Transcriptional Gene Regulatory Network Reconstruction Through Cross Platform Gene Network Fusion

Muhammad Shoaib B. Sehgal[1,3], Iqbal Gondal[1,3], Laurence Dooley[1], Ross Coppel[2,3], and Goh Kiah Mok[4]

[1] Faculty of IT, Monash University, Churchill VIC. 3842, Australia
{Shoaib.Sehgal,Iqbal.Gondal,
Laurence.Dooley}@infotech.monash.edu.au
[2] Department of Microbiology, Monash University Clayton VIC. 3800, Australia
Ross.Coppel@med.monash.edu.au
[3] Victorian Bioinformatics Consortium Wellington Road, Clayton VIC. 3800, Australia
[4] SIMTech Institute of Technology, Nanyang Drive, Singapore
kmgoh@SIMTech.a-star.edu.sg

Abstract. Microarray gene expression data is used to model differential activity in *Gene Regulatory Networks* (GRN) to elucidate complex cellular processes, though network modeling is susceptible to errors due to both noisy nature of gene expression data and platform bias. This intuitively provided the motivation for the development of an innovative technique, which effectively integrates GRN using cross-platform data to minimize the two aforementioned effects. This paper presents a GRN integration (GeNi) framework that fuses cross-platform GRN to remove platform and experimental bias using the *Dempster Shafer Theory of Evidence*. The proposed model estimates gene co-regulation strength by using mutual information and removes spurious co-regulations through data processing inequality. The method automatically adapts to the data distribution using Belief theory, which does not require a preset threshold to accept co-regulated links. [1]GeNi is applied to identify common cancer-related regulatory links in ten different datasets generated by different microarray platforms including *cDNA* and Affymetrix arrays. Experimental results demonstrate that GeNi can be effectively applied for GRN reconstruction and cross-platform gene network fusion for any gene expression data.

1 Introduction

Gene expression analysis has been widely used for different biological studies. Several statistical and computational intelligence modeling methods have been applied for this purpose. While these techniques provide biologists with valuable insights of

[1] GeNi Software and Supplementary Material can be downloaded from www.gscit.monash.edu.au/~shoaib or can be requested by emailing at {Shoaib.Sehgal AT gmail.com}.

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 274–285, 2007.

different biological processes, most analyses are based on over/under expressed genes studies despite the fact that differential expression analysis do not fully harness the potential of microarray gene expression data because genes are treated independently and interactions between them are overlooked [1].

Gene Regulatory Networks (GRN) model how genes regulate different metabolism and can map the casual pathways. GRN reconstruction is however error prone due to the noisy nature of microarray data and microarray platform and experimental bias. One possible solution is to integrate networks constructed through different microarray platforms, e.g. *cDNA* or Affymetrix high density oligonucleotide arrays, under either different or similar experimental conditions, though this is a challenging task because data generated by different platforms is not directly comparable. The objectives include being able to construct models capable of inferring knowledge from thousands of genes at a time, assist in understanding complex genetic interactions and to integrate networks constructed from heterogeneous microarray datasets [2, 3].

Previous attempts to integrate cross platform GRN include Zhou *et al.* [3] who proposed cross platform GRN fusion using second-order expression analysis while Choi *et al.* [1] studied different types of cancer links using cross-platform analysis. In both studies the regulatory pathway was only considered if it was present in more than a certain number of experiments T, where the selection of this threshold T was empirically derived with no formal mathematical foundations so that selection of an incorrect T could inevitably lead to erroneous results. Furthermore, most GRN modeling techniques incur several limitations, including exponential time space complexity, unrealistic GRN assumptions such as acyclic network by bayesian networks, overfitting and under constrained regression analysis [4]. This has created a need for suitable techniques that are scaleable and do not impose unrealistic assumptions on the network structure.

