Discovering Predictive Gene Signatures in BRCA Patients via a Novel Hierarchical Bayesian Framework

Eduardo Ramirez[1], Lifan Liang, BS, MS[2] and Songjian Lu, PhD[2]

[1]Archbishop Stepinac High School, White Plains, New York, 10605, United States

[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15206, United States

Abstract:

Pathway level comprehension plays a key role in precision oncology. Although there have been multitudinous technological advancements in recent years, little is known about the signaling pathways that promote cancer development. In this study, we applied a novel hierarchical Bayesian Framework on breast invasive carcinoma (BRCA) data from the TCGA database to find predictive gene signatures and then compared the results to an artificial neural network (ANN). Through subsequent gene ontology (GO) analysis, we discovered multiple signaling pathways and identified one perturbed pathway associated with poor clinical outcomes to be enriched in genes associated with metastatic cancer. Such identification of perturbed gene signatures consisting of less commonly studied somatic genome alterations (SGAs) (e.g., ERCC6L, TPX2, KIF24) could potentially lead to greater studies on these SGAs and lead to novel treatments in personalized medicine.

Introduction:

Cancer is known to be commonly caused by a variety of SGA's, such as somatic mutations, somatic copy number alterations, and epigenomic alterations(Knudson, 2002). Driver SGAs, when present in important cell activity pathways, such as cell proliferation, commonly lead to the development of cancer. Thus, the importance of cell signaling pathways ushers a better understanding of cancer development.

Further analysis of tumor samples, has led to the conclusion of the heterogeneity in cancer. For example, Sørlie (2001) has discovered the subclasses that lie within breast cancer. Through the emergence of High-throughput DNA sequencing data, subsequent research revealed that somatic genome alterations (SGAs) activate additional signaling pathways in the genome (Hanahan & Weinberg, 2011). Thus, the hallmarks of cancer have been examined and explained through such mutations as seen in Figure 1. There have been numerous ways to discover clusters of alterations or gene signatures, such as the use of statistical differential gene analysis tools (Gentleman et al., 2004). Although these techniques have helped with the development of targeted anti-cancer therapies, such therapeutics only serve subpopulations of those with the same type of cancer. For instance, lung cancer patients acquire a resistance to gefitinib or erlotinib due to the EGFR mutation in the genome (Pao et al., 2005).

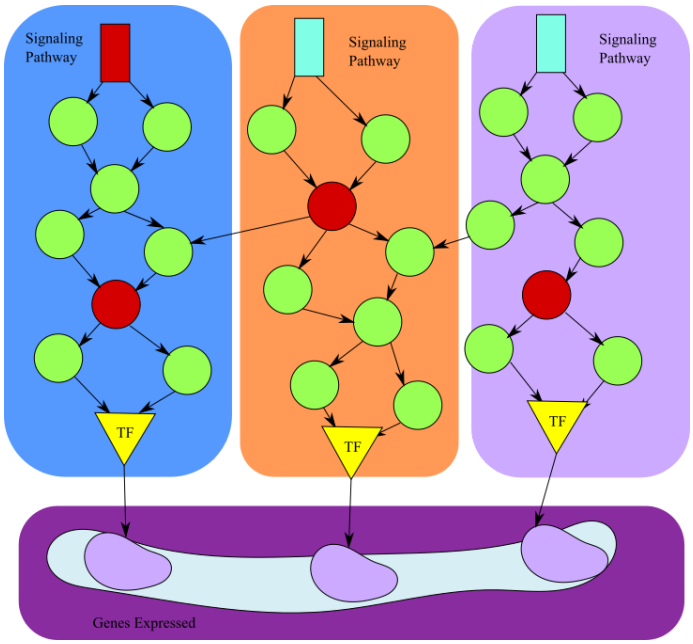

Figure 1. ***Diagram of cellular mechanisms within a cancer cell.*** *A signaling pathway consists of a receptor (rectangle), intracellular nodes (circles), and a transcription factor (triangle). Where if a perturbation occurs (red), the regulated genes are differentially expressed, and multiple signaling pathways can be activated due to co-expression at the pathway level.*

Recently there have been innumerable new frameworks created to address such a problem; however, most studies handle the problem through assumptions; in specific, feature selection, which assumes the transcriptomic factor. Even so, newer frameworks have been deployed to solve such an NP-hard problem through low-rank factor matrices which minimize the problem into one of the latent factors (Liang et al., 2020; Lu et al., 2015). The significance is that these frameworks can now be used to discover gene signatures in cancer patients without assuming transcriptomic factors from previous studies. This can potentially lead to the uncovering of novel and over-looked genes in cancer.

