Computational Proteomics: Session Introduction

Bobbie-Jo Webb-Robertson, William Cannon, Joshua Adkins, and Deborah Gracio

Pacific Symposium on Biocomputing 11:212-218(2006)

THE CHALLENGE OF PROTEOMIC DATA, FROM MOLECULAR SIGNALS TO BIOLOGICAL NETWORKS AND DISEASE

BOBBIE-JO WEBB-ROBERTSON, WILLIAM CANNON, JOSHUA ADKINS, DEBORAH GRACIO

Pacific Northwest National Laboratory
Richland, WA 99352

Mass spectrometry (MS) based proteomics is a rapidly advancing field that has great promise for both understanding biological systems as well as advancing the identification and treatment of disease. Breakthroughs in science and medicine due to proteomics, however, are coupled with our ability to overcome significant challenges in the field. These challenges are multi-scalar, spanning the range from the statistics of molecules and molecular signals, to the phenomenological characterization of disease. The papers presented in this section are a representative snapshot of these challenges that span scale and scientific disciplines.

The multi-scalar challenges are hinted at in figure 1, which depicts a typical MS-based proteomics analysis that is performed in many laboratories. Proteins are first extracted from cells and then may be cut at defined locations in the sequence by adding enzymes called peptidases to the protein extract. The solution of peptides is then partially separated by the use of liquid chromatography. The partially separated peptides are shown in the chromatogram in the top panel of figure 1. Each peak in the chromatogram consists of multiple peptides. A peak is then analyzed by the mass spectrometer attached at the end of the chromatography system. The co-eluting peptides are then introduced into the gas phase and, as shown in the middle panel of figure 1, separated by their mass-to-charge ratios in the mass spectrometer. Ideally, the peptides have been completely separated from each other at this stage. The analysis may stop at this stage, or peaks from the initial mass spectrum may be isolated, and those peptides can be subject to a second round of analysis where the isolated peptides are vibrationally excited by collision with an inert gas. The peptides then fragment at labile bonds and a subsequent mass spectrum is obtained of the fragments of the peptide, shown in the bottom panel of figure 1. Because the peptides tend to fragment into recognizable patterns, the identity of the peptide can frequently be determined from this mass spectrum.

