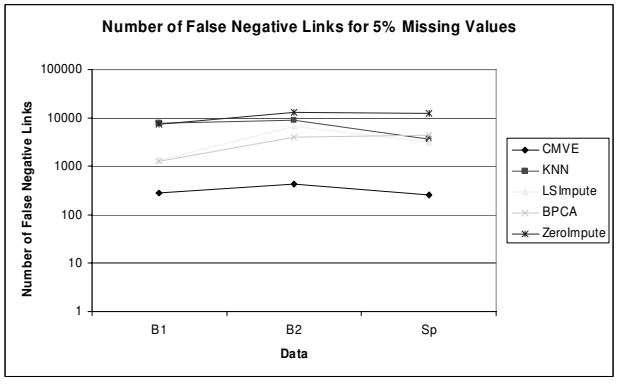

Figure 12: Number of False Negative Gene Regulatory Links for 5% Missing Values

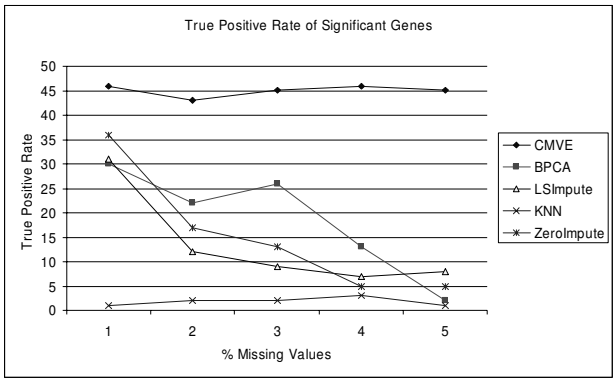

Figure 13: True Positive Rate of the 50 Most Significant Genes

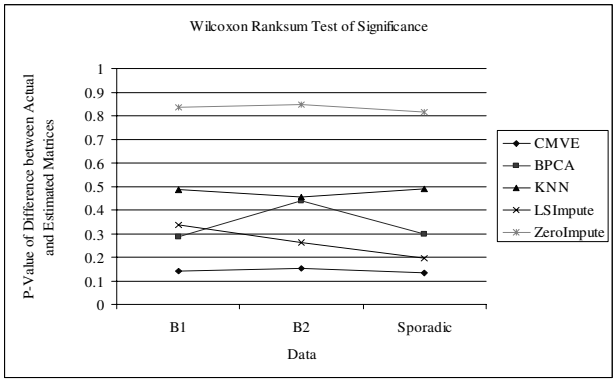

Figure 14: Statistical Significance Test Results of Various Imputation Methods

V. CONCLUSION

This paper has presented a novel *Collateral-Fuzzy Gene Regulatory Network Reconstruction* (CF-GeNe) framework for *Gene Regulatory Network* (GRN) inference, employing the *Collateral Missing Value Estimation* (CMVE) algorithm as its core, to accurately impute missing values in the gene expression data. CF-GeNe mimics the inherent fuzzy nature of gene co-regulation by applying fuzzy clustering principles based upon the *fuzzy c-means* algorithm to automatically adapt to the underlying data distribution. Empirical results confirmed that CF-GeNe can infer most co-regulated links even in the presence of a high number of missing values, compared to other commonly applied imputation techniques including *Least Square Impute*, *K-Nearest Neighbour* and *Bayesian Principal Component Analysis*. This striking performance improvement has been further endorsed in *gene selection* and the *Wilcoxon Ranksum Significance Test* results, both of which substantiated the effectiveness of the CF-GeNe model to infer GRN interactions in noisy genetic expression data**.**

REFERENCES

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasen-beek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Down-ing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lan-der, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, pp. 286(5439):531-537, 1999.

[2] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Statistical Neural Networks and Support Vector Machine for the Classification of Genetic Mutations in Ovarian Cancer," *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)'04, USA.*, pp. 140-146, 2004.

[3] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Missing Values Imputation for DNA Microarray Data using Ranked Covariance Vectors," *The International Journal of Hybrid Intelligent Systems (IJHIS)*, vol. ISSN 1448-5869, 2005.

[4] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, pp. 77-78, 2002.

[5] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Collateral Missing Value Estimation: Robust missing value estimation for consequent microarray data processing," *Lecture Notes in Artificial Intelligence (LNAI), Springer-Verlag*, pp. 274-283, 2005.

[6] M. S. B. Sehgal, I. Gondal, and L. Dooley, "A Collateral Missing Value Estimation Algorithm for DNA Microarrays," *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), USA*, pp. 377-380, 2005.

[7] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian Missing Value Estimation Method for Gene Expression Profile Data," *Bioinformatics*, vol. 19, pp. 2088-2096, 2003.

[8] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, pp. 4348-4355, 2005.

[9] M. S. B. Sehgal, I. Gondal, L. Dooley, and R. Coppel, "AFEGRN: Adaptive Fuzzy Evolutionary Gene Regulatory Network Re-construction Framework," *World Congress on Computational Intelligence: Fuzzy Systems.*, 2006.

[10] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper, "Clustering of gene expression data using a local shape-based similarity measure 10.1093/bioinformatics/bti095," *Bioinformatics*, vol. 21 http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/7/1069, pp. 1069-1077, 2005.