This paper proposes a novel *Gene Regulatory Network Integration* (GeNi) Framework to model GRN. The proposed model integrates GRN generated from different platforms using *Dempster Shafer Theory of Evidence* (DSTE) [5]. The GeNi computes gene to gene co-regulation using mutual information. Mutual Information is selected due to its proven improved performance compared to commonly used correlation based methods and Bayesian Networks [4]. The other advantage of using mutual information is that it does not enforce the acyclic assumption as posed by Bayesian networks and is more scaleable than dynamic Bayesian networks, which remove this acyclic restriction. Mutual Information for GRN reconstruction has been used by Baso *et al.* [4] and Zhao *et al.* [6] though these methods can only be used for single data and doesn't reconstruct network through cross platform network integration to remove bias and minimize the impact of noise. Also, GeNi has added advantage over other mutual information based techniques that it does not require threshold to select co-regulated links because it uses belief theory to accept/reject co-regulated links. After mutual information computation, GeNi then prunes the network using data processing inequality to remove the spurious co-regulations. Finally, the fusion of different gene networks is performed by using the belief theory.

The proposed model is tested for its application to find tumor specific links in various cancer datasets generated by different cDNA and Affemtrix microarray platforms. The results corroborate that GeNi can be effectively used to fuse cross-platform GRN.

The rest of the paper is organized as follows: Section 2 presents GeNi model in detail. Analysis of results is presented in Section 3 while conclusions are drawn in Section 4.

2 Gene Regulatory Network Integration (GeNi) Model

The complete GeNi framework is formalized in Fig. 1. Gene expression data is firstly preprocessed to remove noise and outliers followed by gene to gene mutual information computation to measure gene co-regulation strength. The network is then pruned using data processing inequality, before network fusion is undertaken using DSTE. Each of these constituent blocks is now considered in the following sub-sections, with the rationale for the choice of each algorithm being delineated.

2.1 Pre-processing

The data is preprocessed to minimize the affect of noise on subsequent analysis. Negative values in Microarray data are considered as missing and genes with greater than 70% missing values and less than 4 observations are filtered out. Gene expression data is then re-parameterized using rank transformation to convert each gene into equally spaced expressions between the interval [0 1] [4] (Step1 - Fig. 1). Missing values in the data are then imputed by their gene averages. Finally, each clone was mapped to UniGene accession build # 162 to manage heterogeneous data.

Pre Condition: Gene expression matrices Y_N and Y_T for normal and cancerous data.

1. Preprocess (Section 2.1)
2. Construct GRN using mutual information (Section 2.2).
3. Remove spurious gene links using data processing inequality (Section 2.3).
4. Fuse cross platform networks using Dempster Shafer theory (Section 2.4).
5. Compare Normal and Cancerous fused networks to find out *Conserved*, *Broken* and *Tumor links* (Section 2.5).
6. Stop

Post Condition: Gene Regulatory Networks N_n, N_t for normal and cancerous data.

Fig. 1. GRN integration algorithm

The pre-processing step is followed by GRN reconstruction for each data set (See Step2 - Fig. 1). Following sub-section explains this step in detail.

2.2 GRN Reconstruction

After pre-processing, GeNi computes pair wise mutual information for all genes to construct gene networks. The mutual information $I(g_1,g_2)$, between two genes g_1 and g_2 is computed using *Gaussian Kernel Estimator* as:

$$I(g_1, g_2) = \frac{1}{m} \sum_{i=1}^{m} log \left[\frac{p(g_{1i}, g_{2i})}{p(g_{1i}) p(g_{2i})} \right] \tag{1}$$

where

$$p(g_{1i}) = \frac{1}{\sqrt{2\pi} N \alpha_1} \sum_j e^{\frac{(g_{1i} - g_{1j})}{2\alpha_1^2}}, \tag{2}$$

$$p(g_{2i}) = \frac{1}{\sqrt{2\pi} N \alpha_1} \sum_j e^{\frac{(g_{2i} - g_{2j})}{2\alpha_1^2}}, \tag{3}$$

$$p(g_{1i}, g_{2i}) = \frac{1}{\sqrt{2\pi} N \alpha_2} \sum_j e^{\frac{(g_{1i} - g_{1j}) + (g_{2i} - g_{2j})}{2\alpha_2^2}}. \tag{4}$$

where α_1 and α_2 are tunable parameter and computed by *Monte Carlo Simulations* [7] using bi-variate normal probability densities [4].

Mutual information computation step is followed by network pruning using Data Processing Inequality which is explained in the next sub-section.

