Construction of Co-expression and Co-regulation Network with Differentially Expressed Genes in Bone Marrow Stem Cell Microarray Data

Paramita Biswas, Bandana Barman and Anirban Mukhopadhyay

Abstract It is important to understand the interaction mechanism among co-expressed and co-regulated genes in stem cell to restrict the abnormal growth of cell tissues (tumor) which may lead to cancer. In this article, differentially co-expressed and co-regulated genes exist in normal stem cells and stem cell derived tumors are identified from sample Bone Marrow microarray data. By performing statistical t-test between sample groups, first we have identified differentially expressed genes (DEG). Then up-regulated (UR) and down-regulated (DR) genes are separated by setting a p-value cutoff at 0.001. After identifying the differentially expressed genes, distinguished co-expressed up-regulated and down-regulated genes are found. Subsequently, we have constructed pair-wise co-expression networks with the co-expressed genes. Finally, we have studied the significance of co-expressed genes with gene ontology (GO) and we have found significant GO-ids. This study is expected to lead to finding of pathways for diseases.

Keywords Stem cell · Differentially expressed genes · t-test · Gene ontology · Co-regulated and co-expressed genes

P. Biswas · A. Mukhopadhyay (✉)
Department of Computer Science and Engineering, University of Kalyani,
Kalyani 741235, West Bengal, India
e-mail: anirban@klyuniv.ac.in

P. Biswas
e-mail: paramita.biswas1991@gmail.com

B. Barman
Department of Electronics and Communication Engineering, Kalyani Government
Engineering College, Kalyani 741235, West Bengal, India
e-mail: bandanabarman@gmail.com

© Springer India 2015

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_76

1 Introduction

Stem cells [1] have remarkable potential i.e. atypical characteristics to develop into many different cell types in the body during growth of early life. Sometimes, the unrestricted growth of stem cells or the abnormal growth of cell tissue causes cancer. Stem cells are distinguished from other cell types by two important characteristics. First, they are unspecialized cells capable of renewing themselves through cell division or sometimes after long periods of inactivity. Second, under certain physiologic or experimental conditions, they can be induced to become tissue or organ-specific cells with special functions. Researchers primarily worked with two kinds of stem cells: embryonic stem cells and non-embryonic "somatic" or "adult" stem cells. The induced pluripotent stem cells (iPSCs) [2–4] are some specialized adult cells which "reprogrammed" genetically to assume a stem cell like state in special condition. This property is unique property of stem cell it can be explored by gene expression analysis. Microarray gene expression data to predict and analysis of cancer disease becomes very important. These data can be characterized by genome variables and with their corresponding observations (experiments) in a experimental limitations [5]. To discover co-regulated genes, analysis of gene expression data [6] is required. Previously, it has been assumed that similar patterns in gene expression profiles usually suggest relationships between the genes. Now it is proved genes targeted by same transcription factors, tend to show similar expression patterns along time. Analyzing expression profiles of genes, targeted by same transcription factors revealed complex relationships between co-regulated gene pairs and it also includes co-expression relationships. In this article, we developed a simple algorithm to find differentially co-expressed and co-regulated genes, and then, to construct pairwise co-expression network. We applied the algorithm on sample gene expression microarray data of normal stem cells (*nscr*), stem cells derive tumors cells (*scdtr*) and patch tumor cells (*ptr*).

2 Materials and Methods

For finding and analyzing the relationship between differentially co-expressed genes (DCEG), many techniques have been developed in the literature. Here, we applied statistical t-test [7, 8] and Benjamini Hochberg method [9] to identify differentially expressed genes (DEG) [10] within sample groups of *nscr* and *scdtr*, *nscr* and *ptr* (http://www.ncbi.nlm.nih.gov/GSE20948). In addition to this, we have studied the significance of up-regulated and down-regulated genes. We code the algorithms with using Matlab. The main steps of proposed algorithms are discussed in the following subsections.

2.1 Preprocessing of the Dataset

Our sample Bon Marrow microarray gene expression data is in normalized form. Normalized data for each gene is typically known as an ‘expression ratio’ or as the logarithm of expression ratio. We have done data preprocessing with filtered out low expressed values and null values from the datasets.

