New Algorithm for Label-Free Protein Quantification

Weiwu Chen[1]; Baozhen Shan[1]; Jing Zhang[1]; Eric Bonneil[2]; Janine Voyer[1]; Gilles Lajoie[3]; Pierre Thibault[2]; Bin Ma[4]

[1]Bioinformatics Solutions Inc., Waterloo, ON. [2]University of Montreal, Montreal, QC.
[3]University of Western Ontario, London, ON. [4]University of Waterloo, Waterloo, ON.

Introduction:

Label free quantitative proteomics analysis is a flexible approach enabling the profiling of protein expression across different datasets. The success of this approach relies not only on the efficient detection of peptides over a wide range of ion abundance but also on the capability of correlating their precise coordinates in different LC-MS runs. Several approaches have been previously studied to achieve these goals including the use of normalized LC retention time for data acquired on high resolution mass spectrometry instruments [1,2]. PEAKS™ Q offers this new algorithm as its approach for label-free quantification. We report a new approach termed "feature vector" that analyzes multiple samples simultaneously to increase the accuracy of feature detection and the protein coverage.

Method:

Our quantification workflow is illustrated in Figure 1. The input is several LC-MS/MS datasets from multiple samples. The data analysis steps are summarized in the following:

Retention time correction. Retention time (RT) recorded in the data are normalized to correct the variation between different LC runs. This is done with a new dynamic programming algorithm to maximize the overall similarities between the MS scans with similar normalized RT in different data sets.

Feature detection. By using the isotope distribution and the peak shape similarities between adjacent MS scans, the peptide features (MS peaks that possibly formed by peptides) are selected.

Feature vector formation. Features from different samples defined by the same mass and similar normalized retention time form a feature vector. The concept of feature vector allows the features from different samples to cross-validate each other and therefore false positive features are removed.

Protein identification. PEAKS™ Studio 5.1 is used to do protein identification by combining all of the data sets together. The combining of all datasets together has a potential advantage of increasing protein identification confidence and coverage. For example, when two peptides of the same protein are identified separately in two different samples, running protein identification in any of the samples alone may result into a low-scoring protein that risks being dropped by the protein identification algorithm. Combining all the datasets together can avoid this pitfall. The excellent scalability of PEAKS™ Studio 5.1 is essential for such analysis because the combined dataset is usually large (over 100GB for some of our data).

Another important task in this step is the protein clustering. Normally a family of similar proteins can be identified as long as one protein presents in the sample. They need to be clustered together to avoid false positives which complicated the ratio calculation step.

Ratio calculation. A feature vector containing at least one identified peptide (from any of the input samples) determines the ratio of the peptide. For each protein cluster, the peptide features unique to this cluster are used to calculate the protein ratio. Outliers are removed before the peptide ratios are averaged together to compute the protein ratio.

The J774 mouse proteome was digested with trypsin and measured with a hybrid LTQ-Orbitrap mass spectrometer in a nanoLC-MS/MS experiment. In addition, the tryptic peptides from six standard proteins, DHE3_BOVIN, PERL_BOVIN, ALDOA_RABIT, GLPK_ECOLI, CAS1_BOVIN, and ALBU_BOVIN, were spiked at nine concentration levels of 0-50 fmol/uL into the J774 protein digest. This gives nine samples, each should have the same concentrations of the J774 proteins and different concentrations of the six spiked protein. The resulting nine datasets were analyzed with our method. In addition, the above experiment was conducted with 3 replicates to study the reproducibility.

Result:

Approximately 1000 features were detected in each sample, 85% of which were consistently observed across all other samples and corresponded to tryptic peptides from J774 mouse macrophage proteins. By grouping these features into feature vectors, about half of the vectors contained at least one peptide identified by protein database search. A small fraction of the feature vectors containing different peptide identification results were removed. The others were used to compute the peptide ratios and then protein ratios. By selecting one of the 9 samples as the standard, Table 1 shows the spiked and calculated protein ratios in the other 8 samples relative to the standard. We note that most calculated protein ratios are well conformed with the spiked protein ratios.

Figure 2 shows the relationship between the reproducibility of the ratio calculation on spiked proteins and the spiked concentration. Most spiked proteins give low relative standard deviation (RSD). The three larger ones are because of the low spiked concentration. When a peptide has extremely low concentration, on one replicate it may be detected with correct ratio; but on the other replicates it may not be detected, giving a ratio 0. This causes the increased RSD.

Figure 3 shows the distribution of the calculated ratio for J774 peptides. The J774 peptides are supposed to have ratio 1 if their corresponding peaks are identical in all of the nine samples (meaning the LC-MS experiments are fully reproducible). The figure shows that, our software reported ratios between 1/1.5 and 1.5 for most (93.6%) of the J774 peptides.

Figure 1. Workflow of the quantification method.

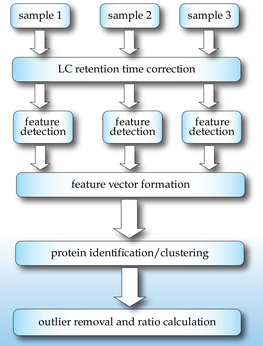


Table 1. Spiked ratios and calculated protein ratios in 8 samples relative to the standard sample.

