Topic Modeling on Purification of Heterogeneous LC/GC-MS based Proteomics and Metabolomics Data

Minkun Wang1,2, Tsung-Heng Tsai1, Cristina Di Poto3, Alessia Ferrarini1, Guoqiang Yu2, Habtom W. Ressom1
1 Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC; 2 Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA

Objective
This study aims to address the data heterogeneity issue in analyzing data generated through liquid (or gas) chromatography coupled with mass spectrometry (LC/GC-MS) in a variety of omic studies, e.g., proteomics and metabolomics. Purification of LC/GC-MS based biomolecular expression profiles is highly desired prior to subsequent analysis, that is, quantitative comparison of the abundance of biomolecules in clinical samples. We applied a topic model to computationally deconvolute samples of LC/GC-MS based cancer expression profiles and infer the underlying sample-specific pure cancer profiles.

Introduction
- Identification of disease-related alterations in molecular and cellular mechanisms helps reveal useful biomarkers for human diseases including cancers.
- High-throughput omic technologies for identifying and quantifying multi-level biological molecules (e.g., proteins, glycans, and metabolites) have facilitated the advancement of biological researches.
- LC/GC-MS allows quantitative comparison of biomolecular abundance in clinical samples to help with the discovery of candidate biomarkers for complex diseases.

Challenges
- Clinical samples collected from patients usually exhibit some degree of heterogeneity.
- The proportion of cancerous, other disease-related, and healthy components varies across individual samples preselected using pathological estimates.
- The cancerous profiles of interest are typically contaminated by other components, leading to unreliable results in differential analyses.
- Experimental methods for cleaning samples and isolating tissue-specific constituents are costly and time-consuming.

Methods
- Data profile \( p_{ij} \) is characterized by a probability distribution across topics.
- Topic/source is probability distribution (normalized intensities) over biomolecules.
- Hierarchical Bayesian model - a variant of latent Dirichlet allocation (LDA)
- Complete likelihood function
- Infer the latent variables \( y_i^*, y_i, \beta, \gamma \) (underlying pure sources and proportions)
- Estimate the hyper-parameters in the model

Experimental Data
1. 116 HCC patients enrolled at MedStar Georgetown University Hospital (MGUH) are diagnosed with liver cirrhosis and 57 of them are developed with HCC.
- 101 proteins were quantified through MRM.

2. 15 GC MS based tissue metabolomics profiles
- 15 liver tissues were collected from 10 participants recruited at MGUH.
- 150 metabolites were identified and quantified after pre-processing the GC-MS raw data.

Conclusion & Future Work
- We apply a topic model based inference method to computationally address heterogeneity in clinical samples analyzed by LC/GC-MS.
- This model gives a probabilistic explanation on the corpus of LC/GC-MS based profiles.
- Simulation demonstrated the model’s capacity of estimating mixture proportion and retrieving underlying pure cancer profile.
- Increased discrimination between case and control groups is observed. More biologically meaningful pathways are found.

Acknowledgements
This work was supported by NIH Grants R01CA143420 and R01GM086746.