2.3 Network Pruning

When two genes g_1 and g_2 are interacting through a third gene G_3 and $I(G_1, G_2 | G_3)$ is zero then these genes are directly interacting with each other if:

$$I(g_1, g_3) \leq I(g_1, g_2) \text{ and } I(g_1, g_3) \leq I(g_2, g_3). \tag{5}$$

As this property is asymmetric it has the possibility of rejecting some of the loops or interactions between three genes whose information may not be fully modeled by pair wise mutual information. The introduction of a tolerance threshold addresses this problem as well as provides the advantage of avoiding rejection of some of thetriangular links and loops [4].

2.4 Cross Platform GRN Fusion

GeNi fuses cross platform networks using DSTE (Step 4 - Fig. 1.) [5], as alluded to in Section 1. The theory extends Bayesian theory to evaluate beliefs from different evidences. The DSTE allows beliefs to be represented by upper and lower probability intervals normally referred to as belief and plausibility respectively [5].

The DSTE assumes that the information sources are independent of each other. This assumption makes it further feasible to use in GeNi as it first constructs GRN independent of each other using data generated by heterogeneous platforms under independent studies.

The application of this theory for cross platform GRN fusion requires a definition of the degree of belief (mass functions) to assign masses, normally referred to as probability value. The DSTE doesn't mandate the method of computing these masses (probabilities) [8], which makes it a more generalized approach than Bayesian theory [9-11]. It adds flexibility to GeNi, as belief masses can be calculated using any GRN reconstruction method (Correlation, Probability value or Mutual Information) to compute the gene co-regulation.

For GRN fusion, the $\Omega = \{R, NR\}$ represents mutually exclusive event space for *Co-Regulated* (R) and *Non Co-Regulated* (NR) links, called frame of discernment or universe of discourse. The $2^{\Omega} = \{ \phi, \{R\}, \{NR\}, \{R, NR\}\}$ represents the set of all subsets of Ω, and classes in Ω are considered mutually exclusive. Let A be a non-zero degree of belief in 2^{Ω}, called the focal element where:

$$\sum_{A \subseteq \Omega} m(A) = 1 \text{ and } m(\phi) = 0 \tag{6}$$

Focal elements and their masses construct an evidence structure, which can be expressed as:

$$\{(A, m(A)) \mid A \subseteq \Omega, m(A) > 0\} \tag{7}$$

The value $m(A)$ represents the weight of evidence in favor of complete set A. The belief function, which is a sum of masses of all subsets of hypothesis for R and NR, can be computed as:

$$Bel(R) = \sum_{A \subseteq R} m(R) \text{ and } Bel(NR) = \sum_{A \subseteq NR} m(NR) \tag{8}$$

The same information can be computed by calculating plausibility or upper probability value, which is the sum of the masses of all sets whose intersection with the hypothesis is empty [12]. The plausibility of R can be defined as:

$$Pl(R) = \sum_{A \cap R = \phi} m(R) \text{ and } Pl(NR) = \sum_{A \cap NR = \phi} m(NR) \tag{9}$$

Similarly plausibility for ϕ is

$$Pl(\phi) = 0 \tag{10}$$

The relation between plausibility and belief masses can be expressed as: $Bel(R) \leq Pl(R) \text{ and } Pl(R) = 1 - Bel(\overline{R})$ where $\overline{R} = \Omega - R$.

The belief masses are the gene co-regulation probability or correlation values computed in STEP 2 (Fig. 1). Figure 2 represents a schematic diagram for the fusion of belief masses where g_1, g_2, g_3 and g_4 are the genes sets. These genes are triggered by different regulation weights (P_1, P_2 …P_n) that represent belief masses $m(R)$) and $m(NR)$ for co-regulated and not co-regulated weights respectively, in experiments $\{E_1, E_2 \ldots E_n\}$. The combined belief F_k for the gene co-regulation is computed by:

$$Bel(F_k) = \frac{\sum_{R_i \cap NR_j = F_k : F_k \neq \phi} m(R_i) \oplus m(NR_j)}{1 - \sum_{R_i \cap NR_j = \phi} m(R_i) \oplus m(NR_j)} \tag{11}$$

where $\oplus$ represents the orthogonal sum and can be computed for n experiments as:

$$\bigoplus_{i=1}^{n} m_i(A) = \sum_{A_1 \cap \ldots \cap A_n = A} \prod_{i=1} m_i(A_i) \tag{12}$$

The combination can be normalized by introducing normalization factor N_b such that:

$$\bigoplus_{i=1}^{n} m_i(A) = N_b \sum_{A_1 \cap \ldots \cap A_n = A} \prod_{i=1} m_i(A_i) \tag{13}$$

