

A Strategy to Analyze Temporal Bottom-up Proteomics Data

Xiuxia Du¹, Stephen J. Callister¹, Nathan P. Manes¹, Joshua N. Adkins¹, Roxana A. Alexandridis², Xiaohua Zeng³, Jung Hyeob Roh³, William E. Smith³, Timothy J. Donohue², Samuel Kaplan³, Richard D. Smith¹, Mary S. Lipton^{1*}

1. Fundamental Science Division, Pacific Northwest National Laboratory, Richland, WA 99352

2. Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706

3. Department of Microbiology & Molecular Genetics, Medical School, University of Texas-Houston, Houston, Texas 77225

*email: mary.lipton@pnl.gov

Introduction

Performing systems biology studies with the goal of developing predictive biological models requires detailed knowledge of the dynamics of biological systems. Most quantitative proteomics efforts using mass spectrometry have focused on static protein abundance measurements. Recently however, liquid chromatography coupled to mass spectrometry (LC-MS) based proteomics has witnessed considerable progress in instrumentation technology, throughput, and data analysis. With these technological advances, the study of protein abundance dynamics has become feasible.

Nevertheless, the potential benefits from studying proteomic dynamics are accompanied with a number of challenges that include how to efficiently analyze the resulting large amounts of data. Among the prominent data analysis challenges are how to handle the extraneous variability and missing abundance values, and how to identify significant temporal patterns.

This article describes an analytical strategy developed specifically to confront data analysis challenges inherent to dynamic bottom-up proteomics studies. To evaluate its utility, this strategy was applied to data from a time-course study of *Rhodobacter sphaeroides* 2.4.1 transitioning from aerobic to photosynthetic metabolism.

Experimental and data analysis framework

Cell samples were collected at ten time-points over a period of 18 hours. Each sample was analyzed by LC-MS five times (50 analyses total) and abundance values of peptides were obtained. A procedure designed to extract significant temporal protein abundance patterns is outlined in Figure 1.

Peptide filtering: Due to random events that occur during LC-MS and subsequent data processing, peptides that might have been observed by chance were of questionable identity and were filtered out using peptide observation counts among the technical replicates.

Abundance normalization: The question of how to normalize the time course data and how

to select the optimal reference for the normalization is answered.

Missing-value imputation: Missing abundance values for each peptide are imputed based on available time-course measurements from all of the technical replicates. The imputed abundance values are required to abide by the existing temporal trend in the available abundance values.

Signal reconstruction: The original continuous peptide abundance values vs. time are reconstructed from the measured abundance values at discrete time points.

Significance analysis of peptide and protein dynamic abundance patterns: Identifying proteins that have significant temporal expression patterns, and thus are of potential biological importance, is inferred from abundance change patterns of at least two peptides belonging to the protein.

Figure 2 shows the results of abundance normalization, missing value imputation, and signal reconstruction for a representative peptide.

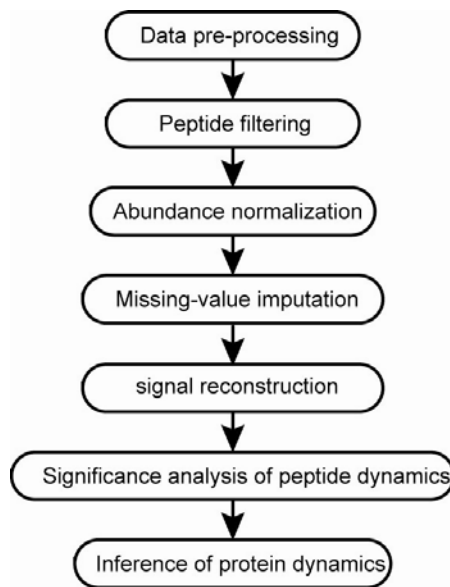


Figure 1. Flow chart of the data analysis procedure.

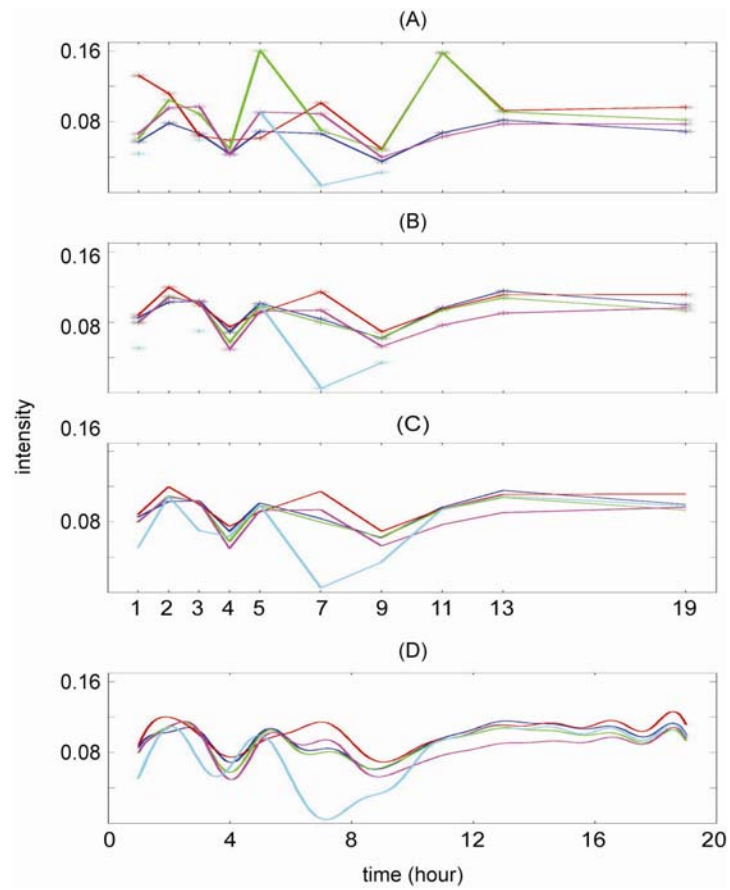


Figure 2. Results of abundance normalization, missing value imputation, and signal reconstruction for one representative peptide. (A) before normalization. (B) after normalization. (C) after imputation. (D) after reconstruction.