2.2 Identification of Differentially Co-expressed Genes

In this section, we discuss the steps to identify DEG. First, a standard statistical t-test is performed for detecting significant changes between measurements of genes in sample microarray groups. It may occurs two types of errors, 1: a false positive by declaring that a gene is differentially expressed when it is not, and 2: a false negative when the test fails to identify a truly differentially expressed gene. Second, Benjamini Hochberg method [9] is used for choosing significantly differentially expressed genes. It is done by the following equation:

$$P \leq y * \frac{\alpha}{m}, \tag{1}$$

where P is the largest p-value called significant, y is the number of genes called significant and m is the total number of genes tested, α is false discovery rate (FDR) [11, 12], defined as the expected ratio of the number of false positives to the total number of positive calls in a differential expression analysis between two sample groups [11]. FDR can be measured [11] as,

$$FDR = Err\left[\frac{F}{F+T}\right] = Err\left[\frac{n_0 \cdot [1 - specificity]}{n_0 \cdot [1 - specificity] + n_1 \cdot sensitivity}\right], \tag{2}$$

where F is the number of false positives, T is the number of true positives, and S is the total number of features called significant. Also, n, number of p-values is seen more clearly, n_0 is the number of truly null features in the study, and $n_1 = n - n_0$ is the number of truly alternative features. Regardless of whether the p-value threshold is fixed or data-dependent, the quantities F, T and S are random variables. Therefore, it is common statistical practice to write the overall error measure in terms of an expected value, which we denote by Err [11].

2.3 Proposed Algorithm

Input: phenotype1: gene expression values of normal stem cell (*nscr*), phenotype2: gene expression values of stem cell derive tumor (*scdtr*), phenotype3: gene expression values of patched tumor (*ptr*).

Output: All differentially expressed genes.
Step1: Perform two-sample t-test to evaluate differential expression of genes from phenotype1 and phenotype2. Their p-values and t-scores are stored.
Step2: Perform a permutation t-test to compute the p-values of 10,000 permutations by permuting the columns of the gene expression data matrix of phenotype1 and phenotype2.
Step3: Determine the number of genes considered to have statistical significance at the p-value 'cutoff' of 0.001.
Step4: Estimate FDR for the genes with statistically significant p-values.
Step5: Create a scatter plot of gene expression data, plotting significance versus fold change of gene expression ratios of phenotype1 and phenotype2.

2.4 Separation of Up-regulated and Down-regulated Genes

We plot the volcano plot of identified all differentially expressed genes as volcano plot (Fig. 4) of two phenotypes returns a structure containing information for genes that are considered to be both statistically significant and significantly differentially expressed. This information helps us to identify co-regulated genes, specially up-regulated and down-regulated genes. Now based on cutoff value (0.001), up-regulated and down-regulated genes are separated from the total set of differentially co-expressed genes.

2.5 Extraction of Co-expressed and Co-regulated Genes

To separate co-expressed and co-regulated genes from the identified DEG, we generate a algorithm and code the algorithm with using Matlab. DEGs which have the same p-value are called co-expressed genes. Up-regulated and down-regulated genes with similar p-values are known as co-expressed and co-regulated genes. Our proposed algorithm is as follows:

Algorithm

Input: Exp: differentially expressed genes or up-regulated or down-regulated genes.
Output: All Co-expressed and Co-regulated genes.
Step1: Each distinct p-value of Exp and its respective position are stored in separate data vector and index vector.

Step2: These data vector p-values are compared with the p-values of the original data, and if they match then their corresponding information is extracted from Exp with the help of the index vectors.

Step3: The process is repeated until all distinct data vector p-values are compared.

2.6 Visualization of Differentially Co-expression and Co-regulation Network

We constructed pair-wise co-expression network to understand the relationship among co-expressed and co-regulated genes. Co-expression networks are built depending on their corresponding p-values. We have used cytoscape software to visualize the pair-wise network. The genes having same p-value construct a paired structure.