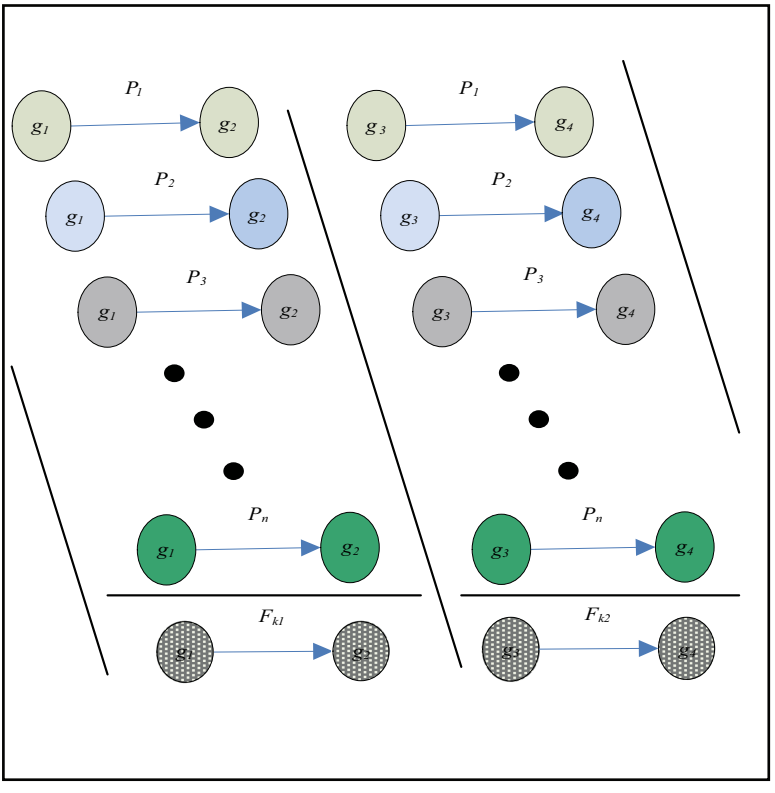

Fig. 2. A schematic diagram presenting fusion of belief masses in GeNi. The probabilities P_1 … P_n are the co-regulation weights of $g_1 \rightarrow g_2$ links computed from heterogeneous datasets. The F_k is the final co-regulation weight calculated using DSTE. The link is accepted or rejected based on the value of F_k.

The value of N_b was set to 0.5; however, GeNi doesn't restrict the use of any normalization value. The denominator in (11) normalizes the output to the safer belief function and serves to distribute any mass associated with ϕ intersections of beliefs to the non empty intersections [13]. The genes co-regulation link is added to the fused network if the combined belief of *R* is higher than *NR* where F_k represents the link weight.

After the integrated network is constructed, the network is pruned to remove the platform bias and the links that occur by chance. Only those regulatory links that are present in more than E_n experiments where $E_n > n/4$ and $E_n > 8$ [3] are added to the final fused network. It should be noted, however, that network pruning is an optional step in GeNi and this step is different from threshold-based integration based methods, as they don't consider the link co-regulation weight-age while selecting the links for the final integration, as mentioned earlier, GeNi adds/removes links, primarily based on belief masses *m(R)* and *m(NR)*.

2.5 Network Comparison

Once the fused networks have been constructed they are compared for Broken, Conserved and Tumor links (See Step 5 - Fig. 1). The precise definition for each of these links is now given:

Definition 1. A link is a *Conserved Link* if it is present in both normal and tumor networks.

Definition 2. A link is a *Broken Link* if it is present in normal network and is missing in tumor network.

Definition 3. A link is a *Tumor Link* if it is not present in normal network but exists in tumor networks.

The next section provides analysis of GRN constructed using GeNi.