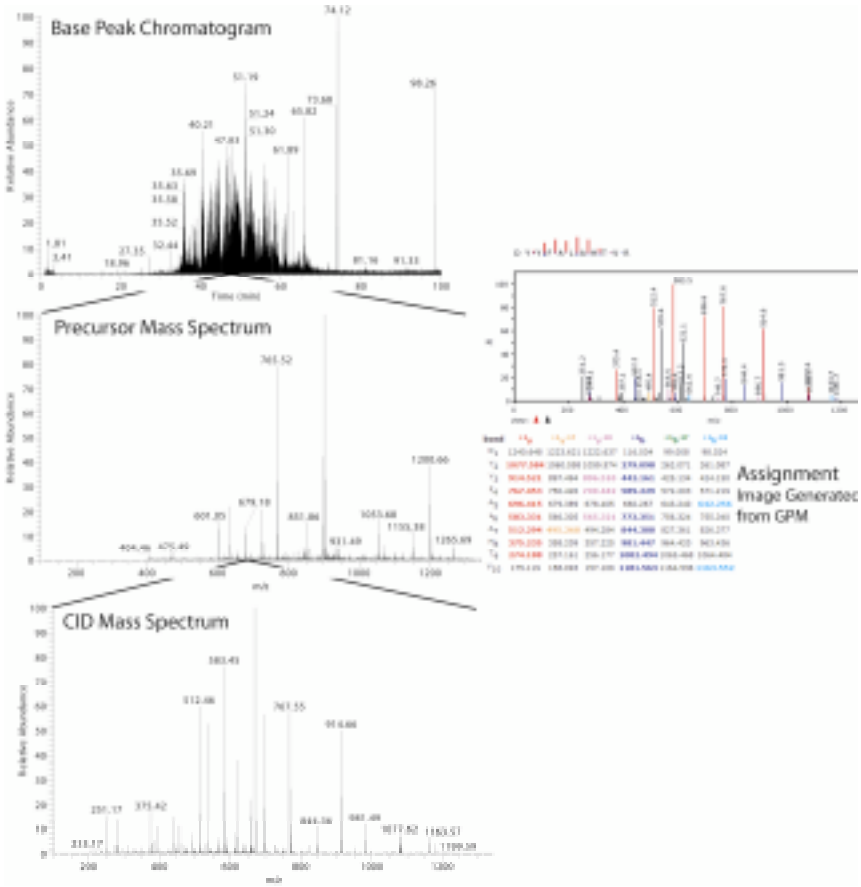

Figure 1. (Top). Peptides are partially separated from each other using liquid chromatography, resulting in a chromatogram in which each peak consists of one or more peptides. (Middle) The peptides co-eluting in a peak from the chromatography system are further separated in the mass spectrometer by their mass-to-charge ratio. The peak intensities reflect the number of molecules that have been isolated. A peptide in the MS spectrum can be isolated and selected for further analysis in which the peptide is collisionally-activated by an inert gas. (Bottom) The peptide may then fragment apart and the resulting mass-to-charge and abundances of the fragments are measured by a subsequent round of mass spectrometry

Currently, many of the instruments use proprietary software for transforming raw spectra into sets of peak locations and intensities. This results in difficulties when gauging the accuracy of the resulting peak intensities and locations; ultimately this limits the interpretations that lead to biological results improvements can be made in the existing methods. Many aspects of this multi-scalar problem can be studied independently, but many more are dependent on

the approaches taken earlier in data processing. In this regard, surveying the depth and scope of computational proteomics is key to understanding the information content of each proteomic experiment. We start more detailed discussions with an example to a critical step in computation proteomics in which **Lange *et al.***[b] use open source wavelet approach to peak picking, the process of transforming the raw spectra into the familiar impulse plots used to display intensity and mass-to-charge location information.

Comparative Proteomics

Following quantification from raw mass spectra, a wide number of approaches may be employed to interpret them. In comparative proteomics, a first step is the normalization of the data to enable comparison between experiments and determination of differential abundance levels of the proteins in a specific proteome. In this area, **Wang, *et al.***[b] introduce a rank-inspired probability method for normalizing ion peaks between samples. The ultimate goal is to increase the reliability of the quantitation process so that differential expression of proteins can be measured under various growth conditions.

Another challenge in comparative proteomics is aligning the spectra so that the correct peaks can be compared and the ratio of the normalized intensities determined. A problem that has been encountered with shotgun proteomics experiments in applying this method, however, is that the alignment of the spectra in both elution time and mass-to-charge values is difficult due to the number data points. A typical shotgun experiment in which the cell lysate is digested with trypsin and then separated using liquid chromotagraphy may generate 70,000 data points or more. An alternative method is to label the samples using isotopes that are normally in low abundance in nature. Samples can then be mixed and the alignment step essentially is obviated because isotopic labeling has only a small affect on the elution properties of the peptides and proteins [1]. Further improvements may be made by reducing the complexity of the peptide mixture and isotopically labeling peptides using isotope-encoded affinity tags (ICAT) [2]. Recent advances in isotope labeling technology have enabled the comparison of differential protein expression for up to four samples[3]. **Michailidis and Andrews**[b] present a statistical framework for the analysis of variance of multiple isotopically-labeled proteins and formal hypothesis testing as to whether the proteins are differentially expressed.

---

[b] Session Publication.