3 Results and Discussion

The Bon Marrow microarray data consists of 45,101 genes with having normal stem cell tissue (*nscr*) responses (5no.), stem cell derived tumor tissue (*scdtr*) responses (5no.) and patched tumor cell tissue (*ptr*) responses (4no.) are taken for the study. After analysis, we considered only the genes having p-value $\preceq$ 0.001 as significant genes (Figs. 1 and 2). Out of 45,101 genes from the dataset, 2,325 genes for sample first group (*nscr–scdtr*), 2,056 genes for sample second group (*nscr–ptr*) of data have been extracted on the basis of p-value which is approximately 5 % of total number of genes. Now, estimated false discovery rate (FDR) for statistically significant p-values are computed by using Eq. 2. As a result, we get 1,110, 792

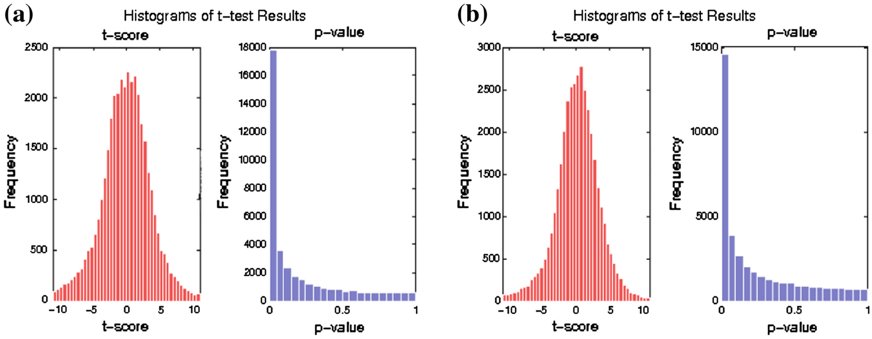


Fig. 1 **a** Histogram plot of sample group *nscr* and *scdtr*. **b** Histogram plot of sample group *nscr* and *ptr*

up-regulated and 1,215, 1,264 down-regulated genes for two sample groups (Figs. 3 and 4) respectively. It implies that 2–3 % genes of total 45,101 genes are up-regulated and down-regulated.

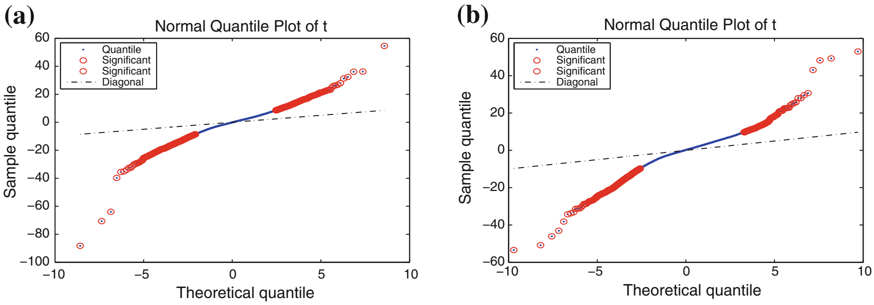


Fig. 2 **a** Quantile plot of sample group *nscr* and *scdtr*. **b** Quantile plot of sample group *nscr* and *ptr*

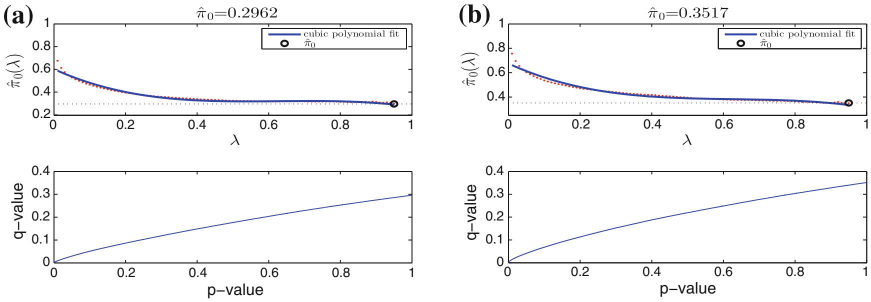


Fig. 3 **a** Fold change plot of sample group *nscr* and *scdtr*. **b** Fold change plot of sample group *nscr* and *ptr*

