Clin Proteom (2009) 5:1–2
DOI 10.1007/s12014-009-9026-3

Bioinformatics for Comparative Proteomics

C. Nicole Rosenzweig · Zhen Zhang

Published online: 1 April 2009

Keywords Bioinformatics · Proteomics · Mass spectrometry

Since the introduction of the concept of the "proteome" 10 years ago, the volume of data collected during a proteomic experiment has dramatically increased. Proteomic experimental methods have expanded to include multi-dimensional protein separation, evaluation by mass spectrometry, protein identification through tandem mass spectrometry, and quantitation through multiple reaction monitoring. In order to characterize these results, diverse bioinformatics tools have been developed. In the broadest sense, the goal of these tools is to alleviate the bottleneck resulting from the volume of data collected.

Bioinformatics has provided significant contributions to the area of proteomics. One area that has contributed significantly to the analysis of proteomic data is the development of common data formats for proteomic mass spectrometric data, such as mzXML and mzData, and the recent release of mzML as a joint format [1]. With the availability of such common data formats, open-source tools have become not only available, but useful to a wide audience. Availability and utility of data repositories has provided greater accessibility to data during both publication review as well as providing additional data for comparison in subsequent data analysis. The value of these repositories will only increase as additional data is deposited over time.

C. N. Rosenzweig (✉) · Z. Zhang
Center for Biomarker Discovery, Department of Pathology,
Johns Hopkins University,
Baltimore, MD, USA
e-mail: CN.Rosenzweig@gmail.com

While bioinformatics tools have contributed to the common research goals, several significant hurdles remain in proteomics where bioinformatics could offer help. As an example, even with the recent advances in sample processing and instrumentation technologies, reproducibility of experimental data remains an issue that hinders meaningful interpretation of the results. Bioinformatics tools that incorporate knowledge of the physical process of proteomic data collection to improve protein identification, quantitation, and data normalization would be very valuable in studies involving differential analysis of protein expressions. Tools that assess quality of proteomic data based on sound statistical principles are also needed. Such assessment will allow us draw conclusions from proteomic studies, from simple fold-changes to networks and models, with clearly stated limitations in terms of statistical confidence.

In the current special issue, we selected several papers with novel approaches which could result in improvement in reproducibility of proteomics results. Topics covered include improvements in quality assessment of tandem mass spectrometry, combining search results to improve the reliability of the predicted identities, and steps to improve the reliability of MRM results. We hope these papers will serve as examples to provide insights into how proteomic analysis pipelines can be improved in the near future.

Quality assessment of tandem mass spectrometry has been evaluated previously, but the methods have focused primarily on supervised machine-learning approaches. The limitations to these approaches are twofold: (1) they require manually validated data to develop a classifier and (2) it is difficult to create a classifier which is robust enough to perform reasonably, given the inherently inconsistencies observed when tandem mass spectrometers are utilized under different conditions. Furthermore, the tandem mass

spectra returning predicted protein identifications can be as low as 10–20%. In an effort to resolve these issues, an alternative approach for discriminating between poor quality and high-quality spectra using an unsupervised method, K-Means Clustering, is presented [2]. This method can be trained to the specific data collected from an experiment. The method reduced the database search time by greater than 50% with a loss of potential identifications between 2% and 14% in the dataset evaluated. Improving the quality of the spectra submitted for identification serves to reduce the time required to return the peptide identification results as well as reduce the probability of a published results relying on low-quality data.

Once a set of reasonable spectra have been selected for analysis, protein identification is pursued. Both commercial and freeware search engines are available for this task. Each method provides a score and associated rank of the predicted peptide sequence. Many also provide an estimate of statistical significance. However, a direct comparison of the predicted matches demonstrates that these methods can present inconsistent results. Most of the search engines agree on approximately 80% of the peptide identifications. Unfortunately, the remaining predicted identifications vary depending on the search engine used. As a result, effort has been dedicated to improving the reliability of the results. Statistical significance re-estimation, supervised machine learning for scoring and prediction, and combining multiple search engines' results have been pursued. PepArML [3] provides an underlying framework to combine each of these improvements in a model-free, unsupervised manner.

In contrast with the unbiased methods utilized in the discovery phase, which is characterized by analysis of many proteins in few samples, the validation phase of biomarker research focuses on measuring a few biomarkers in many samples. Validation and quantitation in the clinical environment traditionally is performed using immunoassay, rather than the platform used during the discovery phase. Unfortunately, immunoassay is a cumbersome tool when tasked with measuring multiple proteins simultaneously. Consequently, technical advances were pursued which allowed for the discovery and validation to be performed on the same technology. Multiple Reaction Monitoring (MRM) has emerged as a sensitive approach for quantitation of proteins [4]. The final contribution of this special issue is focused on analysis of data collected through targeted quantitative MS. Two significant contributions to the field of MRM research are pursued: the optimization of MRM-transition selection, and a novel scoring function called T_{corr} [5]. Unlike S_{corr}, which is a prominent method utilizing spectrum information for database searching algorithms in the analysis of tandem MS data, T_{corr} uses a small number of transition ions to predict peptide ID, thus providing the possibility of improving the reliability of the result.

While bioinformatics tools are necessary for the successful completion of a study, they cannot substitute for statistically and scientifically sound study and experimental design. A confounding variable, once introduced, distorts data irreparably. Post-collection analysis simply cannot remove limitations introduced due to a lack of randomization, replication, or avoidance of systematic bias. In this special edition of Clinical Proteomics, we evaluated the state of the art and advances in bioinformatics tools that could be used to improve the reliability of proteomics results collected from data utilizing good study design. We are very grateful to the authors for their excellent contributions, the reviewers for their timely evaluations, and the Editor-in-Chief and staff of *Clinical Proteomics* for the support provided throughout the construction of this special issue.

References

1. Deutsch E. mzML: a single, unifying data format for mass spectrometer output. Proteomics. 2008;8(14):2776–7.
2. Ding J, Shi JH, Wu FX. Quality assessment of tandem mass specra using weighted k-means. Clin Proteomics. 2009;5(1). doi:10.1007/s12014-009-9025-4
3. Wu X, Tseng CW, Edwards N. An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. Clin Proteomics. 2009;5(1). doi:10.1007/s12014-009-9024-5
4. Anderson L, Hunter CL. Quantitative mass spectrometric MRM assays for major plasma proteins. Mol Cell Proteomics. 2006;5(4):573–88.
5. Liu J, Hewel JA, Fong V, Chan-Shen-Yue M, Emili A. Critical evaluation of product ion selection and spectral correlation analysis for biomarker screening using targeted peptide multiple reaction monitoring. Clin Proteomics. 2009;5(1). doi:10.1007/s12014-009-9023-6