Boolean Matrix Factorization via Expectation-Maximization (BEM), a newly deployed novel hierarchical Bayesian framework, solves Boolean matrix factorization (BMF) in bioinformatics problem in a way which best suits gene-causing diseases(Liang et al., 2020). Previously deployed BMF frameworks (Neumann, n.d.; Ravanbakhsh et al., n.d.) have achieved robustness and efficiency through uniform latent factor shapes; however, in cancer, such is not an efficient way in discovering cancer-causing gene signatures. Instead, BEM is free of assumptions while retaining the robustness of previously deployed BMF frameworks.

In this study, we used BEM to further investigate The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) BRCA data to gain further insights regarding cancer-causing genome alterations. We aim to compare BEM to an artificial neural network (ANN) to see whether its robustness and efficiency is retained when compared to another technique.

Methods:

Data Preprocessing

In this study, we used BRCA patient data from TCGA. This data set consisted of 1075 tumor samples and 114 normal samples. Each transcriptomic profile comprised of 19,665 gene expression values where the values were log2(+1) normalized. Thus, we inversed it and got the original expression value. From such, we created a differentially expressed gene (DEG) matrix by calculating the Z score for each gene in the tumor transcriptomic profiles by:

$$Z_x = (E_x - \mu_x)/\sigma_x$$

where E_x is the value of the gene expression value of gene x in the tumor profile, μ_x is the mean of gene *x* in normal transcriptomic profiles, and σ_x is the standard deviation of gene x in the normal transcriptomic profiles. From there, each gene was determined to be differentially expressed based on two criteria: whether (1) the Z score value was greater than or equal to 1.645, corresponding to the one-sided normal P-value of 0.05; (2) the fold change between x and the mean of normal samples was greater than to 2. If a gene matched both criteria, they were given a 1, if not a 0.

Model Evaluation

Latent factors were used as a way to simulate transcriptomic factors found in signaling pathways. After data preprocessing, the number of latent factors were determined through BEM's Akaike in-formation criterion (AIC) function, where a latent factor size of 15 had the minimum output.

In order to extract the genes in each signature, we first extracted the second edges, or pathway, and co-expressed differential levels matrix where the absolute value of a gene's edge weight determined their significance. The absolute values of the matrix were ranked, and the 99th percentile was used as the cut-off value; only the values greater than it were kept.

Correlation with the signature and the clinical outcomes was based on the first edge or patient-pathway expression matrix. Therefore, we used all patients which exhibited an edge weight value greater than 0.5 in the first edge, where we then correlated it with each patient's disease-specific survival (DSS) and progression-free interval (PFI).

Analysis of each gene signature was through GO enrichment analysis, where we could note the function of each signature, and conclude what each perturbed signature meant on the pathway level.

Results:

BEM Discovers Gene Signatures of Greater Significance

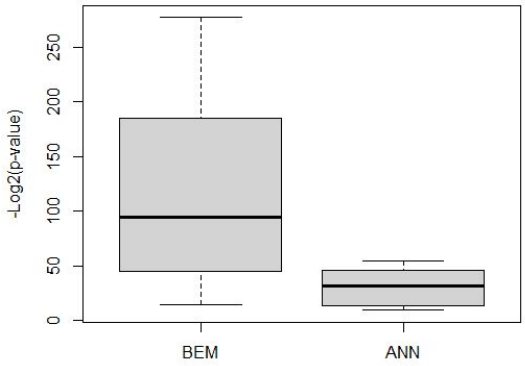

Figure 2. Comparing the GO enrichment analysis p-values of the most significant gene signatures of each ML technique.

In order to analyze BEM's performance, we similarly constructed an ANN to compare performances. In the ANN, we used the latent factor size as in BEM, and also used ReLU as the activation function; in order to get a value between 0 and 1, such as in BEM.

BEM's performance can be noted when analyzing the p-values of the enriched GO terms of each technique (Figure 2.). BEM cannot solely be concluded to discover enriched terms of greater significance, but also minimize redundancy (Figure 3.). The ANN does not get rid of redundant terms, however, BEM does, in order to push for sparsity.