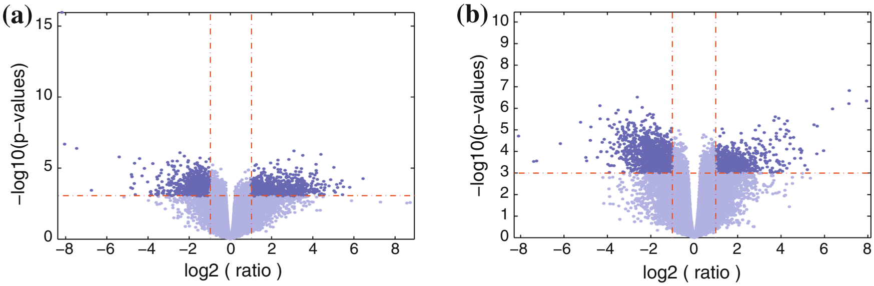


Fig. 4 **a** Volcano plot of sample group *nscr* and *scdtr*. **b** Volcano plot of sample group *nscr* and *ptr*

Further, with finding similar *p*-values co-expressed and co-regulated genes are extracted from all differentially co-expressed genes (DCEG), up-regulated (UR) and down-regulated (DR) genes. We get 8, 4 numbers of up-regulated and down-regulated genes for group1 i.e. approximately 0.01 and 0.008 % of total 45,401 genes. Again we get 6, 14 numbers of up-regulated and down-regulated genes for group2 i.e. approximately 0.01 and 0.03 % of the total 45,101 genes. It is shown in the Table 1. The interaction among co-expressed and co-regulated genes depending on corresponding *p*-values, 6 paired networks are developed (Figs. 5 and 6).

Table 1 The table shows resulting differentially expressed and co-expressed genes of all sample groups

Samples	DEG	DCEG	UR-DEG	UR-DCEG	DR-DEG	DR-DCEG
nscr and *scdtr*	2,326	20 pairs	1,110	4 pairs	1,215	2 pairs
nscr and *ptr*	2,056	21 pairs	792	3 pairs	1,264	7 pairs

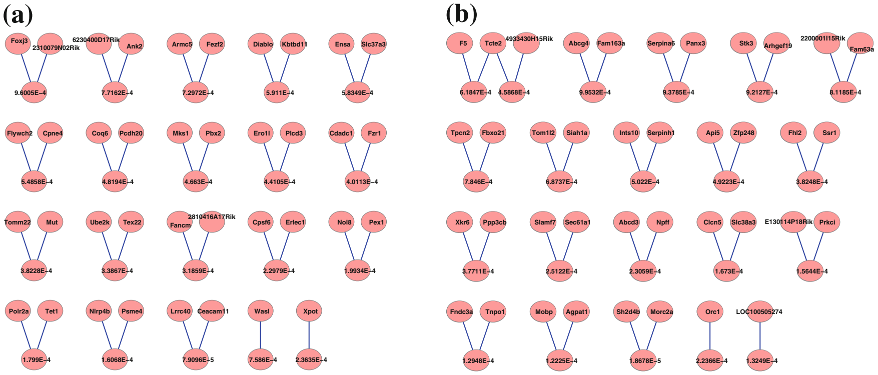


Fig. 5 **a** Pairwise network of differentially co-expressed genes of sample group *nscr* and *scdtr*. **b** Pairwise network of differentially co-expressed genes of sample group *nscr* and *ptr*

