Software platforms for quantitative proteomics
— Dagstuhl Seminar —

Ole Schulz-Trieglaff

International Max Planck Research School
for Computational Biology and Scientific Computing & Free University Berlin,
Department of Computer Science, Takusstrasse 9, D-14195 Berlin, Germany
trieglaf@inf.fu-berlin.de

Abstract. In recent years, it has become obvious that mRNA expression does not always correlate with protein expression. It seems that a full understanding of the complexity of life can only be obtained by examining abundances of proteins under varying conditions. Accurate measurements of these expression values is crucial. This field of research also requires new computational efforts since the data, often from mass spectrometry experiments, is very complex. We present two academic software platforms that offer means to reduce, analyse and compare protein expression data gained from liquid chromatography coupled with mass spectrometry. We outline their methodology and compare them to our own project, OpenMS, which is currently developed in our research group at the Free University Berlin in collaboration with the Kohlbacher group at Tuebingen University.

Keywords. Proteomics, quantification, software platforms

1 Introduction

Liquid chromatography coupled with mass spectrometry (LC-MS) has been used extensively to identify and quantify proteins in a sample [1]. Recently, the development of software tools for the quantitative analysis of data from mass spectrometry experiments has aroused much interest. Many different tools are now available that implement various pre-processing such as denoising and peak picking and finally allow to detect the proteins that were contained in the sample together with an estimate of their abundance. However, these tools differ in approach and scope. Some focus on the management and annotation of proteomics data, others also offer means to further process and analyse the data sets.
In general, the challenge of proteomic data mining is that the data is very complex and noisy. Currently, many projects rely on rather heuristic techniques that do not have a sound theoretical foundation.
Our aim is to give a brief review of two programs representing the current state-of-the-art in the development of proteomics software tools. They both implement a workflow starting from the mass spectrometer that ends with quantitation and identification of the peptides contained in the sample. Nevertheless they are very different in their approach and the assumptions they make.

Dagstuhl Seminar Proceedings 05471
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2006/537

2 The Trans-Proteomic Pipeline

The Trans-Proteomic Pipeline (TPP) [2] is a project of the Proteome Center at the Institute for Systems Biology in Seattle. This pipeline makes use of open XML file formats for storage of data at the raw spectral data, peptide, and protein levels. The TPP integrates other tools developed at the ISB into a coherent framework. Among these tools are *PeptideProphet* [3] which validates peptides assigned to MS/MS spectra, *XPRESS* [4] and *ASAPRatio* [5] that quantify peptides and proteins in differentially labelled samples, *Pep3D* [6] enables a view of the raw spectral data, and *ProteinProphet* [7] infers sample proteins.
An example workflow would consist of the computation of probabilities provided by PeptideProphet and ProteinProphet serving as guides for interpretation of peptide and protein identifications, respectively. These probabilities can be used to predict the false positive error rates. The error rates can be used to compare the results from different peptide identification algorithms but also for the comparison of data sets generated by different researchers. Following this refinement of the results of standard identification algorithms, quantification on the peptide and protein level can be performed by ASAPRatio or XPRESS. Results at each step can visualised by Pep3D. The software *SeachCombiner* implements a voting scheme which takes the results from several peptide search engines into consideration. It assigns a score to each search result reflecting whether the corresponding peptide was also contained in the result lists of other search engines. Figure 1 (left) shows the accuracy of PeptideProphet-computed peptide probabilities for an example data set in sliding window of 50 search results. The results show that for all three search engines, the probabilities estimated by PeptideProphet represent accurate estimates of the likelihood that search results in the data set are correct. The right plot shows the numbers of search results for an example data set filtered at a minimum PeptideProphet probability to achieve a predicted 2.5% error rate. The inset shows the numbers using Mascot results with probabilities adjusted by SearchCombiner to take into account the results of SEQUEST and COMET applied to the same data set. We can see that the application of SearchCombiner increased the number of peptides that were predicted by all three engines but also the number of peptides passing the error threshold of 2.5%. A major obstacle to the uniform proteomic analysis is the great heterogeneity of data formats. The TPP tries to deal with this problem by converting the data from different mass spectrometry instruments into a common data exchange format called mzXML. This XML implementation can deal with raw data but also picked peak data sets.
To summarize, the Trans-Proteomic Pipeline is a software platform that integrates several independent tools into a common framework. It implements visualisation tools and data exchange formats for each level of analysis. Nevertheless, the main emphasis of this project is on peptide identification and quantification. No pre-processing steps of the data are implemented but the software deals only with the data as it leaves the mass spectrometer.