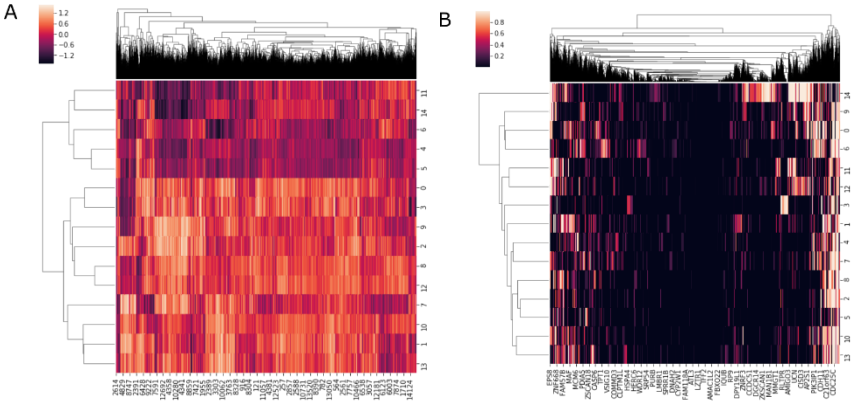

Figure 3. Heatmap representation of Second edge matrix, or the relationship matrix between the perturbed pathways and co-expressed differential levels, ANN (3A), BEM (3B). A large amount of redundancy can be noted in the ANN.

BEM Discovers Gene Signatures Closely Related to Clinical Outcomes

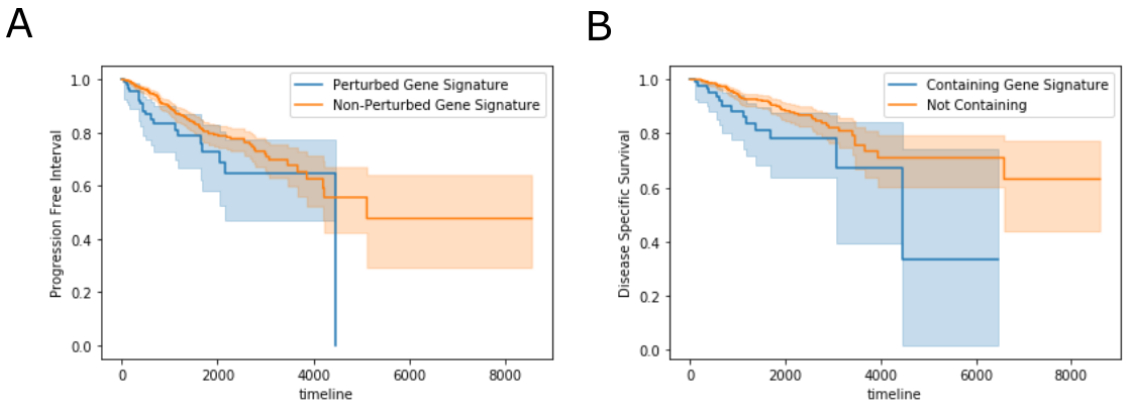

Figure 4. Patients with the perturbed gene signature discovered in our 8th latent factor have worse overall survival compared to others. Figure 4A shows the progression-free interval probability in days, and Figure 4B shows the disease-specific survival in days.

BEM's robustness can be noted in analyzing the gene signatures discovered. We discovered a gene signature in our 8th latent pathway which leads to sizable patient decline ~4100-day mark (Figure 4.). Through GO analysis, this can be noted since the signature is enriched in cytoskeleton and microtubule genes.

Further analysis of the gene signature led to the realization that our signature was enriched in metastatic genes. Fife (2014) discovered that the cell cytoskeleton played a significant part in lymphatic metastatic cancer due to the fact that the microtubule structures play an emerging role in helping to mediate cell movement; thus, helping better understand the poor outcomes in the patients who exhibit the signature.

BEM's Signatures are Enriched in Genes Significant in BRCA

Further analysis of the gene signature noted in our 8th latent factor can be characterized to consist of masses of genes that are under-studied in BRCA. Genes such as ERCC6L, TPX2, KIF24 are often overlooked in BRCA, however, recent studies prove that these genes serve as prognostic biomarkers and a potential future for being therapeutic targets (Jiang et al., 2019; Li et al., 2020; Sebastian-Leon et al., 2014).

However, other gene signatures discovered through BEM also consist of commonly known BRCA-causing genes, such as BRCA1, BRCA2, CHEK2, BARD1. We can conclude that these genes were not part of the signature associated with the poorest clinical outcomes due to the immense number of studies that exist on such genes. Instead, those with the poorest outcomes suffer due to the under-studied mutations.

