Yang Pan
UID: 504567911

***Running title*: Uncovering hidden patient-gene architectures of cancer via network-guided factorization-based 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], or expression profiles[4]. Hierarchical clustering (HC) is one of the most widely used methods in many cancer studies[5] like TCGA for its simplicity. It is based on global similarities thus fails to capture hidden local structures. Biclustering, on the other hand, promises to find subgroups of driver genes and derived patients. Factorization-based biclustering is considered more 'natural'[6] in contrast to other biclustering methods applying greedy search and imposing constrains on size or number of biclusters. In 2013, Hofree et al.[7] proposed an interesting framework, named network-based stratification (NBS), incorporating known molecular networks with mutational profiles to perform a part-based clustering. Follow-up studies[8] confirmed that their network strategy has enhanced the mutation-based patient stratification. However, there is a remaining issue in their proposed network-regularized NMF, which may cause insufficient patient stratification. According to previous literatures[6,9], constrains added on the vectors of standard NMF will either not promoting sparseness efficiently or decrease the model's goodness of fit. In NBS paper, the network constrains on basis vectors $\boldsymbol{W}$ will cause smoothness on encoding vectors $\boldsymbol{H}$ which will compromise the stratification.

Novelty: Here we proposed a new method combining both a biclustering version of NMF and network-based regularization to take the benefits from network without increasing the ambiguity of stratifying patients and genes into substructures.

2.Proposed Method:

Input: First we need to format a matrix $\boldsymbol{V}$, where values in it can be mutation or expression measurements. If it is mutational profile, then it should be smoothed by network as NBS[7]; if it is expression profile, positive and negative values for each gene should be separated into two rows and use absolute value only[10].

Objective function: 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:

$$\min_{W,H} \|\boldsymbol{V} - \boldsymbol{WSH}\|^2 + \mathrm{trace}(\boldsymbol{W^T K W})$$

where basis vectors $\boldsymbol{W} \in \mathbb{R}^{n\times q}$ and encoding vectors $\boldsymbol{H} \in \mathbb{R}^{q\times p}$ are just defined as standard NMF, $\boldsymbol{S}$ 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 [9]. 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. Another layer of constrain is added as a $\mathrm{trace}(\boldsymbol{W^T K W})$ function contains term $\boldsymbol{K}$ which is derived from an influence matrix used in Hotnet2 or its former version[11]. Thus $\boldsymbol{K}$ can be estimated based on the given network and numbers of neighbors can be adjusted. And the rank q is also needed to be estimated via a consensus clustering method that commonly used for selection of the rank[12].

As we can see from our modified NMF equations, it minimizes the sum of the norm of the residual which serves to give sparseness simultaneously on basis vectors $\boldsymbol{W}$ and encoding vectors $\boldsymbol{H}$ and the quadratic regularizing term which serves to constrain basis vectors $\boldsymbol{W}$ with local network neighborhoods. Then we just follow the standard NMF updating rule[12] 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[10].

Rationale: The rationale of this method described as following. Being inspired by the non-smooth NMF formulated by Pascual-Montano et al.[9], we proposed this approach specifically for stratifying patients via a biclustering version of NMF while also inheriting NBS's strength by taking network information. Thus, our network-regularized non-smooth NMF model has the potential to provide more compact and less ambiguous stratification.

3.Proposed Data Analysis:

First, we construct simulation dataset (use mutation data as a POC) based on a given network and then analyzing the performance on it. This is to benchmark the proposed method against existing method.
Second, we perform the analysis on real cancer mutation data from TCGA (for comparison, use the same dataset used by Hofree et al.[7]). This analysis should be done in comparison with network-regularized NMF and standard NMF.
Third, an optional analysis could be done with same TCGA cancer type but other data type, like CNV or mRNA expression data. This analysis is to explore the potential of this method on other date type and explore association between mutation/expression/copy number/etc. to explore mechanisms shaping different patient subgroups (patients overlapping could be measured by a hypergeometric test)

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.
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.

Yang Pan
UID: 504567911

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 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).

6 Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M. & Pascual-Montano, A. Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. *BMC Bioinformatics* **7**, 78, doi:10.1186/1471-2105-7-78 (2006).

7 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).

8 Chang, Y. *et al.* COPD subtypes identified by network-based clustering of blood gene expression. *Genomics* **107**, 51-58, doi:10.1016/j.ygeno.2016.01.004 (2016).

9 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).

10 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).

11 Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* **18**, 507-522, doi:10.1089/cmb.2010.0265 (2011).

12 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).