3 Analysis of Results and Discussion

For cross platform GRN fusion, 10 different datasets under 11 different experimental conditions (Table 1), designed for the comparison of primary cancer and non cancer counterpart, were used. These datasets were generated using different microarray platforms including cDNA and Affmetrix GeneChip. The datasets were collected from breast, pancreas, colon, brain, bladder, ovary, uterus, kidney, liver, lung, lymphoma, stomach and prostate tissues and had 5603, 17660, 3697, 3732, 5575, 13171, 12065, 5983, 4615, 24822, 6593 genes respectively (Table 1). The total number of genes in all experiments were 103,516 (Choi *et al.* [1] for further details). To construct the fused network, we selected 61 commonly present, regulated genes from the above datasets. The gene networks were first individually constructed using (Steps 1-3 - Fig. 1) and then these network were integrated using belief theory to form fused

Table 1. Datasets

Tissues	Platforms	Normal Samples	Tumor Samples
Breast [14]	cDNA	13	13(72)
Colon [15]	Hu6800	22	22
Kidney [16]	cDNA	81	81
Liver [17]	cDNA	76	76(104)
Lung [18]	U95A	17	17(127)
Lymphoma [19]	cDNA	31	31(77)
Pancreas [20]	cDNA	14	22
Prostate [21]	U95A	50	52
Stomach [22]	cDNA	29	29(103)
Brain,	Hu6800	8	20
Bladder,	Hu35KSubA	7	11
Ovary	Hu35KSubA	3	11
Uterus [23]	Hu35KSubA	6	10

networks for both normal and tumor data (Figs. 3-5). These fused normal and tumor networks were then compared to search for *Broken*, *Conserved* and *Tumor* links.

Table 2 shows selected Broken, Conserved and Tumor links. The results demonstrate that 52% of the links were broken links in tumor tissue samples which were present in normal tissues while only 2% links were newly created in tumor cells compared to normal tissues. Only 45% of the links were conserved between normal and tumor tissues. These links can be used to monitor patient's response to certain treatment. For instance, if the response of patient to the treatment is positive then the number of conserved links should increase while concomitantly decreasing the broken and tumor links.

Figure 3 plots a selected section of normal and tumor networks for comparison where complete normal and tumor networks for commonly selected genes are shown in Figs. 4 and 5 (Individual networks can be downloaded from www.gscit.monash. edu.au/~shoaib/GeNi.html). It is evident from Figs. 3-4 that normal data has high percentage of connected nodes compared to tumor network. Figure 3 shows several inserting observations for instance, a link from *nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3* (HS.172674) to *protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform* (HS.431156) is present in normal network but is broken in the tumor network. A new link is created between *protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform* (HS.431156) and *Sulfotransferase family, cytosolic, 1A, phenol-preferring, member* (HS.368950) in tumor network, which was not present in the normal network. Figure 3 also shows the conserved link between *nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3* (HS.172674) and *Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1* (HS.368950), which is present in both datasets.

