Yang Pan
UID: 504567911

***Running title*: Patient stratification of cancer via integrated genomic and transcriptomic profiles coupling with network-aware biclustering**

1.Background and Significance:

Significance: As a realization of Precise Medicine (PM), Stratified Medicine (SM) has been recently pointed out by Willis and Lord [1] in a recent review in 2015. Cancer, as a complex (in mechanism) and heterogeneous (among patients) disease, can be better understood and cured by gaining insights of the hidden samples structures driven by shared underlying mechanisms. Patient stratification can be considered and solved as an unsupervised clustering problem.

Existing works: Many efforts have been made on patient stratification in cancer studies based on individual mutations[2], combination of mutated genes[3], expression profiles[4], or the combination of several data types[5 6]. In terms of clustering approach, hierarchical clustering (HC) is commonly seen in many cancer studies[7] like TCGA, it is simple to perform but insufficient to capture local similarities among a subset of subjects as it based on global similarities. Other advanced unsupervised methods including probabilities models[5,8], factor analysis[9], SVD[10] and self-organizing map (SOM)[11] are also employed to this problem, however they suffer various limitations like instable result, single data source and so on. After reviewing those methods, the consensus is that a desirable stratification should be comprised of 1) an efficient clustering method, 2) a comprehensive integration of different level of molecular data and 3) a proper use of biological networks for interactions among molecular features. Following this line, a recent deep learning-based work developed a deep belief network (DBN)[12] to stack multiple data types, however it ignores biological networks when constructing abstract features. Another popular method, called iCluster[9], also provides good data integration but ignores the feature correlations like gene-gene network. A successful work based on non-negative matrix factorization (NMF) named NBS provides a network-aware clustering. However, it only uses mutation data which limits the comprehensiveness of stratifying patients, and the stochastic nature of NMF may not be effectively constrained by only the network penalty applied on one of the two decomposition matrices[13], which may further affect consensus clustering results. In another study using network-based NMF, only expression level datasets are integrated[14].

Novelty: To our best knowledge, there is no existing model that provides both integration of mutation and expression data and leveraging network of molecular feature interaction for patient stratification problem. Here we present a NMF-based model, performing patient stratification based on patterns from combination of mutation and gene expression coupling with known gene-gene or protein-protein network. On the top of that, technically we optimized the standard NMF allowing a compact and simultaneous clustering between molecular signatures with patients. This could help us reduce subsampling and may also increase accuracy.

2.Proposed Method:

Outline: Performing NMF with patient-cluster matrix $\boldsymbol{H}$ (row as patients and column as clusters) simultaneously with both cluster-mutation matrix $\boldsymbol{W_1}$and cluster-expression matrix $\boldsymbol{W_2}$ (row as clusters and column as mutation/expression values). Integrating network knowledge as a quadratic form penalty term for each input data type. $\boldsymbol{K_1}$can be adjacency matrix of network derived from Hotnet2, $\boldsymbol{K_2}$ can be a gene-gene network. Additional regularization (smooth term $\boldsymbol{S}$) is added to enhance compact patient and molecular feature as biclusters to enhance the result reproducibility, which reduces the time spent on subsampling and enhances clustering efficiency.

Objective function: First we need to format a matrix $\boldsymbol{V}$, where values in it are mutation and expression measurements. Based on this given matrix $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, we proposed the network-regularized non-smooth NMF model to perform clustering. This is achieved by minimizing the following objective function:

$$\min_{W,H} \|\boldsymbol{V_1} - \boldsymbol{W_1 S H}\|^2 + \|\mathbf{V_2} - \boldsymbol{W_2 S H}\|^2 - \operatorname{trace}\left(\boldsymbol{W_1}^{\boldsymbol{T}} \boldsymbol{K_1 W_1}\right) - \operatorname{trace}\left(\boldsymbol{W_2}^{\boldsymbol{T}} \boldsymbol{K_2 W_2}\right)$$

