Pacific Symposium on Biocomputing 12:403-408(2007)

COMPUTATIONAL PROTEOMICS: HIGH-THROUGHPUT ANALYSIS FOR SYSTEMS BIOLOGY

WILLIAM CANNON

Computational Biology & Bioinformatics, Pacific Northwest National Laboratory
Richland, WA 99352 USA

BOBBIE-JO WEBB-ROBERTSON

Computational Biology & Bioinformatics, Pacific Northwest National Laboratory
Richland, WA 99352, USA

High-throughput proteomics is a rapidly developing field that offers the global profiling of proteins from a biological system. These high-throughput technological advances are fueling a revolution in biology, enabling analyses at the scale of entire systems (e.g., whole cells, tumors, or environmental communities). However, simply identifying the proteins in a cell is insufficient for understanding the underlying complexity and operating mechanisms of the overall system. Systems level investigations generating large-scale global data are relying more and more on computational analyses, especially in the field of proteomics.

1. Introduction

Proteomics is a rapidly advancing field offering a new perspective to biological systems. As proteins are the action molecules of life, discovering their function, expression levels and interactions are essential to understanding biology from a systems level. The experimental approaches to performing these tasks in a high-throughput (HTP) manner vary from evaluating small fragments of peptides using tandem mass spectrometry (MS/MS), to two-hybrid and affinity-based pull-down assays using intact proteins to identify interactions. No matter the approach, proteomics is revolutionizing the way we study biological systems, and will ultimately lead to advancements in identification and treatment of disease as well as provide a more fundamental understanding of biological systems. The challenges however are amazingly diverse, ranging from understanding statistical models of error in the experimental processes through categorization of tissue types. The papers presented in this session are a representative snapshot of this broad field of research that spans scale and scientific disciplines.

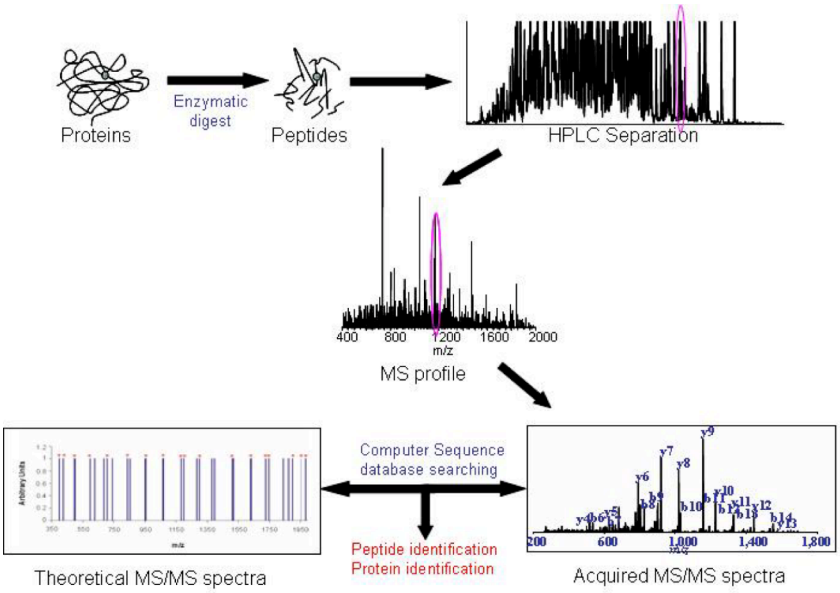

Figure 1. A typical MS proteomics process from protein isolation through peptide identification. Proteins are first isolated from other cellular components (top left) and then cleaved into peptides by enzymatic digestion(top middle). The peptides are partially separated using chromatography (top right) and then further separated by mass-to-charge ratios in the first stage of mass spectrometry (center). In tandem mass spectrometry, the isolated peptide is then collisionally-activated causing it to fragment into pieces. The mass-to-charge ratio of each fragment is measured (bottom right), and this fragmentation pattern is compared to model spectra (bottom left) for the peptide that are derived from training data or expert opinion. The peptide with a model spectrum that best matches the experimental spectrum is a potential match.