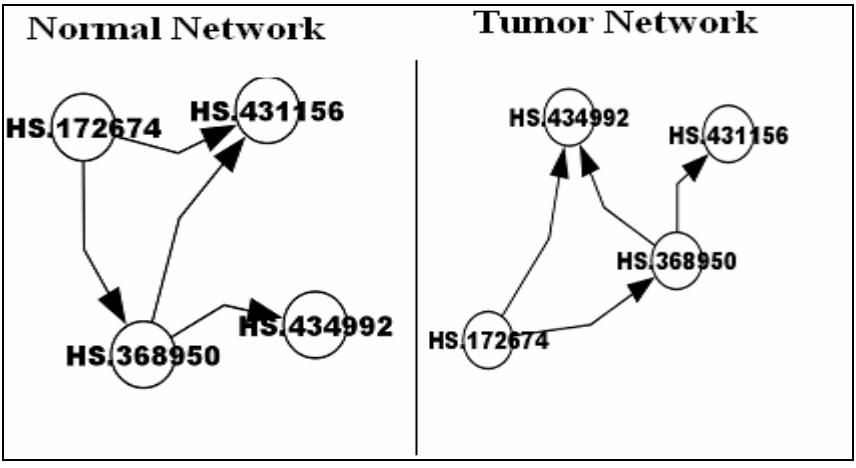

Fig. 3. Cross-section of normal and tumor tissue networks

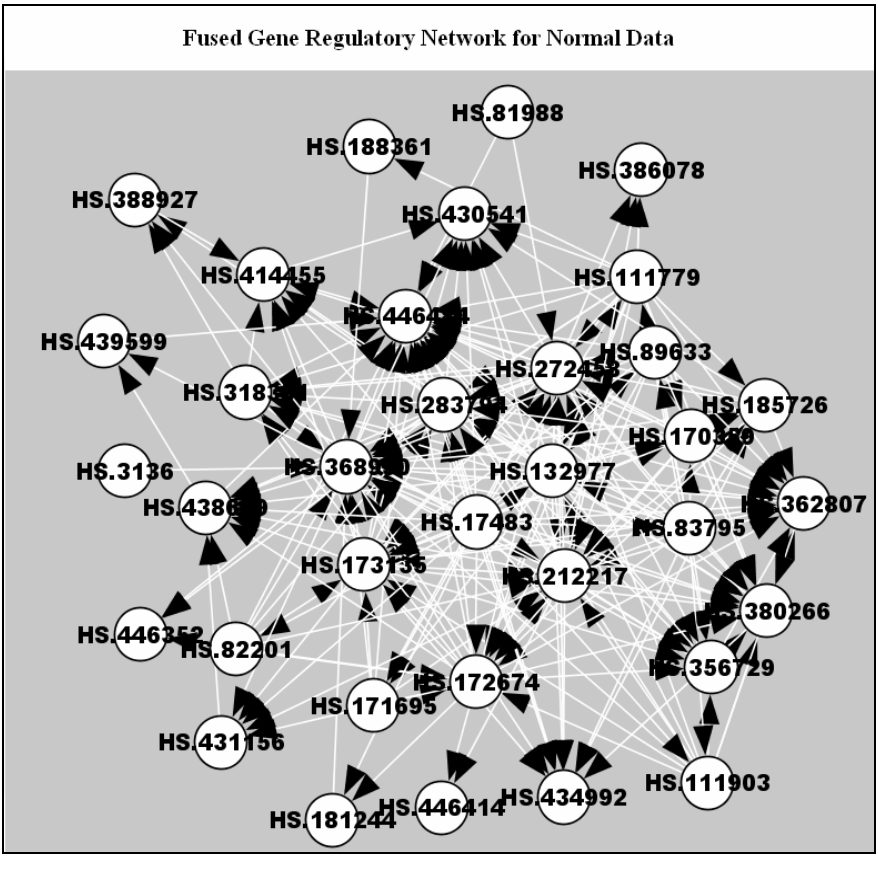

Fig. 4. Complete fused network of normal tissues

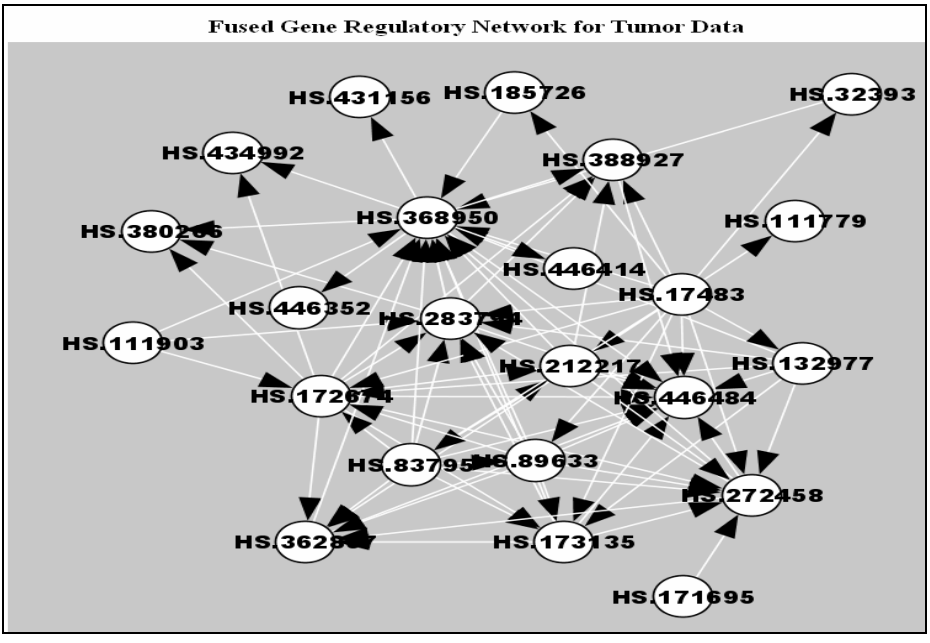

Fig. 5. Complete fused network of tumor tissues

Table 2. Number of genes involved in GRN links

GRN Links	Broken Links	Conserved Links	Tumor Links
% Links	52%	45%	2%

These results above all demonstrate that GeNi can indeed be used for cross-platform network fusion however; further wet laboratory results are required in order to completely verify the model.

