Statistics and Its Interface Volume 5 (2012) 1–2

Editorial

Proteomics is a rapidly developing field. With the help of high-throughput mass spectrometry instruments, researchers are now able to take a shotgun approach to monitoring a vast number of cellular activities at the protein/peptide level across different samples/conditions (e.g., disease versus normal, or drug treatment versus control).

Given the huge amount of data acquired from the shotgun proteomic experiments, data analysis becomes a key element in proteomic studies. Due to the stochastic nature of mass spectrometry instruments in ionizing and digesting biological samples, statistical methods are indispensable to appropriately analyze and interpret proteomics data. We feel that biomedical research will benefit greatly from having more researchers in the statistics community being exposed to the many exciting and challenging problems in proteomics. This has motivated us to edit this special issue on "statistical methods for proteomics".

The key analysis issues in current proteomic studies are to identify which proteins/peptides are present in biological samples, what their expression levels are, what post-translational modifications these proteins have, relationships between protein structures and functions, among others. This special issue covers the following topics:

1. Protein/peptide identification

Popular peptide identification methods include database search, *de novo* sequencing, and library search methods.

Serang and Noble present a nice overview of protein identification methods using tandem mass spectrometry. They describe predominant methods using a common computational framework. For each method, they analyze and evaluate the outcome and methodology of published methods.

Li et al. present a Bayesian perspective on the peptide and protein identification problem. They provide a unified interpretation for both the database search and *de novo* sequencing approaches.

Lam summarizes a new class of library search-based peptide identification methods and provides a high-level description of the underlying algorithms and remaining challenges.

2. Identification of post-translational modifications

In identifying post-translation modifications (PTM), controlling the false discovery rate (FDR) is very important. Fu argues that it is not appropriate to equally consider peptides with and without PTMs in calculating the FDR. He further elaborates on a few factors influencing the PTM-related FDR calculation and overall FDR calculation.

In order to understand more about the regulation mechanism of phosphorylation, people are interested in identifying the common patterns (also known as motif) around the phosphorylation sites in protein sequences. Existing motif predication methods cannot provide theoretical guidance on distinguishing true phosphorylation motifs from false motifs. Gong and He use permutation to calculate p-values of identified motifs and to estimate their statistical significances.

3. Protein/peptide quantification methods

Quantifications are needed in comparative proteomics studies. Existing methods fall into two categories: label-free MS data quantification and labeled MS data quantification. They differ by whether labeling reagents are used in data acquisition.

Generally, label-free LC-MS/MS data is quantified using two measures: the spectral count derived from the identification of MS/MS spectra and the ion abundance derived from the LC-MS data. Milac, Randolph and Wang compare the performance of these two measures by using two case studies. They find that using the ion abundance can reveal more properties in these two data sets than using the spectral count measure.

In another paper, Leitch, Mitra and Sadygov study the label-free shotgun proteomics data analysis frameworks by comparing three popular methods: QSpec, quasi-Poisson, and negative binomial distribution based method. After applying these methods to analyze a control data set and a data set with a known differential expression, they observe that the quasi-Poisson statistical model is significantly more liberal in determining a protein as differentially expressed than the QSpec, and the negative binomial model is more conservative.

In labeled MS data, the iTRAQ (isobaric Tags for Relative and Absolute Quantitation) labeling technique has been widely used in proteomic experiments due to its ability to allow simultaneous labeling of proteins in multiple samples. As iTRAQ labeling is a chemical process with a stochastic nature, there exists a number of statistical challenges. Luo and Zhao review the computation problems in iTRAQ data analysis, especially the nonrandom missing in the iTRAQ data.

4. Protein structure and function prediction

Protein functions are closely related to their structures. Given a protein sequence, it is very challenging to predict the corresponding 3-D structure. While various score functions have been developed, their applications in protein structure predictions are unsatisfactory. He et al. develop a novel two-stage optimization method to combine a set of basic scoring functions for improving the selection performance. The new method has achieved better performance than existing methods.

5. Different feature selection methods and classification methods for proteomics-based disease biomarker discovery

Disease biomarker discovery is of clinical importance for early diagnosis of complex diseases such as cancer. Proteomics is a very promising tool to discover such biomarkers. Morris reviews existing analytical challenges that must be addressed properly for effective comparative proteomics studies to yield potential biomarkers, including experimental design, preprocessing, feature extraction, and statistical analysis accounting for the inherent multiple testing issues.

6. Interaction between proteomics and transcriptnomics

It is well known that proteins are the products of gene expressions. At the same time, it is interesting to note that information about proteins can also be helpful in studying genes. Chen et al. review recent progresses of using a phenotype similarity profile (phenome) and a protein-protein interaction network (interactome) for the prioritization of disease-associated genes.

The above topics are by no means complete. There are other important and interesting topics that are not covered. In this sense, this special issue is just a “sampling of the field”. Given that the journal *Statistics and Its Interface* targets at interdisciplinary problems with interactions between different fields and statistics, we hope that this issue can invoke more interest and enthusiasm in the statistics community to help tackle the proteomic problems.

Finally, we would like to thank the Editor-in-Chief of *Statistics and Its Interface* Professor Heping Zhang at Yale University for his tremendous help during our editing process. We hope you like this special issue.

Weichuan Yu
Hongyu Zhao