where basis vectors $\boldsymbol{W} \in \mathbb{R}^{n\times q}$ ($\boldsymbol{W_1}$, $\boldsymbol{W_2}$ are cluster-mutation and cluster- expression matrix respectively) and encoding vectors $\boldsymbol{H} \in \mathbb{R}^{q\times p}$ is patient-cluster matrix which is just defined as standard NMF. Besides network constraints elaborated above, another layer of constrain is $\boldsymbol{S}$ as a smoothing matrix is defined as $\boldsymbol{S} \in \mathbb{R}^{q\times q}$, $\boldsymbol{S} = (1-\theta)\boldsymbol{I} + \frac{\theta}{q}\boldsymbol{11^T}$. θ is used to control the sparseness of the model, ranging between 0 to 1 [13]. This allows the sparseness generated on both $\boldsymbol{W}$ and $\boldsymbol{H}$ as demonstrated by Pascual-Montano et al..Note that the parameter θ cannot be learned from the data and need to be experimented. And the rank q is also needed to be estimated via a consensus clustering method that commonly used for selection of the rank[15].

For optimizing the objective function, we will just follow the standard NMF updating rule[15] with the modification corresponding to added constrains to iterate until convergence. Since the randomized initiation is commonly

adopted by variations of NMF, it is always preferred to repeat NMF multiple times for robustness[16] and we will use subsampling to enhance the clustering reproducibility.

3.Proposed Data Analysis:

First, we construct simulation dataset based on a given network and then analyzing the performance on it. This is to benchmark the proposed method against existing methods including standard NMF, NBS and iCluster.
Second, we perform the analysis on real cancer mutation data from TCGA (for comparison, use the same dataset used by Hofree et al.[17]). This analysis should be done in comparison with standard NMF, NBS and iCluster. We will use survival time as a benchmark to evaluate performance. We will explore mutation-expression signatures for stratified patients, with the hope of being confirmed with known mechanism and discover new ones.

4.Expected Timeline:

The proposed work should be finished following timeline below:
Use 5 days to prepare data and implement the proposed method in python or R.
Use 5 days to perform analysis and debug.
Use 5 days to finish the comparison and evaluation.
Use 10 days to draft manuscript and address any remaining problems in the project.

Reference

1 Willis, J. C. & Lord, G. M. Immune biomarkers: the promises and pitfalls of personalized medicine. Nat Rev Immunol 15, 323-329, doi:10.1038/nri3820 (2015).
2 Dulak, A. M. et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet 45, 478-486, doi:10.1038/ng.2591 (2013).
3 Bertrand, D. et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res 43, e44, doi:10.1093/nar/gku1393 (2015).
4 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330-337, doi:10.1038/nature11252 (2012).
5 Cho, D. Y. & Przytycka, T. M. Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. Nucleic Acids Res 41, 8011-8020, doi:10.1093/nar/gkt577 (2013).
6 Zhang, S. et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 40, 9379-9391, doi:10.1093/nar/gks725 (2012).
7 Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339, doi:10.1038/nature12634 (2013).
8 Khakabimamaghani, S. & Ester, M. Bayesian Biclustering for Patient Stratification. Pac Symp Biocomput 21, 345-356 (2016).
9 Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25, 2906-2912, doi:10.1093/bioinformatics/btp543 (2009).
10 Sun, J., Bi, J. & Kranzler, H. R. Multi-view singular value decomposition for disease subtyping and genetic associations. BMC Genet 15, 73, doi:10.1186/1471-2156-15-73 (2014).
11 Borkowska, E. M. et al. Molecular subtyping of bladder cancer using Kohonen self-organizing maps. Cancer Med 3, 1225-1234, doi:10.1002/cam4.217 (2014).
12 Liang, M., Li, Z., Chen, T. & Zeng, J. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. IEEE/ACM Trans Comput Biol Bioinform 12, 928-937, doi:10.1109/TCBB.2014.2377729 (2015).
13 Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. & Pascual-Marqui, R. D. Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Trans Pattern Anal Mach Intell 28, 403-415, doi:10.1109/TPAMI.2006.60 (2006).
14 Zhang, S., Li, Q., Liu, J. & Zhou, X. J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics 27, i401-409, doi:10.1093/bioinformatics/btr206 (2011).
15 Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101, 4164-4169, doi:10.1073/pnas.0308531101 (2004).
16 Pascual-Montano, A. et al. bioNMF: a versatile tool for non-negative matrix factorization in biology. BMC Bioinformatics 7, 366, doi:10.1186/1471-2105-7-366 (2006).
17 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. Nat Methods 10, 1108-1115, doi:10.1038/nmeth.2651 (2013).