Overview

This study aims to develop an integrated pipeline for analysis of liquid chromatography-mass spectrometry (LC-MS) based glycomic data. The proposed pipeline consists of a collection of algorithms to extract meaningful features and cluster the features that represent the same glycan. This helps to summarize a list of glycans comprising unique natural masses annotated by putative glycan compositions.

Introduction

- Glycomics helps investigate the role that glycosylation plays in complex diseases.
- LC-MS allows us to evaluate the levels of N-glycans released from proteins in biological samples to identify candidate biomarkers.
- However, individual N-glycans can generate a set of ion species with various charge and adduct states, which impedes the interpretation of LC-MS data.
- The proposed pipeline integrates peak detection and feature clustering to generate a list of unique neutral masses representing putative glycan compositions.
- We demonstrate the performance of the pipeline in detecting features and recognizing ions derived from the same glycan through analysis of LC-MS data from a serum biomarker discovery study.

Peak Detection

- Deisotoped [DeconTools]: Each data point is ion sum with mass value, charge state and intensity
- Ion tracing and interpolation: ion list → trace list
- Iteratively find trap deisotoped [DeconTools] with highest intensity
- Record for mass (m), scan number and charge
- Based on the desired precision, define the mass range (m=Δm, m=2Δm)
- Link ions within the mass range in adjacent scans

Feature Clustering

- Clustering charge states: split peak list into blocks, collecting peaks with close retention time
- Determine maximum boundary of peaks within each block, extract their EICs respectively from raw data (contain profiling shape information)
- Create connected graph for each block using pairwise Pearson correlation coefficients of EICs as edge weights and peaks as nodes
- Apply the Highly-Connected-Subgraphs algorithm to cluster peaks with high similarities
- Re-group unclustered peaks for a second round clustering

Methods

The proposed pipeline consists of two major components: peak detection and feature clustering, both of which incorporate chromatographic profiles (extracted ion chromatogram, EIC) as additional information to address analytical and computational uncertainties.

Peak Detection

Deisotoped [DeconTools]
- Each data point is ion sum with mass value, charge state and intensity
- Ion tracing and interpolation: ion list → trace list
- Iteratively find trap deisotoped [DeconTools] with highest intensity
- Record for mass (m), scan number and charge
- Based on the desired precision, define the mass range (m=Δm, m=2Δm)
- Link ions within the mass range in adjacent scans

Feature Clustering

Clustering charge states:
- Split peak list into blocks, collecting peaks with close retention time (RT)
- Determine maximum boundary of peaks within each block, extract their EICs respectively from raw data (contain profiling shape information)
- Create connected graph for each block using pairwise Pearson correlation coefficients of EICs as edge weights and peaks as nodes
- Apply the Highly-Connected-Subgraphs algorithm to cluster peaks with high similarities
- Re-group unclustered peaks for a second round clustering

Experimetnal Results

- From each LC-MS run, DeconTools deisotoped ~150000 peaks detected
- 2000 good-quality traces generated
- ~3500 peaks detected
- ~80 clusters had putative glycan compositions.

Table II: the list the results of annotation done within one cluster. Among 24 peaks, 22 had been assigned to one of the three glycans and annotated by different charge and adduct states. Proton-adducted peak was selected and matched to composition (e.g., G1→HexNAc_HexHexNAc.)

Summary

- The proposed pipeline characterizes LC-MS based glycomic data through computational elucidation and integration of chromatographic profiles of co-eluted features.
- Incorporating profiling shape information gives high confidence in grouping features.
- The redundant and more informative peak list is generated, facilitating downstream matching across runs and statistical work.

Future Work

- Extend the pipeline by incorporating models for retention time alignment and normalization of peak intensities.
- Utilize tandem MS to improve the accuracy of annotation.

Acknowledgements

This work was supported by NIH Grants R01CA143420 and R01GM086746.