Purification of LC/GC-MS based biomolecular expression profiles using a topic model

Kevin Minkun Wang

CBIL, Virginia Tech
Ressom Lab, Georgetown University

Aug. 14th, 2015
Outline

1. Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2. Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
Glycomics: LC-MS profiled N-glycans in sera

Biomarker discovery in hepatocellular carcinoma (HCC) cases vs. liver cirrhosis (CIRR) controls in an Egyptian (TU) and US (GU) cohort.

- **Experimental design:**
  - Analyzed sera in four batches (~24 samples each batch) for each cohort
  - Balanced assignment of HCC and CIRR into each batch in terms of age, race, gender, smoking, alcohol, and BMI.

- **Sample Preparation:**
  - Release → Purification → Reduction → Permethylation of N-glycan

- **Global Profiling:**
  - LC-ESI-MS data acquisition
  - Data preprocessing (de-isotoping, peak detection, alignment and normalization)

- **Targeted Analysis:**
  - Targets include N-glycans (i) detected by the Orbitrap or QqQ instrument in our previous studies, (ii) selected as potential HCC biomarkers in previous studies, and (iii) involved in Golgi apparatus retrieved from KEGG GLYCAN database.
  - LC-ESI-MRM-MS data generation
  - MRM transition curation

- **Statistical Analysis**
  - ANOVA test
  - Fold change consistence

Tsai et al. J Proteome Res. 2014
Glycomics: LC-MS data preprocessing

- Deisotoping → Peak Detection → Alignment → Adducts Clustering → Normalization

1) Trace the deisotoped ions with the same molecular weight (with 10 ppm tolerance) across retention time.
2) Interpolate missing values using corresponding extracted ion chromatograms. Smooth trace by convoluting with a Savitzky-Golay filter.
3) Take a first-order derivative of a Gaussian kernel to identify the position and boundaries.
4) Characterize each peak by monoisotopic mass, charge state, intensity (area under curve within boundary), and retention time.

Wang et al. IEEE BIBM 2013
Glycomics: biomarker discovery

Identification based on tandem mass spectra

Statistical analysis for identified glycan across samples

- Many of the candidate biomarkers for HCC are closely related in their structures. Within each cluster the glycans show consistent changes in their levels.

- Further elucidation of the relationship between the identified complex N-glycans can be obtained by referring to their biosynthesis process.
Background: related projects in Ressom Lab
- Glycomic Study
- Proteomic Study
- Multi-omic Study

Topic Model based Computational Purification
- Motivation
- Probabilistic Modeling: Topic Model
- Simulation
- Experimental Results
Proteomics: LC-MS based global and targeted analysis in serum proteins

- Biomarker discovery in hepatocellular carcinoma (HCC) cases vs. liver cirrhosis (CIRR) controls.

Tsai et al. Proteomics 2015.
286 proteins were identified by global proteomic analysis of the sera samples also used in glycomics study.

101 target proteins were further analyzed by MRM.

61 proteins showed significant changes in HCC vs. CIRR.

52 are potential glycoproteins based on our search for glycosylation sites.
Outline

1. Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2. Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
Multi-omic study: pilot project

★ Analysis of multiple cellular biomolecules in liver tissues

- 15 liver tissues from 10 subjects
  ✓ 5 tumor and 5 adjacent cirrhotic tissues from 5 HCC cases
  ✓ 5 cirrhotic tissues from 5 CIRR controls
- Goal is to identify biomolecules significantly altered in the following two comparisons:
  • Tumor vs. adjacent cirrhotic tissues (5 vs 5)
  • Tumor vs. independent cirrhotic tissues (5 vs. 5)
- Multi-omic analysis of
  ✓ mRNAs by RNAseq
  ✓ metabolites by LC/GC-MS
  ✓ proteins by LC-MS
  ✓ N-glycans by LC-MS
  ✓ N-glycopeptides by LC-MS

