Feature Decomposition from LC-MS Data using Non-negative Matrix Factorization

Yong Fugui Li1, Randy J. Arnold2, Predrag Radivojac1,4, Haixu Tang1,3
1School of Informatics and Computing, 2Department of Chemistry, 3Center for Genomics and Bioinformatics and 4Department of Statistics, Indiana University, Bloomington, IN 47405

OVERVIEW

- Peak intensity information is commonly used in MS based quantification studies.
- Feature extraction from LC-MS spectra is non-trivial due to the presence of excessive noise, contaminants and overlapping signals.
- We propose DECIPHER, a probabilistic unsupervised feature extraction method.
- DECIPHER is closely related to the recently developed NMF (Non-negative Matrix Factorization) method for pattern recognition.

MOTIVATION

- Inherent regularities of LC-MS signals
LC and MS are orthogonal by design. Our aim here is to extract features use (only) this property.
What is a feature?
A feature does not necessarily come from a peptide. It may correspond to a non-peptide bio-molecule, chemical contaminant, systematic noise, or a peptide (possibly with multiple isotope states).

FULL BAYESIAN FORMULATION

- DECIPHER model
  Given a feature (\(z\)), the elution profile (\(t\)) and m/z measurement (\(m\)) are independent.
  \[
  P(X,W,S,H|\alpha, \beta, \gamma) = P(X|WSH^T) \cdot P(W|\beta) \cdot P(H|\gamma) \cdot (S|\alpha)
  \]
  \[
  W = (W_{z}) = (P(t_{z}); H = (H_{mz}) = (P(m_{z}); S = \text{diag}(S_{z}) = \text{diag}(P(z))
  \]
  Maximum likelihood – equivalent to NMF
  Bayesian – selects the number of features

RESULTS – REAL DATA

- Data: yeast N15 labeling, analyzed by LTQ-FT
- Model: DECIPHER-NPB, with Gaussian error, with or without Gaussian priors, ALS updating algorithm

RESULTS – SIMULATED DATA

- Data: randomly generated features with Gaussian elution profiles, isotope-distribution-convoluted Gaussian m/z peaks and random noise
- Model: DECIPHER-NPB, with Gaussian error, ALS updating algorithm

DISCUSSIONS

We present a general machine learning method for feature extraction in LC-MS. Encoded with barely any prior knowledge, DECIPHER-NPB generates surprisingly meaningful features. More accurate feature extraction can be achieved by the parametric version of the model. Our approach may serve as a preprocessing step for labeled/label-free quantitative proteomics

REFERENCES

2. A.T. Cemgil, Computational Intelligence and Neuroscience, 2009.

This project was funded by the NIH grant R01 RR024236-01A1 (PI: Radivojac) and NCI grant U24 CA126480-01 (PI: Regnier).