Peptide Identification Using Protein Sequence Databases Derived from RNA-Seq Data

Lei Xin, Lian Yang, Baozhen Shan
Bioinformatics Solutions Inc, Waterloo, ON

Bioinformatics Solutions Inc.

Introduction

Protein sequence databases are widely used for peptide identification with MS/MS spectra. However, only a fraction of the proteins in protein sequence databases, such as UniProt/Swiss-Prot, are expressed in a specific sample. Larger databases yield more distraction and lower signal-to-noise ratios. As a result, sensitivity is reduced to maintain a low false discovery rate. RNA-Seq provides a far more precise measurement of levels of transcripts and their isoforms than other methods. There is an increasing interest in incorporating RNA-Seq and protein sequence databases to improve peptide/protein identification by database searching in shotgun proteomics. In this study, we reported the performance of incorporating RNA-Seq and protein databases for peptide identification.

Method

1. All spectra were searched against a public protein sequence database. Identified spectra were filtered out at 1% of the false discovery rate (FDR).
2. To reduce the effect of signal-noise ratio, unidentified spectra were searched against the subset of the database in step 1, containing proteins for which there was evidence of expression based on the RNA-Seq data.
3. To improve the identification of peptide variants, unidentified spectra were searched against the database in step 1, plus additional sequences containing single amino acid substitutions derived from identifying single nucleotide variants in the RNA-Seq data.

Results

The public dataset from ABRF iPRG 2013 study was used [1], which contains LC-MS data with 133533 MS/MS spectra obtained from a LTQ-Orbitrap and three sequence databases. Database FPKM contains sequences of all non-redundant proteins derived from commonly used protein databases. Database FPKM_tg0 is a subset of FPKM, containing proteins for which there was evidence of expression based on the RNA-Seq data. Database FPKM_SNV contains all entries in FPKM plus additional sequences containing single amino acid substitutions derived from identifying single nucleotide variants in the RNA-Seq data.

PEAKS 6 software was used for data analysis. In the database search, precursor mass error tolerance was set to 20 ppm and a fragment ion mass error tolerance was set to 0.05 Da. Fixed modifications were TMT6plex on Lys and N-term and Carbamidomethylation on Cys. Variable modifications were Deamidation on Asn and Gln and Oxidation on Met. A target-decoy approach was used to control the FDR at 1% in this study.

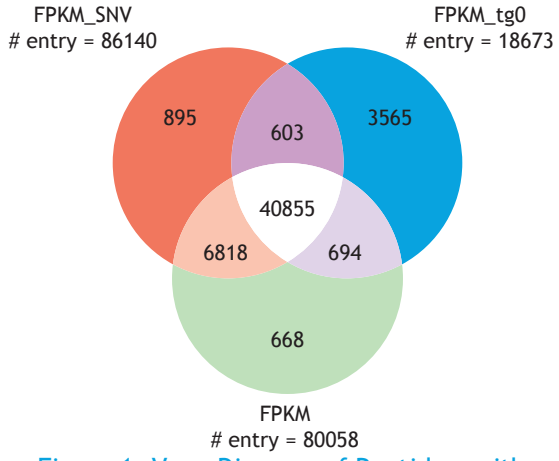

Figure 1. Venn Diagram of Peptides with FPKM, FPKM_tg0 & FPKM_SNV

A comparison of identified peptides with three databases was shown in Figure 1. The majority of identified peptides (40855) were common to each database. There were 3565 peptides identified exclusively by FPKM_tg0, significantly lower than 1% FDR. This was in contrast to FRKM, which identified 80058 proteins and FPKM_SNV with 86140 proteins. FPKM_tg0 contained only 18673 proteins for which there was evidence of expression based on the RNA-Seq data. There was an 8% improvement towards sensitivity for peptides exclusively identified by FPKM_tg0, due to the effect of signal-noise ratio [2].

PEAKS DB (database search) includes SPIDER homology search, designed to identify homologous peptides from commonly used protein databases. Compared to FRKM, FPKM_SNV contains an additional 6082 sequences, each with single amino acid substitutions and variants derived from the RNA-Seq data. There were 895 peptides found independently by FPKM_SNV, slightly high than the 668 peptides found autonomously by FRKM; however, both are close to 1% FDR. There were 6818 peptides shared solely between FPKM_SNV and FPKM_tg0, well below 1% FDR. It appeared that the majority of homologous peptides were identified using SPIDER (homology matching) without referencing RNA-Seq data.

To further evaluate the performance of peptide identification from the protein sequence databases derived from RNA-Seq data, a workflow in Figure 2 was proposed. All spectra were searched against FPKM first. Then, unidentified spectra were searched against FPKM_gt0, and finally against FPKM_SNV. All identified peptides were summarized with 1% FDR.

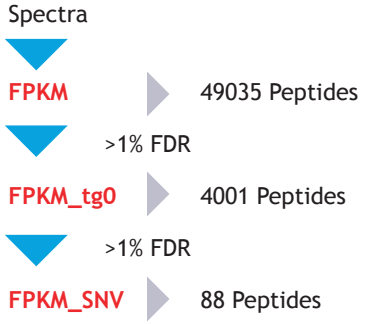

Figure 2. Workflow of Data Analysis with Databases Derived from RNA-Seq Data

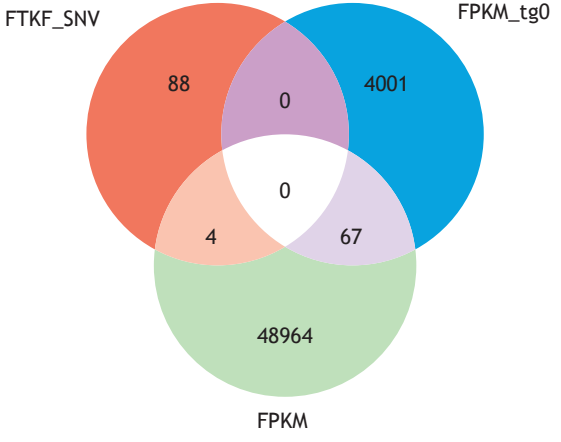

Figure 3. Venn Diagram of Peptides in the Workflow

The performance was shown in Figure 3. By incorporating RNA-Seq data into protein sequence databases, 4089 more peptides were identified.

Conclusion

The sensitivity of the peptide identification was improved by incorporating RNA-Seq data into proteomics.

Reference

[1] http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm
[2] Wang X, et al. J Proteome Res. 2012 Feb 3;11(2):1009-17