★ 15 → 105 tissue samples
Multi-omic study: pilot project

- Prioritize targets based on previous and ongoing untargeted multi-omics studies and literature survey.
- Acquire quantitative information of targeted metabolites, N-glycans, proteins, deglycosylated glycopeptides, intact glycopeptides in liver tissues and sera.
- Perform integrative analysis of multi-omic data.
Outline

1. Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2. Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
Motivations

Sample heterogeneity → biomarker discovery

- Specimens (e.g., tumor tissues) are typically mixtures of cells with distinct states and types, and usually part of the constituents is relevant to the biological question of interest.
- In some cancer studies, heterogeneity is due to the co-existence of multiple cancerous subtypes.
- The proportion of cancerous, other disease-related, and healthy components varies across individual samples preselected using pathological estimates.

Experimental purification methods: costly and time-consuming.

Computational purification methods*: inexpensive and efficient to implement

*(available for data already generated without any modifications on experimental procedures).
Outline

1. Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2. Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
Probabilistic modeling: hypothesize a way to generate the heterogeneous profiles

- **Terminology**
  - \( \{t_d\}_{d=1,\ldots,D} \): expression profile of a heterogeneous sample. [Observed]
  - \( \{\gamma_d\}_{d=1,\ldots,D} \): pure/cancerous origin. [Latent]
  - \( \{\beta_m\}_{m=1,\ldots,M} \): non-cancerous contaminants/unfavorable source. [Observed]
  - \( \{\theta_d\}_{d=1,\ldots,D} \): sample-specific mixture proportion. [Latent]
  - \( \{z_{d,n}\} \): source indicator for each ion in each sample. [Latent]
  - \( \gamma' \): average cancer origin (whole-collection-level). [Latent]
  - \( \alpha, \eta, \kappa \): hyperparameters of Dirichlet priors. [Latent]
  - Mixture of multiple sources: \( \beta, \gamma, \theta \Rightarrow t \)

- **Intuitive explanation**
  - \( t \) is treated as an article with \( N \) words, representing \( N \) measured ions.
  - \( \{\gamma_d\}, \{\beta_m\} \) play a role of underneath "topics" in generating each article in the corpus.
Probabilistic modeling: hypothesize a way to generate the heterogeneous profiles

- Three assumptions in our study
  - \( \{\beta_m\} \rightarrow t_d \): The source contaminants in each expression profile \( \{t_d\} \) are coming from the control group \( \{\beta_m\}, m=1, \ldots, M \). It has been observed that the cancerous tissues within tumor samples are typically surrounded by adjacent non-cancerous tissues.
  - \( \gamma' \rightarrow \gamma_d \): Corresponding cancerous origins \( \{\gamma_d\}, d=1, \ldots, D \) share an average cancer profile \( \gamma' \). Individual cancerous profile can be treated as a noisy version of the average cancer profile in the same group (i.e., HCC group).
  - \( \{\beta_m\} \rightarrow \gamma' \): Average cancer profile has similar patterns as non-cancerous profiles, except for some sites (biomolecules) which are differentially expressed between case and control groups – holds in the same cohort.

- Mathematically, \( \{\beta_m\}, \gamma', \{\gamma_d\} \) represent multinomial (probabilistic) distribution over vocabulary/biomolecules.
Probabilistic modeling: Topic Model

- 3-level generative probabilistic model derived from latent Dirichlet allocation (LDA model)
- (1) Deterministic ↔ Stochastic ; (2) Frequentist ↔ Bayesian
- Inference & Estimation: maximizing complete/joint likelihood function via variational expectation maximization (variational EM) algorithms.

Fig. 1: Graphical representation of the generative probabilistic model: hyperparameters $\eta$, $\kappa'$ together with sources of contaminants $\{\beta_m\}$ determine an average cancer profile $\gamma'$. Each of the $D$ profiles is associated with a mixture proportion $\theta_d$ (regularized by hyperparameter $\alpha$) and a topic panel consisting of $\{\beta_m\}$ and $\gamma'$ (generated from the average cancer profile). Each of the $N$ ions in a profile $t_{n,d}$ is sampled from a topic indicated by $z_{n,d}$.