		S1	S2	S3	S4	S5	S6	S7	S8
DHE3_BOVIN	spiked ratio	0	20.5	9.8	4	2.1	0.5	0.2	0
	calculated ratio	≤0.1	≥10	≥10	5.79	2.49	0.57	0.24	≤0.1
PERL_BOVIN	spiked ratio	0	20.5	9.8	4	2.1	0.5	0.2	0
	calculated ratio	≤0.1	≥10	≥10	7.23	2.67	0.66	0.29	≤0.1
ALDOA_RABIT	spiked ratio	0	1	1	1	1	1	1	1
	calculated ratio	≤0.1	0.96	1.04	1.13	1.09	1.08	0.99	1.34
GLPK_ECOLI	spiked ratio	0	1	1	1	1	1	1	1
	calculated ratio	≤0.1	0.73	0.85	0.96	0.89	1.02	0.91	0.78
CAS1_BOVIN	spiked ratio	0	0	0.01	0.25	0.5	2	5.1	10
	calculated ratio	≤0.1	≤0.1	≤0.1	≤0.1	0.21	2.2	6.51	≥10
ALBU_BOVIN	spiked ratio	0	0	0.01	0.25	0.5	2	5.1	10
	calculated ratio	≤0.1	≤0.1	≤0.1	0.22	0.44	2.5	6.3	≥10

Figure 2. The relationship between the reproducibility of ratio calculation and the absolute protein concentration. Each data point represents one spiked protein in one sample. RSD is calculated across the three replicates. Y-axis is the spiked concentration. Most spiked proteins have low RSD while the larger errors are all due to the very low absolute concentration of the spiked protein.

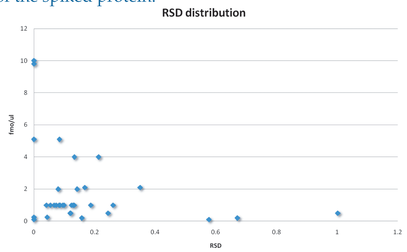
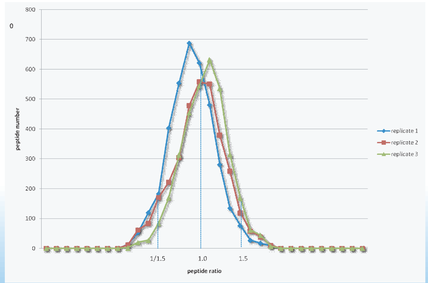


Figure 3. The ratio distribution for the J774 peptides. Y-axis is the percentage of peptides at the given ratio.



Software:

PEAKS™ Q label-free quantification also provides an interactive graphical user interface for users to examine each peptide feature vector included in the ratio calculation. Figure 4 shows a 3D view of the peptide feature vector in 9 samples (with no feature detected in the second sample). Figure 5 shows the chart view of the feature vector, where each coloured curve shows the change of the peptide intensity over the retention time in one sample. PEAKS™ Q is available as an optional module within the PEAKS™ Studio 5.1 package.

Figure 4. 3D view of a peptide feature vector.

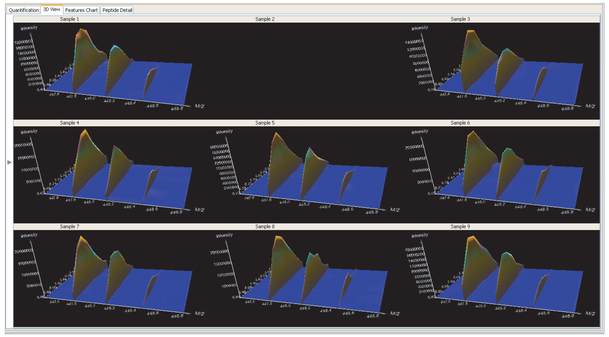
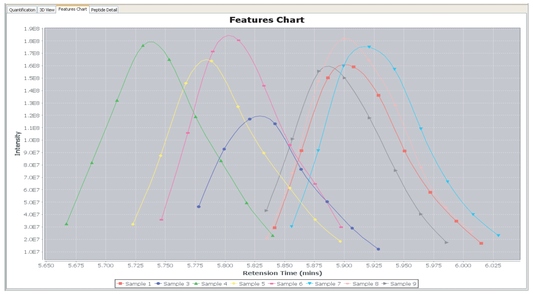


Figure 5. Chart view of a peptide feature vector.



References

E. Bonneil, G. Jaitly, N. Jaitly, C. Pomies, P. Thibault. Comprehensive expression profiling and trace-level identification of unlabeled peptides ions in 2DLC-MS proteomics experiments using integrated detection and clustering software. ASMS 2007 poster.

G. Jaitly, N. Jaitly, E. Bonneil, M. Ghitun, C. Pomies, M. Fortier, P. Thibault. MassSense, a new software for peptide detection and abundance comparison to monitor protein expression changes using unlabeled peptides. ASMS 2006 poster.