1.1. *HTP Mass Spectrometry*

The application of high-throughput (HTP) liquid-based separations and mass spectrometry (MS) to global profiling of proteins is providing an essential component to the challenge of understanding biology at a systems level (1). Figure 1 depicts a typical MS-based proteomics analysis that is performed in many laboratories. Enzymatic digestion of proteins extracted from cells results in the lysis at defined locations in the proteins producing peptides of predictable length (when derived from a known protein sequence). Reversed phase high performance liquid chromatography is used to partially separate the peptides in the solution. The eluting peak consists of a population of peptides which are analyzed by a mass spectrometer interfaced with the chromatography system. The electrospray process aerosolizes and ionizes the peptides into the gas phase and the charged particles are propelled into the mass spectrometer for analysis. The mass spectrometer scans the population of eluting ions, measures the mass

to charge ratio, and in the case of tandem mass spectrometry proceeds to the fragmentation step. This step consists of the capture of all ions in a narrow mass-to-charge range in the ion trap of the mass spectrometer, the 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. Because the peptides tend to fragment into recognizable patterns, the identity of the peptide can frequently be determined from this spectrum.

1.2. *HTP Yeast-Two-Hybrid*

In contrast to the destructive technique of HTP MS-based approaches, two-hybrid assays (2) are used for assessing protein-protein interactions in live cells. A typical implementation of the two-hybrid assay involves the attachment of bait and prey proteins to separated binding and activating domains of a transcription factor, typically GAL4, that controls for the production of a reporter protein. In principle, if the bait and prey interact then the modularized domains of the transcription factor are brought together and the newly combined transcription factor can both bind DNA and activate the gene coding for the reporter protein. Conversely, if the reporter protein is present in the assay, then it is presumed that the bait-prey pair interact. False negatives can occur for several reasons, such as when the covalent attachment of the transcription factor domain to the bait protein interferes with interaction with the prey protein. Likewise, false positives can occur if adaptive mutations or auto-activation result in expression of the reporter protein regardless of interaction between the bait and prey proteins. While issues regarding the interference of the binding of the bait to the prey due to blocking by the transcription factor modules may represent a random, independent source of errors, auto-activation is a systematic error affecting the entire screening process. Two-hybrid methods are most frequently used in genetically-tractable organisms such as yeast, *C. elegans*, and *Drosophila*. Recent development of bacterial two-hybrid systems may eventually result in the expansion of this method to many other genomes.

2. Challenges

2.1. *Accuracy of Peptide Identification*

Intrinsic to the MS-based proteomics measurement process is the comparison of MS/MS fragmentation patterns to model fragmentation patterns derived from the predicted peptides of a sequenced genome, which provides the basic peptide identification upon which all other evaluations are based (3-5). The common approach is to search the experimental spectra against a database of computationally generated model spectra in a database representing the constituent peptides of the entire genome. The computational peptide

identification process measures how well the mass peaks in an experimental spectrum match those in the model spectrum of a candidate peptide (3, 5-7). However, these database search routines are known to return both correct identifications against the experimental spectra, as well as a similar number of false positives. Considerable work on the data analysis front is still required (8).

The false positive problem of MS-based proteomics is largely due to the introduction of many sources of errors through the entire experimental and identification stages. Since the experimental observation of peaks introduces a mass error, a mass error distribution is often used in this matching process (9-11), i.e., the peaks of the experimental and theoretical are not expected to match up exactly. However, in most computational identification methods to-date these error models have followed simple statistical distributions. In fact, by far the most widely used distribution is the uniform distribution. **Fu et al.**[a] present, improved estimations of mass error distributions that can be incorporated into the identification process. Another major technical challenge lies downstream from the database search routines. These database search procedures typically return several metrics associated with the match, creating a challenge in separating true from false identification. To counter this problem there has been a fair level of effort placed on the development of probability-based scores(12-15). These statistical metrics have alleviated some problems with false positives, but make uniform assumptions about the identifiability of each peptide. That is, it is generally universally assumed that all peptides are equally detectable. **Alves et al**.[b] revisit the protein inference problem accounting for this assumption on peptide detectability.

2.2. *Network Inference*

Rapid advances are currently being made in the determination of protein networks (16-19). Typical approaches use either a two-hybrid screening of an entire genome, or affinity purifications to pull-down a pre-selected *bait* protein and the *prey* proteins that interact, directly or indirectly. This information can be used to construct a protein interaction network in which all discovered interactions are laid out. A common challenge in both affinity-based methods and two-hybrid screens is that of estimating and reducing error rates. The paper

[a] Y. Fu. W. Gao, S. He, R. Sun, H. Zhou, R and R.Zeng. Mining tandem mass spectral data for more accurate mass error model for peptide identification. *PSB 2007*.

