PEAKS - A Software Tool for Shotgun Label Free Proteomics with High Sensitivity and High Accuracy

Lei Xin[1], Hao Lin[1], M. Ziaur Rahman[1], Weiwu Chen[1], Baozhen Shan[1], Bin Ma[2]

[1] Bioinformatics Solutions Inc, Waterloo, ON, Canada, [2] University of Waterloo, Waterloo, ON, Canada

Overview

Label free shotgun proteomics has been used for protein identification and quantification, which are the two most fundamental applications in proteomics. For system-wide application, the sensitivity and accuracy are of the challenge in label-free shotgun proteomics due to the broad dynamic range of protein abundance and stochastic identification of peptides between samples. For intensity-based approach, the information for quantification (MS1) and identification (MS2) is measured in a single run. The deconvolution of overlapped peptide features and retention alignment between runs are the key factors for the data analysis, because the overlapped peptide feature clusters cannot be avoided even with today's high resolution MS instruments and LC separation techniques. Here we present a software tool, PEAKS Q, for accurate label-free quantification with high sensitivity.

Methods

To deconvolute overlapped peptide features, an EM algorithm was used to auto-fit a distribution model for each peptide feature as a component in the presented isotopic clusters.

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, which represents a component in the cluster.
3. Use EM iteration to auto-fit the distribution for each component.

To align retention time among sets of runs, a maximum weighted matching algorithm was used [1].

1. The file of a run which shares most features with the files of rest runs was chosen as the reference.
2. Given two sets of features, the retention time alignment algorithm is based on an optimization model, which works on feature matching and retention time alignment simultaneously.

Figure 1. Deconvolution of Two and Three Overlapped Peptide Isotopic Clusters

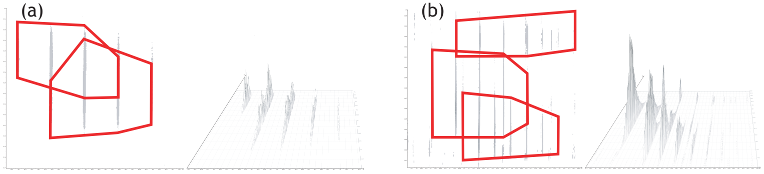
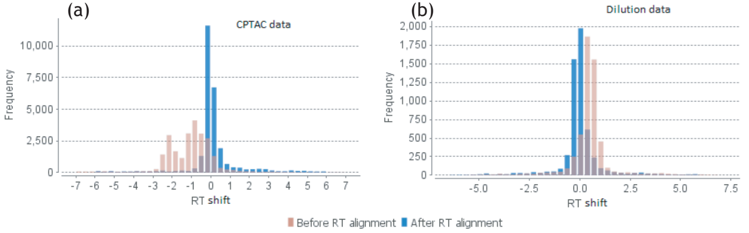


Figure 2. Distribution of Retention Time Shift

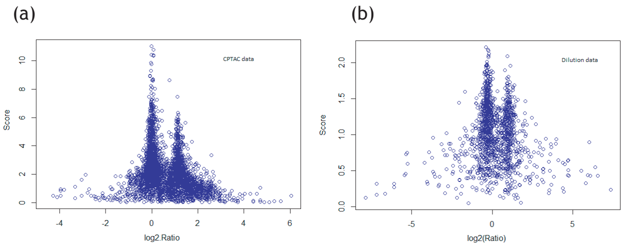


Experimental

To validate our label-free quantification tool, two types of data sets were used to compare the quantitative results to a ground truth:

1. Spike-in data from CPTAC study [2] data 6: Four datasets from Orbitrap instruments (Orbi, OrbiO, OrbiP, and OrbiW) were used. In each dataset, sample A, B, C, D, and E were yeast spiked with a mixture of 48 proteins (UPS1) at 0.25, 0.74, 2.2, 6.7, and 20 fmol/μL respectively. The sample was analyzed in triplicate.
2. Dilution data from [3]: On the basis of the TICs, the lysate of a bacterium Streptococcus pyogenes was mixed with a lysate of human cells in a dilution series with the ratios 0/100, 20/80, 40/60, 60/40, 80/20, and 100/0%. Each sample was measured a single time on an LTQ Orbitrap instrument.