Wang et al. IEEE BIBM 2015 (under review)
Outline

1. Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2. Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
Purification on LC-MS profiles: simulation

Generate a set of synthetic data by artificially mixing real ‘homogeneous’ LC-MS data.
Estimated mixture proportions $\theta^*$

**Top:** radar charts with 10 spokes, each representing a source in topic panel. The proportion of each source is delimited by the length of lines with color.

**Bottom:** scatter plots of corresponding proportions in $\theta$ and $\theta^*$
Inferred pure cancer profile $\gamma^*$

Scatter plots of unpurified and purified cancer profiles vs. true cancer profiles

Define estimation error ratio for a single sample:

\[
\zeta_d(\theta^*, \theta) = \frac{||\theta^*_d - \theta_d||_1}{||\theta_d||_1} \times 100\%, \quad d = 1, \ldots, 30
\]

- \(\zeta_d(\theta^*, \theta) = 2.33\%
- \(\zeta_d(\gamma^*, \gamma) = 6.51\% < \zeta_d(t, \gamma) = 16.57\% \)
Outline

1 Background: related projects in Ressom Lab
   - Glycomic Study
   - Proteomic Study
   - Multi-omic Study

2 Topic Model based Computational Purification
   - Motivation
   - Probabilistic Modeling: Topic Model
   - Simulation
   - Experimental Results
116 LC-MS based serum proteomic profiles
- All 116 patients recruited at MedStar Georgetown University Hospital (MGUH) were diagnosed with liver cirrhosis and 57 of them developed with HCC.
- It is not clear how the development of tumor in liver directly affect the alterations in blood.
- We hypothesize that there are some impacts from cirrhotic constituents contributing to the HCC profile in serum. The contamination may occur in an indirect way.
- 101 proteins were quantified through LC-MRM-MS.

15 GC-MS based tissue metabolomic profiles
- 15 liver tissues were collected from 10 participants recruited at MGUH.
- 559 metabolites were identified and quantified after preprocessing the GC-MS raw data.
**Purification on LC/GC-MS profiles: experimental data**

- **LC-MS proteomics**

![Plot showing principal components analysis](image)

Left panel: ROC curves for each of 43 significant proteins before purification (AUC = 0.706, 95%CI [0.606, 0.795]). Right panel: ROC curves for each of 75 significant proteins after purification (AUC = 0.793, 95%CI [0.700, 0.863]).

**TABLE I: Signaling Pathways (number of biomarkers hit)**

<table>
<thead>
<tr>
<th>Without Purification</th>
<th>With Purification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complement and coagulation cascades (13)</td>
<td>Complement and coagulation cascades (18)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus (5)</td>
<td>Systemic lupus erythematosus (6)</td>
</tr>
<tr>
<td>Prion diseases (4)</td>
<td>PPAR signaling pathway (5)</td>
</tr>
<tr>
<td>-</td>
<td>Prion diseases (4)</td>
</tr>
</tbody>
</table>

- **GC-MS metabolomics** (1) Treat independent cirrhotic profiles as contaminants of HCC profiles: $0 \rightarrow 7$ (FDR adjusted $p$-value $\leq 0.05$); (2) Treat HCC profiles as contaminants of adjacent cirrhotic profiles: $\tilde{\zeta}(\psi, \beta) = 28.3\% \rightarrow \tilde{\zeta}(\psi^*, \beta) = 24.9\%$
We have applied a **topic model** based inference method to computationally address heterogeneity issue 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 was observed. More biologically meaningful pathways were found.

**Ongoing/Future work**

- Find appropriate forms of regularization on parameters to address the limitation due to small sample size.
- Consider label information to endow prediction function to the model.
- Apply to clustering of subtype diseases.
Acknowledgment

http://omics.georgetown.edu

Welcome To Ressom Lab

Thank you for your attention and food!