Network Inference

If the differential expression can be done in a reliable and informative manner as **Michailidis and Andrews**[b] present, then it would be conceptually possible, with proper experimental design, to infer networks of protein regulation from the mass spectrometry results. One of the persistent questions regarding the use of graphical models to infer networks from RNA or protein expression data is how to optimally design experiments in order to achieve the greatest resolution of the network. Typically this argument is cast in terms of whether one should study cell populations under different treatment or environmental conditions, or whether one should use time series data. The former are attractive because each data series represents an independent observation, while in the latter the observation of expression of each protein is correlated in time and thus is not an independent observation. On the other hand, time-series experiments that focus on a specific sub-network are easier to design. This trade off is considered in detail by **Page and Ong**[b]**.**

Regulatory networks are not the only networks of interest. Rapid advances are currently being made in the determination of protein interaction networks [4-9]. A typical approach in these assays is to use affinity purification techniques to pull-down a preselected *bait* protein and the *prey* proteins that interact, directly or indirectly, with the bait protein. This information can be used to construct a protein interaction network in which all discovered interactions are laid out. However, a protein-protein interaction network is not as informative as actually determining the protein complexes or machines that carry out the biological function in the cell. Determining these complexes from the pull-down data is computationally challenging because multiple-complexes may be present in any given pull-down data set. **Chu *et al.***[b] propose a solution to this problem that combines a kernel method to identify potential complexes, a latent feature model to address the number of complexes, and Bayesian statistics that can ultimately be used to bring in informative prior knowledge

Peptide Identification

Network analyses and any other analyses that seek to determine biological knowledge from proteomics experiments rely on the initial correct identification of peptides. The automation of peptide identifications using computers took a step forward in the late1980s and early 1990s in work by several groups that laid a foundation for the next several years[10-12]. In 1994-1995, the *SEQUEST* method was developed and published [13]. *SEQUEST* was the first example of high-throughput processing of proteomic data and has since become one of the

standards of the field. Although the code was developed in a relatively short time, the wide spread use of the tool is a testimonial to its utility.

Now that proteomics is being widely used in industry and research labs, however, there is a pressing need to solve the many of the peptide identification challenges that remain. Currently, only 25% or fewer of peptide spectra are identified with a peptide. There are many reasons for this. First, many fragmentation pathways are poorly understood and the current set of patterns that are searched for is limited. Second, most peptide identification tools only consider parent ions that have charges of +3 or less. Incomplete digestion of peptides by peptidases, such as trypsin, is likely to result in an abundance of higher charge state ions leading to a decrease of spectra that are currently identifiable. In addition, tools such as *SEQUEST* search genome sequence databases for peptides that are likely to result in a spectrum similar to the experimental spectrum under consideration.

In *SEQUEST* and most peptide identification tools, a fragmentation pattern is used in one way or another to determine this similarity. Typically, the peptide sequence is used as a template to generate the pattern, and the pattern or *model spectrum* is compared to the actual spectrum. The accurate development of these patterns is the topic of the work by **Arnold, *et al.***[b]**,** in which they employ a neural network to learn peptide fragmentation patterns from a training database of peptide spectra. This work is significant because it is believed that non-classical fragmentation patterns are largely missed in the identification process. **Wang *et al.***[b] take a different approach in which they use a minimal fragmentation model in a simple scoring function and then analyze the score and properties of the candidate peptides using a support vector machine (SVM). This approach extends the use of SVMs in peptide identification [14, 15] to not only choose spectra that have been correctly matched to a peptide, but to also choose the best candidate peptide from a sequence database for a given spectrum.

Alternate Spectra Assignment Algorithms

Although more and more genome sequences are becoming available, by far the majority of genomes have not been sequenced nor are they likely to be sequenced in the next 5-10 years. The alternative to searching a sequence database is to determine the peptide sequence *de novo* from the spectrum[11, 16, 17], or to use optimization to evolve a peptide sequence to match a spectrum [18]. *De novo* methods are attractive because the idealized problem, that of essentially spelling out a peptide from a set of mass peaks, intuitively corresponds to a problem that can be solved using graph theory. The devil, as usual, lies in the details. Real spectra are noisy and missing peaks from key

[b] Session Publication.

fragments are the rule rather than the exception. Graph theory analyses often result in a series of graphs, not all of which are compatible. **Liu, *et al.*** extend recent advances in this area by the application of tree decomposition to the problem that allows for compatible graphs to be found more readily and in faster time.