[b] P. Alves, R.J. Arnold, M.V. Novotny, P. Radivojac, J.P. Reilly and H. Tang. The protein inference problem in shotgun proteomics – revisited. *PSB 2007*.

by **Sontag et al.**[c] describes a novel approach to estimating errors in two-hybrid experiments which could be adopted also for affinity purifications. Appropriately modeling this error allows better use of the data leading to better identification of interacting proteins. Due the error in the experimental system, it is common to utilize multiple sources of diverse information in defining protein interactions, for example, cell location or sequence motif composition. However, a common challenge in integrating all this information with the experimental data is that each source has varying levels of reliability. Computing reliability metrics for multiple sources of information to infer networks is the topic of **Leach et al.**[d]**.**

3. Final Thoughts

Ultimately, a major motivation for investments into the development of proteomics and systems biology is to develop advanced methods of disease diagnosis, understanding disease processes, and remedies. Each experimental approach offers a unique view, for example unlike MS/MS or two-hybrid approaches, MALDI-based Imaging MS (IMS) offers an approach to study the spatial distribution of biomolecules, such as proteins, in tissue. However, similar to other imaging based methods, classes in the data associated with the tissue must be identified in order to differentiate diseased tissue. A principal component analysis based approach is taken for IMS data in **Van de Plas et al.**[e].

The field of HTP proteomics is becoming a central component enabling systems level analyses. The level of complexity from inception of the experimental technique through the data analysis and modeling is incredible. As seen in this session research is taking place at the level of errors within a mass spectrum through the study of entire tissues. Proteomics is likely to offer a central role in understanding protein function and complex biological systems leading to a new revolution in advanced targeted therapeutics to treat disease.

Acknowledgments

The session organizers would like to express deep gratitude to the anonymous referees who together volunteered uncountable hours to provide to key feedback to make this session successful. The session chairs were supported through funding provided by the U.S. Department of Energy Office of Advanced

---

[c] D. Sontag, R. Singh and B. Berger. Probabilistic modeling of systematic errors in two-hybrid experiments. *PSB 2007.*

[d] S.M. Leach, A. Gabow L. Hunter and D. Goldberg. Assessing and combining reliability of protein interaction sources. *PSB2007.*

[e] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor and E. Waelkens. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *PSB 2007.*

Scientific Computing Research under contract 47901, and by the Office of Biological and Environmental Research under contract 43930, as well as PNNL laboratory directed research and development funds.

References

1. Nesvizhskii, A. I. & Aebersold, R. (2005) *Mol Cell Proteomics* **4,** 1419-40.
2. Fields, S. & Song, O. (1989) *Nature* **340,** 245-6.
3. Cannon, W. R., Jarman, K. H., Webb-Robertson, B. J., Baxter, D. J., Oehmen, C. S., Jarman, K. D., Heredia-Langner, A., Auberry, K. J. & Anderson, G. A. (2005) *J Proteome Res* **4,** 1687-1698.
4. Pappin, D., Rahman, D., Hansen, H., Bartlet-Jones, M., Jeffery, W. & Bleasby, A. (1996) *Mass Spectrom. Biol. Sci***,** 135-150.
5. Yates, J. R., Eng, J. K., McCormack, A. L. & Schieltz, D. (1995) *Anal Chem* **67,** 1426-1436.
6. Eng, J. K., McCormack, A. L. & Yates, J. R. (1994) *Journal of the American Society for Mass Spectrometry* **5,** 976-989.
7. Mann, M. & Wilm, M. (1994) *Analytical Chemistry* **66,** 4390-9.
8. Patterson, S. D. (2003) *Nat Biotechnol* **21,** 221-222.
9. Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. (1999) *J Comput Biol* **6,** 327-42.
10. Fenyo, D., Qin, J. & Chait, B. T. (1998) *Electrophoresis* **19,** 998-1005.
11. Frank, A. & Pevzner, P. (2005) *Anal Chem* **77,** 964-73.
12. Anderson, D. C., Li, W., Payan, D. G. & Noble, W. S. (2003) *J Proteome Res* **2,** 137-46.
13. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. (2002) *Anal Chem* **74,** 5383-5392.
14. Moore, R. E., Young, M. K. & Lee, T. D. (2002) *J Am Soc Mass Spectrom* **13,** 378-386.
15. Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottaz, H. M., Anderson, G. A., Shen, Y., Jacobs, J. M., Camp, D. G., 2nd & Smith, R. D. (2004) *J Proteome Res* **3,** 760-769.
16. Bader, J. S. (2003) *Bioinformatics* **19,** 1869-74.
17. Butland, G., *et al.* (2005) *Nature* **433,** 531-7.
18. Giot, L., *et al.* (2003) *Science* **302,** 1727-36.
19. Ho, Y., *et al.* (2002) *Nature* **415,** 180-3.