[11] G. Fort and S. Lambert-Lacroix, "Classification using partial least squares with penalized logistic regression," *Bioinformatics*, vol. 21, pp. 1104-1111, 2005.

[12] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis 10.1093/bioinformatics/bti517," *Bioinformatics*, vol. 21 http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/15/3201, pp. 3201-3212, 2005.

[13] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Collateral Missing Value Imputation: a new robust

missing value estimation algorithm for microarray data," *Bioinformatics*, vol. 21(10), pp. 2417-2423, 2005.
[14] P. Du, J. Gong, E. S. Wurtele, and J. A. Dickerson, "Modeling gene expression networks using fuzzy logic," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, pp. 1351-1359, 2005.
[15] X. J. Zhou, Ming-Chih, J. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan, and W. H. Wong, "Functional annotation and network reconstruction through cross-platform integration of microarray data," *Nature Biotechnology*, vol. 23, pp. 238 - 243, 2005.
[16] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, pp. 487-501, 2004.
[17] P. Y. Chen and P. M. Popovich, *Correlation: Parametric and Nonparametric Measures*, 1st edition ed: SAGE Publications, 2002.
[18] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discovery 10.1093/bioinformatics/bti1049," *Bioinformatics*, vol. 21 http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/suppl_1/i213, pp. i213-221, 2005.
[19] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*: Plenum Press, 1981.
[20] M. Kaufmann, *Data Mining: Practical Machine Learning Tools and Techniques*, 2 ed: Elsevier, 2005.
[21] P. Bobko, *Correlation and Regression: Principals and Applications for Industrial/Organizational Psychology and Management (Organizational Research Methods)*, 2 ed: SAGE Publications, 2001.
[22] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.
[23] T. H. Bø, B. Dysvik, and I. Jonassen, "LSimpute: Accurate estimation of missing values in microarray data with least squares methods," *Nucleic Acids Res.*, pp. 32(3):e34, 2004.
[24] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent, "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med*, pp. 22; 344(8):539-548, 2001.
[25] R. P. Abelson, *Statistics as Principled Argument*: Lawrence Erlbaum Associates, 1995.
[26] Z. Sidak, P. K. Sen, and J. Hajek, *Theory of Rank Tests (Probability and Mathematical Statistics)*: Academic Press, 1999.

Muhammad Shoaib B. Sehgal is a PhD student at *Monash University, Australia*. He earned B.Sc. (HONS) degree in Computer Sciences from *University of Engineering & Technology* (UET), Lahore Pakistan (2002) with higher distinction.

He joined *Lahore University of Management Sciences* (LUMS) as a Research Associate where he did a pioneering job to establish *Artificial Intelligence & Natural Language Processing Lab.* He was the first one to publish refereed papers in that Lab. Later on he joined UET Lahore, Pakistan as a Lecturer. During his stay at Monash he has worked on various research grant projects and has published several research papers in Journals and conferences of international repute including *Journal of Bioinformatics*. He is a named inventor of a patent on imputation methods and is serving as a referee for high impact journals/conferences including *Journal of Bioinformatics* (Oxford University Press) and *Journal of Biomedical Informatics* (Elsevier Publishers).

Mr. M. Sehgal's research interests include: Bioinformatics, Drug Discovery, Missing Value Imputation, Gene Regulatory Networks, Machine Learning, Neural Networks, Support Vector Machines, Pharmacogenomics/genetics, Class Prediction and Speech/Image Processing.

Iqbal Gondal is working as a senior lecturer with Monash University Australia and teaches project management and Data communication and Networks. Prior joining Monash, he worked in the capacity of research fellow and senior systems engineer for seven years in Singapore and Australia with General Motors, Singapore Government and other industries. He has experience of ten years in network design and development, project management, System design and integration, SCADA, intelligent techniques, adaptive systems and wireless switches for financial services. He is a member of IEEE for fifteen years and is also member of IEE and Institute of Engineer Australia. He has authored forty research papers in reputable International conferences and Journals. His research interests are: Bioinformatics, Pervasive networks, Technology transfer, Information fusion, On-line education, and Networking and Communication.

Laurence S. Dooley (M'81—SM'93) received his B.Sc.(Hons), M.Sc. and Ph.D. degrees in Electrical and Electronic Engineering from the University of Wales, Swansea in 1981, 1983 and 1987 respectively.

Since 1999, he has been Professor of Multimedia Technology in the Faculty of Information Technology, Monash University, Australia, where his major research interests are in multimedia signal processing, mobile communications, image/video object segmentation, bioinformatics, wireless sensor networks and applied R&D Commercialisation strategies for regional small businesses. He has edited one book and published over 175 international scientific peer-reviewed journals, book chapters and conference papers, and was jointly awarded the IEEE Communications Society sponsored Outstanding Paper Prize at this year's *1st International Conference on Next Generation Wireless Systems* (ICNEWS'06). He has supervised 14 PhD/Masters research students to successful completion and is the co-inventor of two patents, as well as being in receipt of numerous external grants from both government and industry to support his multi-faceted research projects.

Professor Dooley is a Senior Member of the IEEE, a Chartered Engineer (C.Eng), and a corporate member of the British Computer Society.