Post-Translation Modifications

Of the 75% or so of spectra that go unassigned with a peptide in a general database search, many of these are thought to be post-translationally modified peptides. The identification of these peptides is extremely important because the post-translational modifications (PTM) are one of the key mechanisms by which cells respond to external stimuli, resulting in the up and down regulation of genes. **Yan *et al.***[b] describe a point-process model that has the advantage used in dynamic programming models [19] in which the mass offsets for the PTMs are determined automatically, and is deployed in a fast cross-correlation framework.

Final Thoughts

Ultimately, a major motivation for investments into the development of proteomics is to develop advanced methods of disease diagnosis, understanding of disease processes, and remedies. Early detection of disease is important because the clinical outcome is much more favorable, in general, if the disease can be treated in an early stage. As a result, there is much interest in improvements at every level that can yield MS-detectable biomarkers that signal the presence of the disease long before more overt symptoms occur that signal advanced stages of disease. An example of this need and approach is **Pratapa *et al.***[b] present a hierarchical data analysis scheme for the identification of protein biomarkers that are indicative of lung cancer. Using data from mass spectrometric analyses of diseased and normal tissues, they compare a SVM classification with that of a Bayesian sparse logistic regression. They find known biomarkers as well as identify several more candidate biomarkers that may prove to be clinically useful.

The incredible diversity of problems and solutions is well sampled by the efforts of this session authors. Mass spectrometry-based proteomics is likely to offer a central role well into the future for understanding protein function and complex biological systems.

---

[b] Session Publication.

Acknowledgments

The session organizers would like to thank the authors of the 30 submissions to this session and express our regret that only a handful of the excellent papers can be presented. We would also like to express deep gratitude to the anonymous referees who together volunteered uncountable hours to provide the key input to make this session successful.

References

1. L. Pasa-Tolic et al., *J Am Chem Soc*. **121**(34): 7949-7959 (1999).
2. S.P. Gygi et al., *Nat Biotechnol*. **17**(10): 994-999 (1999).
3. P.L. Ross et al., *Mol Cell Proteomics*. **3**(12) : 1154-1169 (2004).
4. J.S. Bader et al., *Nature*. **22**(1): 78-85 (2004).
5. A.C. Gavin et al., *Nature*. **415**(6868): 141-147 (2002).
6. L Giot et al., *Science*. **302**(5651) : 1727 :1736 (2003).
7. Y Ho et al., *Nature*. **415**(6868) : 180-183 (2002).
8. A.J. Link et al., *Nature Biotech*. **17**(7): 676-682 (1999).
9. G. Butland et al., *Nature*. **443**(7025) : 531-537 (2005).
10. K. Biemann, Methods in Enzymology, ed. J.A. McCloskey. Vol. 193. 1990, San Diego: Academic Press, Inc.
11. C. Bartels, *Biomed Env Mass Spectrom*. **19**: 363-368 (1990).
12. M. Mann, C.K. Meng and J.B. Fenn, *Anal Chem*. **61**(15): 1702-1708 (1989).
13. K. Eng, A.L. McCormack and J.R. Yates III, *J Am Chem Soc*. **5**: 976-989 (1994).
14. D.C. Anderson et al., *Proteome Res*. **2**(2) : 137-146 (2003).
15. Cannon et al., *Proteome Res*. Web release September 10, 2005.
16. V. Dancik et al. *J Comput Biol*. **6**(3/4): 327-342 (1999).
17. J.A. Taylor and R.S. Johnson, *Rapid Comm in Mass Spectrom*. **11**: 1067-1075 (1997)
18. A. Heredia-Langner et al., *Bioinformatics*. **20**(14): 2296-2304 (2004).
19. P.A. Pevner et al., *Genome Res* **11**(2): 290-299 (2001)