Application of a Novel Triclustering Method (δ-TRIMAX) to Mine 3D Gene Expression Data of Breast Cancer Cells

Anirban Bhar[1], Martin Haubrock[1], Anirban Mukhopadhyay[2] and Edgar Wingender[1*]

1. Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany

2. Department of Computer Science and Engineering, University of Kalyani, Kalyani, India

1* edgar.wingender@bioinf.med.uni-goettingen.de

Abstract: We have proposed a novel triclustering algorithm δ-TRIMAX to mine 3D gene expression data sets by introducing a mean squared residue (MSR) score as a measure of coherence of the resultant triclusters. Applying our proposed algorithm on a time series gene expression dataset from an estrogen induced breast cancer cell, we identified key drivers for each resultant tricluster and found a number of hub genes that are known to be associated with breast cancer or estrogen responsive elements. Additionally, our coregulation analysis reveals synergistic regulatory effects of transcription factors.

1 Introduction

With the advent of microarray and other high-throughput technologies, it is feasible to measure expression profiles of thousands of genes across a set of samples and a set of time points. Exploratory approaches facilitate to analyze such high-throughput datasets and thus help to understand the phenotype of a cell. Coexpression analysis is instrumental in identifying genes that exhibit similar expression profiles in molecular networks. Highly interconnected genes in such lists of coexpressed genes are often called hub genes, the analysis of which may reveal underlying disease mechanisms. Clustering algorithms are useful to extract groups of genes or samples having similar expression profiles over all samples or genes, respectively. However, genes are not necessarily similarly expressed over all samples. To find local patterns in two-dimensional gene expression datasets, biclustering algorithms are used. However, to detect groups of genes that are coexpressed over a subset of samples during a subset of time points, triclustering algorithms are required. Attempts to apply biclustering approaches to higher dimensional data would result in a disrupture of the time-dependent structure [SSW+07] and in an inappropriate amalgamation of the different dimensions, requiring extra efforts for postprocessing of the results. In a recent work we have proposed one triclustering algorithm δ-TRIMAX that aims to find triclusters from such 3D gene expression datasets [BHM+13]. We have delineated the coherence of a tricluster by introducing a novel measurement, called mean squared residue (MSR) score; each resultant tricluster must have an MSR score below a threshold δ [BHM+13]. In this work we have applied δ-TRIMAX on a time series gene expression dataset from an estrogen-induced breast cancer cell line to apprehend the underlying disease mechanisms, regulatory effects of transcription factors etc. Additionally, we have compared the capability of δ-TRIMAX with that of an existing triclustering algorithm using an artificial dataset and a real life dataset.

2 Method

Suppose D (G × C × T) represents a 3D gene expression dataset containing G, C and T number of genes, samples and time points, respectively. M(I, J, K) is a tricluster where I $\subseteq$ G, J $\subseteq$ C and K $\subseteq$ T. We define Mean Squared Residue (MSR) to estimate the quality of a tricluster, i.e. the level of coherence among the elements of a tricluster as follows [BHM+13]: $\mathbf{MSR} = \frac{1}{|I||J||K|} \sum_{i \in I, j \in J, k \in K} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2$, where each element of the dataset is m_{ijk} and m_{iJK}, m_{IjK}, m_{IJk} correspond to the mean expression value of ith gene, jth sample, kth time point, respectively. m_{IJK} represents the mean over all genes, samples and time points. For further details of the method and for a description of the whole workflow, we refer to the original publication [BHM+13].

