Deconvolution of Overlapped Peptide Isotopic Peak Clusters with EM Algorithm for Label-free Quantification

Lei Xin[1], M. Ziaur Rahman[1], Weiwu Chen[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] University of Waterloo, Waterloo, ON
Bioinformatics Solutions Inc.

Introduction

With the development of high resolution mass spectrometry and liquid chromatographic (LC) separation techniques in recent years, the label-free quantitative method has greatly improved its sensitivity and accuracy for globally profiling changes in protein abundances in biological systems. However, even with today's high resolution instruments and LC separation techniques, for complex mixtures overlapped peptide isotopic clusters cannot be avoided. An example of overlapped peptide isotopic clusters are shown in Figure 1. This kind of overlapping will hamper the quantification accuracy of the label-free method. Expectation maximization (EM) [1] is a statistical algorithm which is used to deal with mixed distributions. In this poster, we present an EM based deconvolution algorithm for the overlapped peptide isotopic clusters.

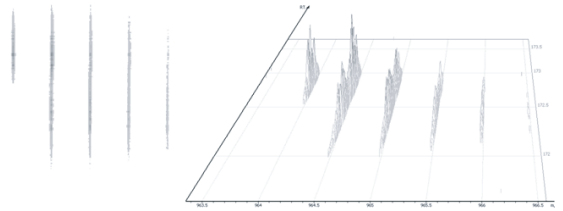

Figure 1. Overlapped Peptide Isotopic Clusters

Method

The observed mixed distribution is usually composed of two or more component distributions, as shown in Figure 2. The EM algorithm was used to auto-fit the component distributions for an observed mixed distribution. Several overlapped peptide isotopic clusters were treated as a mixed distribution. Then we used the EM algorithm to auto-fit its component isotopic cluster. The algorithm is described below:

1. Detect all local maximum points on the LC-MS view.
2. Initialize a distribution model for each local maximum point and each possible charge. Each model represents a component isotopic cluster.
3. Use EM iteration to auto-fit the distribution for each component isotopic cluster.
4. Remove insignificant distribution models.
5. Perform EM iteration again for all remaining distribution models.
6. Report the position, duration and intensity for each component isotopic cluster.

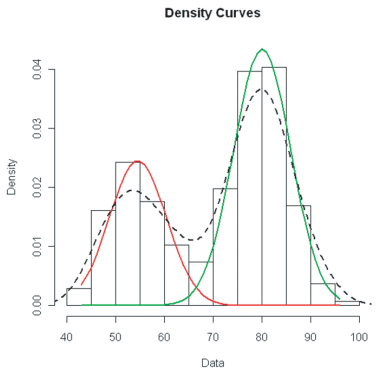

Figure 2. Mixed Distribution.
The dotted line represents the observed mixed distribution while red and green lines represent component distributions.

Results

We tested the EM based deconvolution algorithm on two data sets. The first data set is from samples which were obtained after treating human cells in culture with a drug or a control solution (DMSO) for 1hr or for 4hrs. Cells were lysed and the resulting proteins were reduced and alkylated on cysteine residues, before being digested with trypsin. Phosphopeptides were enriched using a titanium dioxide approach. Every sample was done in triplicate at the biological level. In total 12 files are generated after being analyzed by LC coupled with a Thermo LTQ-Orbitrap instrument. One file from the control sample was picked to test the method proposed here. This file contained 4537 MS scans and 3252 MS/MS scans. After the file was analyzed by the EM based algorithm, 9368 peptide isotopic clusters were reported. Among them 1090 (11.6%) clusters were overlapped with other clusters. The second data set was a public data set from yeast samples [2]. Yeast cells were lysed and the resulting proteins were reduced and alkylated on cysteine residues, before being digested with LysC. 6 files were generated after the sample was analyzed by LC coupled with a Thermo Q Exactive instrument. One file with 35268 MS scans and 68128 MS/MS scans was analyzed by the EM based algorithm. 106245 peptide isotopic clusters were reported. Among them 16583 (15.6%) clusters were overlapped with other clusters. Figure 3 to Figure 5 (red outline shows component clusters) present examples of detected overlapping peptide isotopic clusters.

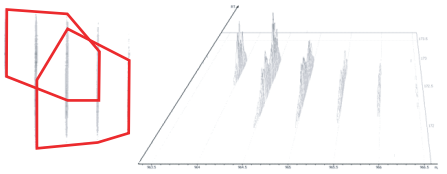

Figure 3. Deconvolution of Two Overlapped Peptide Isotopic Clusters

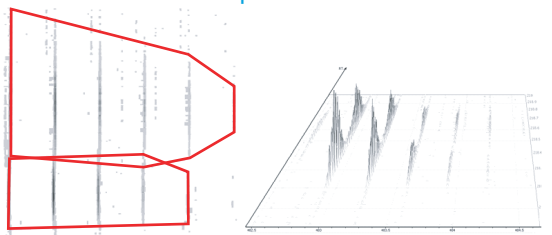

Figure 4. Deconvolution of Two Overlapped Peptide Isotopic Clusters

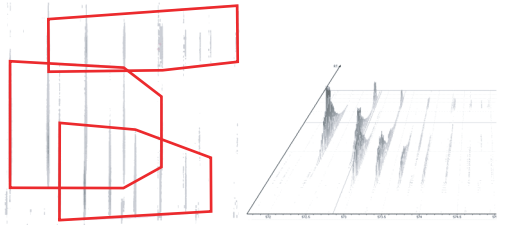

Figure 5. Deconvolution of Three Overlapped Peptide Isotopic Clusters

Conclusion

Our EM based statistical method can effectively detect overlapped peptide isotopic clusters and deconvolute them into their component peptide isotopic clusters.

References

[1]. A.Dempster, N.Laird, D.Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1-38,1977

[2]. N.Nagaraj, N.A.Kulak, J.Cox, N.Neuhauser, K.Mayr,O. Hoerning, O. Vorm, M.Mann. System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. Mol Cell Proteomics. 2012 March; 11(3): M111.013722.