Discussion and Future Directions:

In this study, we present the use of a new novel hierarchical bayesian framework, BEM, in discovering gene signatures within RnaSeq data. In order to analyze its robustness and efficiency, we compared it to a traditional ANN. Moreover, through statistical analysis of enriched GO terms, we can conclude that BEM provides much more efficient and robust findings than the traditional ANN.

Such findings in BEM are not solely statistically significant, but also associated with poor clinical outcomes in BRCA patients. Through BEM, we were able to find gene signatures which have a significant outcome in patients, which through GO analysis, came to be due to being enriched in metastatic pathway mutations.

BEM not solely highlights signatures of genes which are commonly known in BRCA, but instead highlights less commonly studied genes in BRCA (e.g., ERCC6L, TPX2, KIF24), meanwhile also highlighting commonly studied genes in BRCA (e.g., BRCA1, BRCA2, CHEK2, BARD1). Through the use of BEM, we can conclude that consecutive use of it in larger data samples, novel targets for treatments can be found.

Subsequent analysis of our signatures most closely related to poor clinical outcomes can lead to findings of whether or not there is a novel gene mutation in our signature which leads to the metastatic movement as exhibited by the enriched GO terms.

Funding:

This project was supported through funds from the University of Pittsburgh Department of Biomedical Informatics and the National Institutes of Health YES Program, Grant Number PAR-17-059.

Acknowledgments:

I would like to thank my mentors, Dr. Songjian Lu, and Lifan Liang for guiding me throughout this project. Without them, this would not have been achievable. I would also like to thank Dr. David Boone and Solomon Livshits for this experience through the UPMC Hillman Cancer Center Academy.

Bibliography:

Fife, C. M., McCarroll, J. A., & Kavallaris, M. (2014). Movers and shakers: Cell cytoskeleton in cancer metastasis. *British Journal of Pharmacology*, *171*(24), 5507–5523. https://doi.org/10.1111/bph.12704

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li,

C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Jiang, Y., Liu, Y., Tan, X., Yu, S., & Luo, J. (2019). TPX2 as a Novel Prognostic Indicator and Promising Therapeutic Target in Triple-negative Breast Cancer. *Clinical Breast Cancer*, *19*(6), 450–455. https://doi.org/10.1016/j.clbc.2019.05.012

Knudson, A. G. (2002). Cancer genetics. *American Journal of Medical Genetics*, *111*(1), 96–102. https://doi.org/10.1002/ajmg.10320

Li, T.-F., Zeng, H.-J., Shan, Z., Ye, R.-Y., Cheang, T.-Y., Zhang, Y.-J., Lu, S.-H., Zhang, Q., Shao, N., & Lin, Y. (2020). Overexpression of kinesin superfamily members as prognostic biomarkers of breast cancer. *Cancer Cell International*, *20*(1), 123. https://doi.org/10.1186/s12935-020-01191-1

Liang, L., Zhu, K., & Lu, S. (2020). BEM: Mining Coregulation Patterns in Transcriptomics via Boolean Matrix Factorization. *Bioinformatics (Oxford, England)*, *36*(13), 4030–4037. https://doi.org/10.1093/bioinformatics/btz977

Lu, S., Lu, K. N., Cheng, S.-Y., Hu, B., Ma, X., Nystrom, N., & Lu, X. (2015). Identifying Driver Genomic Alterations in Cancers by Searching Minimum-Weight, Mutually Exclusive Sets. *PLOS Computational Biology*, *11*(8), e1004257. https://doi.org/10.1371/journal.pcbi.1004257

Neumann, S. (n.d.). *Bipartite Stochastic Block Models with Tiny Clusters*. 11.

Pao, W., Miller, V. A., Politi, K. A., Riely, G. J., Somwar, R., Zakowski, M. F., Kris, M. G., & Varmus, H. (2005). Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLoS Medicine*, *2*(3). https://doi.org/10.1371/journal.pmed.0020073

Ravanbakhsh, S., Póczos, B., & Greiner, R. (n.d.). *Boolean Matrix Factorization and Noisy Completion via Message Passing*. 10.

Sebastian-Leon, P., Vidal, E., Minguez, P., Conesa, A., Tarazona, S., Amadoz, A., Armero, C., Salavert, F., Vidal-Puig, A., Montaner, D., & Dopazo, J. (2014). Understanding disease mechanisms with models of signaling pathway activities. *BMC Systems Biology*, *8*(1), 121. https://doi.org/10.1186/s12918-014-0121-3

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., & Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10869–10874. https://doi.org/10.1073/pnas.191367098

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764