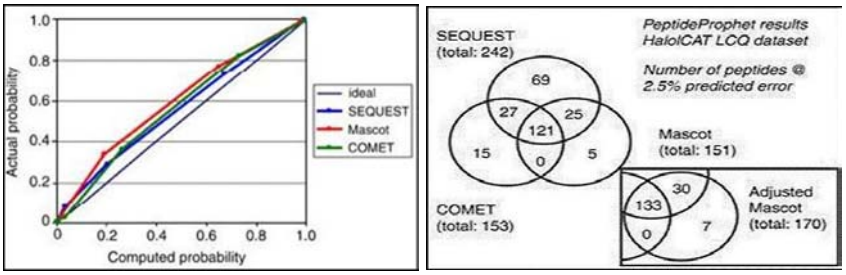

Fig. 1. Accuracy of the PeptideProphet algorithm (left) and PeptideProphet results using the estimates from several peptide identification algorithms (right).

3 Informatics platform for global biomarker discovery

In contrast to the Trans-Proteomic Pipeline, this software does implement the whole workflow from feature detection to quantification and identification of peptides [8]. First, the peaks are rounded to the nearest integer and grouped into bins of ($\pm$ 0.5 Thompson). Smoothing is performed using moving averages. A *threshold-based feature detection* is performed by declaring peaks as features that have a intensity higher than a threshold. This threshold depends on the median of the intensity on the whole data set. To be declared as feature, the intensities of the neighbouring peaks need to be higher than a certain threshold as well.

In the next step, neighbouring peaks are grouped and assigned the same ID. Alignment of peaks is performed by searching for a linear transformation that maximises the number of overlapping peaks between two data sets. To find this transformation, an accelerated random search is performed. A certain "wobble" between the peaks is allowed i.e. a peak is allowed to move (1−2% of total scan headers) in order to find the nearest adjacent peak. Finally, identification of peptides in the sample is performed using the SEQUEST algorithm [9] and quantification is performed by summing the intensities of peaks grouped in the feature detection step. According to Professor Radulovic (personal communication), no public release of this software is planned but a commercial version might be released in the near future.

4 Summary and conclusions

Both software platforms that were presented in this extended abstract are of high quality and can be considered to represent the current state-of-the-art. We noticed that it is very difficult to make comprehensive statements of the performance of different computational methods since every researcher evaluates his or her results on a different data set. Currently, there no gold standard in Computational Proteomics, no reproducible data that is available to everyone and

that could be used to compare different algorithms under the same conditions and on the same data.
If we compare the projects presented here to our own software, OpenMS, we can state that even if the overall aim is similar, the quantification and identification of peptides in a sample, the approaches chosen to achieve this aim are very different. OpenMS implements a hierarchical concept similar to [8] which includes all necessary pre-processing steps such as peak picking and feature detection. But in contrast to their work, OpenMS is available for free and published under an Open Source software licence. The main emphasis of the Trans-Proteomic Pipeline is on quantification and the refinement of results from peptide identification algorithm. It does not perform any feature detection or peak alignment of its own and is therefore not directly comparable to our own work. With OpenMS, we intend to fill a gap between commercial software and software that does not offer the whole workflow such as the Trans-Proteomic Pipeline. We expect that the importance of reliable and flexible software tools for research on proteomics will even increase and that free tools will have a advantage over their competitors.

References

1. Mann, M., Aebersold, R.: Mass spectrometry-based proteomics. Nature 422 **422** (2003) 198 – 207
2. Keller, A., Eng, J., Zhang, N., jun Li, X., Abersold, R.: A uniform proteomics ms/ms analysis platform utilizing open xml file formats. Molecular Systems Biology (2005)
3. Keller, A., Nesvizhskii, A., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. Anal Chem **74** (2002) 5383–5392
4. Han, D., Eng, J., Zhou, H., R, A.: Quantitative profiling of differentiation-induced microsomal proteins using isotopecoded affinity tags and mass spectrometry. Nat Biotechnology **19** (2001) 946–951
5. Li, X.j., Zhang, H., Ranish, J., Aebersold, R.: Automated statistical analysis of protein abundance ratios from data generated by stable isotope dilution and tandem mass spectrometry. Anal Chem **75** (2003) 6648–6657
6. Li, X.j., Pedrioli, P., Eng, J., Martin, D., Yi, E., Lee, H., Aebersold, R.: A tool to visualize and evaluate data obtained by liquid chromatography/electrospray ionization/mass spectrometry. Anal Chem **76** (2004) 3856–3860
7. Nesvizhskii, A., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem **75** (2003) 4646–4658
8. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T.G., Foss, E., Mao, Y., Emili, A.: Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Molecular and Cellular Proteomics **3** (2004) 984–997
9. Eng, J.K., McCormack, A.L., Yates, J.R.I.: An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J. Am. Soc. Mass Spectrom. **11** (1994) 976–989