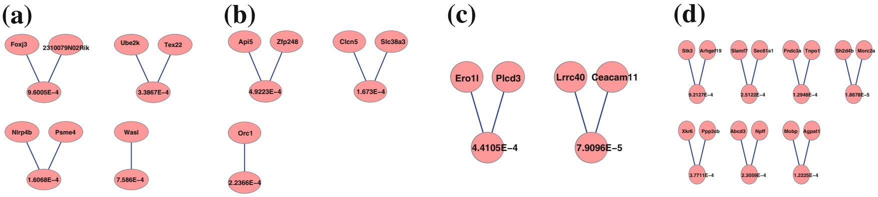
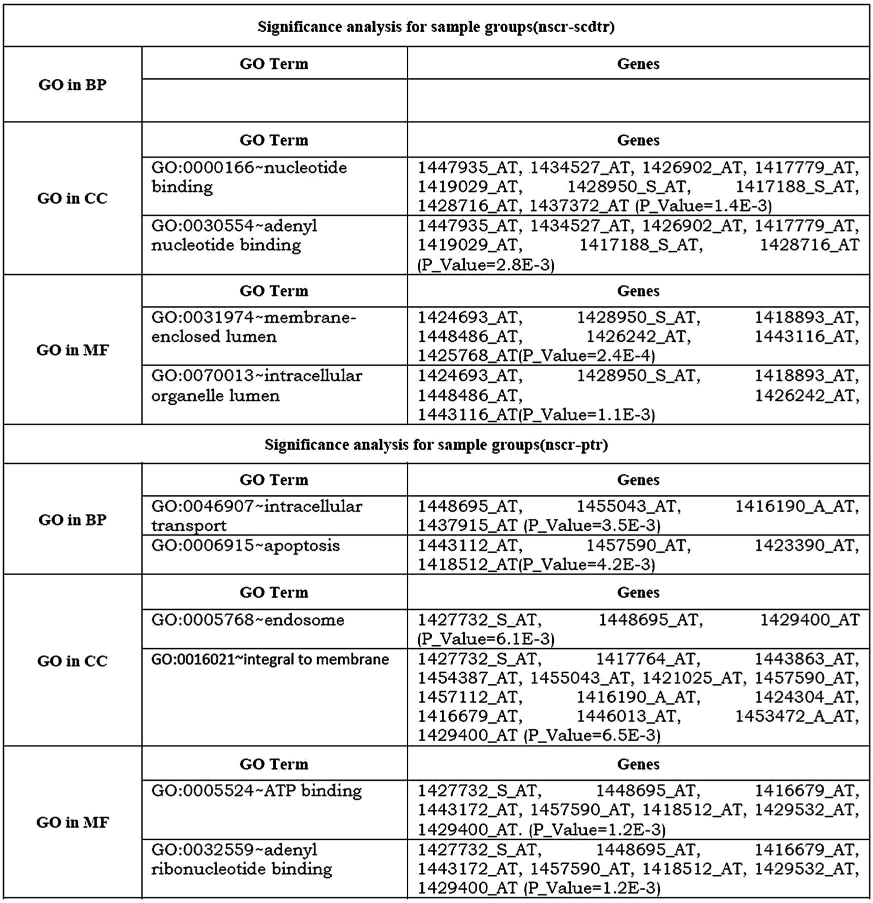


Fig. 6 **a** Pairwise network of differentially co-expressed up-regulated genes of sample group *nscr* and *scdtr*. **b** Pairwise network of differentially co-expressed up-regulated genes of sample group *nscr* and *ptr*. **c** Pairwise network of differentially co-expressed down-regulated genes of sample group *nscr* and *scdtr*. **d** Pairwise network of differentially co-expressed down-regulated genes of sample group *nscr* and *ptr*