4 Conclusions

The paper has presented GRN integration (GeNi) framework to fuse cross-platform GRN in order to remove platform and experimental bias. The proposed model estimates gene co-regulation strength by using mutual information and removes spurious co-regulations by using data processing inequality. The method automatically adapts to the data distribution using Belief theory and hence does not require preset threshold to accept the co-regulated links which makes method more robust for GRN reconstruction. The GeNi was used to find common cancer related regulatory links in ten different datasets generated by different microarray platforms including cDNA and Affymetrix arrays. The experimental results demonstrated that GeNi can be applied successfully for GRN reconstruction and cross-platform gene network fusion for various types of genetic data.

References

[1] Choi, J.K., Yu, U., Yoo, O.J., Kim, S.: Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 21, 4348–4355 (2005)
[2] Fort, G., Lambert-Lacroix, S.: Classification using partial least squares with penalized logistic regression. Bioinformatics 21, 1104–1111 (2005)
[3] Zhou, X.J., Ming-Chih, Kao, J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E., Wong, W.H.: Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature Biotechnology 23, 238–243 (2005)
[4] Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human B cells. Nature Genetics 37, 382–390 (2005)
[5] Shafer, G.: Mathematical Theory of Evidence. Princeton Univ. Press, Princeton, NJ (1976)
[6] Zhao, W., Serpedin, E., Dougherty, E.R.: Inferring gene regulatory networks from time series data using the minimum description length principle. Bioinformatics 22(17), 2129–2135 (2006)
[7] Casella, G., Robert, C.P.: Monte Carlo Statistical Methods. Springer, Heidelberg (2005)
[8] Malpicaa, J.A., Alonsoa, M.C., Sanz, M.A.: Dempster–Shafer Theory in geographic information systems: A survey, Expert Systems with Applications, vol. 32. Elsevier, Amsterdam (2007)
[9] Hegarat-Mascle, S.L., Bloch, I., Vidal-Madjar, D.: Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing. IEEE Trans. Geosci. Remote Sensing 35, 1018–1031 (1997)
[10] Bloch, I.: Some aspects of Dempster-Shafer evidence theory for classification of multimodality medical images taking partial volume effect into account. Pattern Recognition Letters 17, 905–919 (1996)
[11] Rombaut, M., Zhu, Y.M.: Study of Dempster–Shafer for image segmentation applications. Image Vision Comput. 20, 15–23 (2002)
[12] Barnett, J.A.: Calculating Dempster-Shafer plausibility. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 599–602 (1991)
[13] Murphy, R.R.: Dempster-Shafer Theory for Sensor Fusion in Autonomous Mobile Robots. IEEE Transactions on Robotics and Automation 14, 197–206 (1998)
[14] Sorlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., Rijn, M.v.d., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lonning, P.E., Borresen-Dale, A.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. 11, 98(19), 10869–10874 (2001)
[15] Notterman, D.A., Alon, U., Sierk, A.J., Levine, A.J.: Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays. Cancer Res. 61, 3124–3130 (2001)
[16] Boer, J.M., et al.: Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. Genome Research 11, 1861–1870 (2001)
[17] Chen, X., Cheung, S.T., So, S., Fan, S.T., et al.: Gene Expression Patterns in Human Liver Cancers. Mol. Biol. Cell 13, 1929–1939 (2002)

[18] Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.F., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. 13790–13795 (2001)

[19] Alizadeh, A.A., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)

[20] Lacobuzio-Donahue, C.A., et al.: Exploration of Global Gene Expression Patterns in Pancreatic Adenocarcinoma Using cDNA Microarrays. Am. J. Pathol. 162, 1151–1162 (2003)

[21] Singh, D., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1 (2002)

[22] Chen, X., et al.: Variation in gene expression patterns in human gastric cancers. Mol. Biol. Cell 14, 3208–3215 (2003)

[23] Ramaswamy, S., et al.: Multiclass cancer diagnosis using tumour gene expression signatures. Proc. Natl. Acad. Sci. 98(26), 15149–15154 (2001)