3 Results and Conclusions

For validation of our algorithm, we have applied it first on a synthetic dataset with implanted triclusters and different levels of noise. Comparing δ-TRIMAX with another algorithm, we found that δ-TRIMAX was more reliable in re-identifying the artificial triclusters. We have then applied our algorithm on a time series gene expression dataset from estrogen-induced breast cancer cells to understand the underlying mechanisms of transcriptional regulation during different stages of estrogen response [CMS$^+$06]. We have compared the performance of our algorithm with that of an existing algorithm in terms of coverage, statistical difference from background (SDB) and triclustering quality index (TQI) using the real life dataset [BHM$^+$13]. We could demonstrate that δ-TRIMAX outperforms the existing algorithm according to each of these criteria. To assess the biological significance of genes belonging to each resultant tricluster we performed Gene Ontology biological process (GOBP) and KEGG pathway enrichment analysis. We have observed GOBP enrichment for genes belonging to each tricluster. We used the singular value decomposition (SVD) method to represent each tricluster by its eigen gene. Then we detected hub genes for the co-expression network of each resultant tricluster by calculating Pearson correlation coefficients between eigen gene and expression values of each gene of a tricluster over the samples and time points that are present in that tricluster. The genes (more specifically, the probeset IDs) were sorted in descending order of correlation coefficients. From the 10 topmost probeset IDs, hub genes of each tricluster were identified. This way, we have identified *NPC1L1, TMEM161B-AS1, POU5F1P3, POU5F1P4, POU5F1B, CCL2* as those hub-genes that are coexpressed over all-time points and samples. The chemokine *CCL2* has already been reported to play a role in breast cancer development [TCW$^+$12]. Isoforms or pseudogenes of the transcription factor POU5F1 / OCT4, in particular POU5F1P4, have been found to play specific roles in other types of cancer [WGZ$^+$13], while our results suggest for three of them that they are involved in breast cancer as well. The intestinal cholesterol absorption protein *NPC1L1* has already been inferred to be a target of liver X receptors (LXR) which play an instrumental role in breast carcinogenesis [VDI$^+$04,DTT$^+$06]. Additionally, a previous study infers that estrogen plays an important role in the upregulation of *NPC1L1* [VCR07]. It is already known that the intestinal cholesterol absorption can be used as a drug target for reducing the plasma cholesterol level below the threshold where it promotes the development of tumors and aggravates their aggressiveness [LDM$^+$11, TD03]. Thus we can hypothesize *NPC1L1* as a potential drug target to prevent the growth of breast tumors. We have identified such key drivers for other triclusters as well and found that many of those hub genes are already reported to be associated with breast cancer or estrogen responsive elements. Moreover we performed TFBS enrichment analysis to identify statistically enriched transcriptional regulatory elements in the promoter regions of coexpressed genes using the TRANS-

FAC library (version 2009.4). From this analysis potential coregulation of the coexpressed genes could be inferred. The TFBS found to be enriched also suggested synergistic regulatory effects of transcription factors such as CREB, ATF3, Sp1 etc., which are already known to play crucial roles in breast cancer. We thus feel that our triclustering approach is very suitable to provide biologically meaningful hypotheses, in the example shown about the development of breast cancer.

References

[BHM$^+$13] A Bhar, M Haubrock, A Mukhopadhyay, U Maulik, S Bandyopadhyay, and E Wingender. Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. *Algorithms for molecular biology*, 8(1), March 23 2013.

[CMS$^+$06] J S Carroll, C A Meyer, J Song, W Li, T R Geistlinger, J Eeckhoute, A S Brodsky, E K Keeton, K C Fertuck, G F Hall, Q Wang, S Bekiranov, V Sementchenko, E A Fox, P A Silver, T R Gingeras, X S Liu, and M Brown. Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics*, 38(11):1289–1297, November 2006.

[DTT$^+$06] C Duval, V Touche, A Tailleux, J C Fruchart, C Fievet, V Clavey, B Staels, and S Lestavel. Niemann-Pick C1 like 1 gene expression is down-regulated by LXR activators in the intestine. *Biochemical and Biophysical Research Communications*, 340(4):1259–1263, February 24 2006.

[LDM$^+$11] G Llaverias, C Danilo, I Mercier, K Daumer, F Capozza, T M Williams, F Sotgia, M P Lisanti, and P G Frank. Role of cholesterol in the development and progression of breast cancer. *The American Journal of Pathology*, 178(1):402–412, January 2011.

[SSW$^+$07] J Supper, M Strauch, D Wanke, K Harter, and A Zell. EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics*, 8(334), September 12 2007.

[TCW$^+$12] A Tsuyada, A Chow, J Wu, G Somlo, P Chu, S Loera, T Luu, A X Li, X Wu, W Ye, S Chen, W Zhou, Y Yu, Y Z Wang, X Ren, H Li, P Scherle, Y Kuroki, and S E Wang. CCL2 mediates cross-talk between cancer cells and stromal fibroblasts that regulates breast cancer stem cells. *Cancer Research*, 72(11):2768–79, June 1 2012.

[TD03] S D Turley and J M Dietschy. The intestinal absorption of biliary and dietary cholesterol as a drug target for lowering the plasma cholesterol level. *Preventive Cardiology*, 6(1):29–33, Winter 2003.

[VCR07] M A Valasek, S L Clarke, and J J Repa. Fenofibrate reduces intestinal cholesterol absorption via PPARα-dependent modulation of NPC1L1 expression in mouse. *Journal of Lipid Research*, 48(12):2725–2735, December 2007.

[VDI$^+$04] D M Vigushin, Y Dong, L Inman, N Peyvandi, J P Alao, C Sun, S Ali, E J Niesor, C L Bentzen, and R C Coombes. The nuclear oxysterol receptor LXRalpha is expressed in the normal human breast and in breast cancer. *Medical Oncology*, 21(2):123–131, 2004.

[WGZ$^+$13] L Wang, Z Y Guo, R Zhang, B Xin, R Chen, J Zhao, T Wang, W H Wen, L T Jia, L B Yao, and A G Yang. Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. *Carcinogenesis*, 00(00), May 23 2013.