Significance analysis for sample groups(nscr-scdtr)		
GO in BP	GO Term	Genes
GO in CC	GO Term	Genes
	GO:0000166~nucleotide binding	1447935_AT, 1434527_AT, 1426902_AT, 1417779_AT, 1419029_AT, 1428950_S_AT, 1417188_S_AT, 1428716_AT, 1437372_AT (P_Value=1.4E-3)
	GO:0030554~adenyl nucleotide binding	1447935_AT, 1434527_AT, 1426902_AT, 1417779_AT, 1419029_AT, 1417188_S_AT, 1428716_AT (P_Value=2.8E-3)
GO in MF	GO Term	Genes
	GO:0031974~membrane-enclosed lumen	1424693_AT, 1428950_S_AT, 1418893_AT, 1448486_AT, 1426242_AT, 1443116_AT, 1425768_AT(P_Value=2.4E-4)
	GO:0070013~intracellular organelle lumen	1424693_AT, 1428950_S_AT, 1418893_AT, 1448486_AT, 1426242_AT, 1443116_AT(P_Value=1.1E-3)
Significance analysis for sample groups(nscr-ptr)		
GO in BP	GO Term	Genes
	GO:0046907~intracellular transport	1448695_AT, 1455043_AT, 1416190_A_AT, 1437915_AT (P_Value=3.5E-3)
	GO:0006915~apoptosis	1443112_AT, 1457590_AT, 1423390_AT, 1418512_AT(P_Value=4.2E-3)
GO in CC	GO Term	Genes
	GO:0005768~endosome	1427732_S_AT, 1448695_AT, 1429400_AT (P_Value=6.1E-3)
	GO:0016021~integral to membrane	1427732_S_AT, 1417764_AT, 1443863_AT, 1454387_AT, 1455043_AT, 1421025_AT, 1457590_AT, 1457112_AT, 1416190_A_AT, 1424304_AT, 1416679_AT, 1446013_AT, 1453472_A_AT, 1429400_AT (P_Value=6.5E-3)
GO in MF	GO Term	Genes
	GO:0005524~ATP binding	1427732_S_AT, 1448695_AT, 1416679_AT, 1443172_AT, 1457590_AT, 1418512_AT, 1429532_AT, 1429400_AT. (P_Value=1.2E-3)
	GO:0032559~adenyl ribonucleotide binding	1427732_S_AT, 1448695_AT, 1416679_AT, 1443172_AT, 1457590_AT, 1418512_AT, 1429532_AT, 1429400_AT (P_Value=1.2E-3)

Fig. 7 Significant genes analyzed with gene ontology

The significance of the identified DEG are studied and analyzed with Gene Ontology (GO). We have listed it in Fig. 7.

4 Conclusion

In this paper, we first found DEG and then found correlations between gene-pairs for construction of co-expressed and co-regulated networks under diseased conditions that assist the interpretability of network. We also generated pairwise differentially co-expression network and constructed the same for differentially co-expressed and co-regulated genes of Bone Marrow stem cell microarray data. We also analyze the significance of DEG for the same microarray data. From our sample data we found 82 significantly co-expressed genes and 30 co-expressed and co-regulated genes. In future study, we can apply artificial intelligence based sophisticated techniques (fuzzy logic, neural networks, evolutionary computation) for better construction of cancer-specific regulatory networks.

References

1. William, et al, J.B.: Functional recovery of spinal cord injury following application of intralesional bone marrow mononuclear cells embedded in polymer scaffold-two year follow-up in a canine. J. Stem Cell Res. Ther. **1**(3) 2011
2. Hanna, J., et al.: Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. Science **318**(5858), 1920–1923 (2007)
3. Saha, K., Jaenisch, R.: Technical challenges in using human induced pluripotent stem cells to model disease. Cell Stem Cell **5**(6), 584–595 (2009)
4. Jiang, Y., et al.: Pluripotency of mesenchymal stem cells derived from adult marrow. Nature **418**(6893), 41–49 (2002)
5. Vaishali, K., Vinayababu, A.: Application of microarray technology and soft computing in cancer biology: a review. Int. J. Biometr. Bioinform. **5**(4), 225–233 (2011)
6. Chen, J.J.: Key aspects of analyzing microarray gene-expression data. Pharmacogenomics **8** (5), 473–482 (2007)
7. Allen, et al. J.D.: Comparing statistical methods for constructing large scale gene networks. PLoS ONE **7**(1) (2012)
8. Vardhanabhuti, S., et al.: A comparison of statistical tests for detecting differential expression using affymetrix oligonucleotide microarrays. OMICS **10**(4), 555–566 (2006)
9. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. **57**(1), 289–300 (1995)
10. Bandyopadhyay, S., Mallik, S., Mukhopadhyay, A.: A survey and comparative study of statistical tests for identifying differential expression from microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(1), 95–115 (2014)
11. Storey, J.D., Tibshirani, R.: Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. **100**(16), 9440–9445 (2003)
12. Storey, J.D., Taylor, J.E., Siegmund, D.: Strong control conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. Roy. Stat. Soc. **66**(1), 187–205 (2004)