Figure 3. Distribution of Peptide Feature Ratios

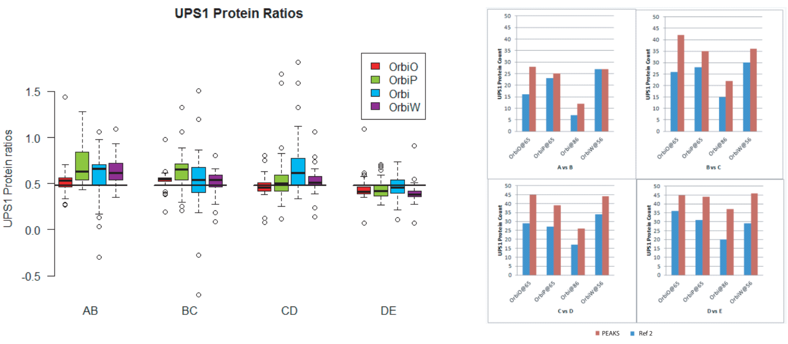


The datasets were analyzed with PEAKS 7. CPTAC data were processed using a database combining Yeast Uniport database and UPS proteins and typical laboratory contaminant proteins, containing 6716 entries in total. Dilution data were processed with Uniprot database containing 526969 entries. Peptides were identified with PEAKS DB and filtered at 1% of FDR. Proteins were filtered at 1 minimum unique peptide.

Results

1. Feature detection Even with high resolution LC separation techniques and mass spectrometer instruments, the overlapped peptide isotopic clusters were commonly observed (Figure 1). 13.2% and 15.6% clusters were overlapped for CPTAC data and Dilution data.
2. Retention time alignment The sample which share most features with rest samples is chosen automatically as the reference sample. All rest samples carry on retention time alignment with this reference sample. Given two lists of feature pairs, the retention time was aligned with a clear optimization model, which works on feature matching and retention time alignment simultaneously. The algorithm is efficient and accurate, even with significant time shifts and distortions (Figure 2). After alignment, features from different runs were mapped into feature vectors/groups. A confidence score was associated with each feature vector.

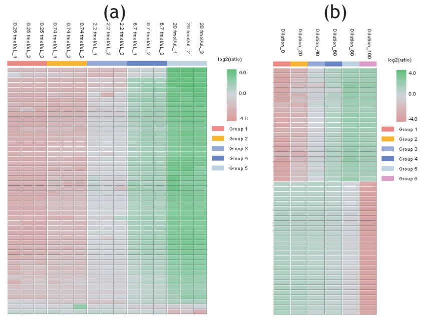
Figure 4. PEAKS Quantified Proteins with High Accuracy and High Sensitivity.
(a) Boxplot of the Log Ratios of UPS1 Proteins (b) The Count of Quantified Proteins



The distributions of intensity ratios of peptides between two runs were shown in Figure 3. Two clusters were observed corresponding to two groups of proteins in the sample.

3. Protein ratio estimation The protein abundances were estimated in each sample by correlation the average of the feature intensities of the three most highly responding peptides per protein. The estimation of spiked-in protein ratios in CPTAC data were shown in Figure 4. PEAKS showed high accuracy of protein ratio estimation (4a) and the number of UPS proteins in the top 50 quantified proteins (4b). The summary of the first quantified proteins was shown in Figure 5.

Figure 5. Summary of Protein Quantification



Conclusions

PEAKS estimates protein ratios for label-free quantification with high accuracy and high sensitivity.

References

1. H. Lin et al. A Combinatorial Approach to the Peptide Feature Matching Problem for Label-Free Quantification, Bioinformatics, 2013, 10.1093
2. Y. Chen et. al., IDPQuantify: Combining Precursor Intensity with Spectral Counts for Protein and Peptide Quantification. Journal of Proteomics Research, 2013, 12(9): 4111-4121.
3. H. Weisser et.al., An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. Journal of Proteomics Research, 2013, 12(4): 1628-1644.