OpenMS – A Framework for Quantitative HPLC/MS-Based Proteomics

Knut Reinert[1], Oliver Kohlbacher[2],Clemens Gröpl[1], Eva Lange[1], Ole Schulz-Trieglaff[1],Marc Sturm[2] and Nico Pfeifer[2]

[1] Algorithmische Bioinformatik, Fachbereich Mathematik und Informatik, FU Berlin, Takustr. 9, 14195 Berlin, Germany,
reinert@inf.fu-berlin.de
[2] Simulation biologischer Systeme, WSI/ZBIT, Eberhard-Karls-Universität Tübingen, Sand 14, 72076 Tübingen, Germany,
oliver.kohlbacher@uni-tuebingen.de

Abstract. One of the main goals of proteomics research is the discovery of novel diagnostic markers and therapeutic targets. Currently, mass spectrometry is the main platform for analyzing complex protein samples. Lately, HPLC/MS-based approaches have gained considerable interest due to their larger potential for full automation when compared to gel-based techniques. Particularly, multi-dimensional HPLC-MS methods have a great potential as a platform for differential quantification of proteins in complex mixtures. However, computational methods to analyze these automated analyses at a large scale are yet to be developed. The development of these methods should encompass new methods for data reduction, data interpretation, data management and visualization. We propose an algorithmic framework for a fully automated differential analysis of HPLC/MS samples, which goes beyond the currently established pairwise comparison of samples towards a statistically sound analysis of larger sample numbers. In this short paper we outline the framework in its current state and lay out future plans.

Keywords. computational proteomics, quantitative analysis, C++, software framework, mass spectrometry

1 Introduction

High-performance liquid chromatography (HPLC) prior to mass-spetrometric (MS) analysis has become a standard technique for high-throughput proteomics [1]. Complex peptide or protein samples can be analyzed, and in combination with labeling techniques and MS-based protein identification, differential studies can be carried out effectively. The data processing pipeline for high throughput analysis of proteins using mass spectrometry requires efficient algorithms to extract the information of interest from a large volume of data (see Fig. 1 for an example of quantitative data obtained from a mass spectrometer).

There are other ideas of software frameworks for computational proteomics. Closest to our ideas is the Trans-Proteomic Pipeline (TPP) [2]. The TPP makes

Dagstuhl Seminar Proceedings 05471
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2006/546

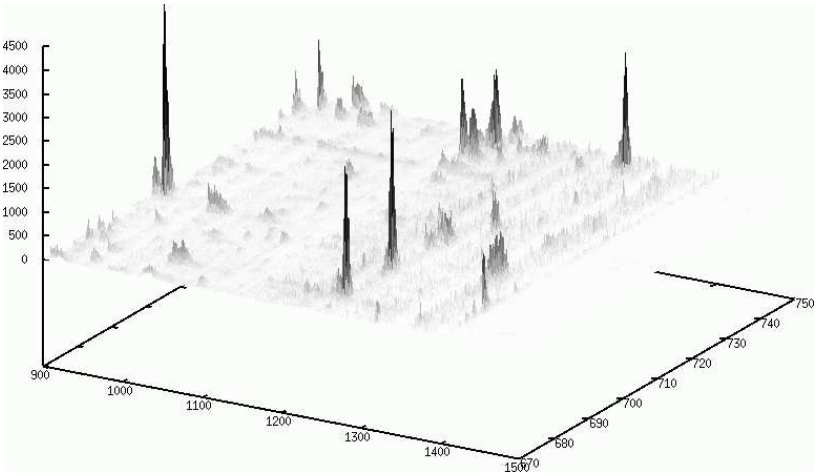

Fig. 1. Part of a raw map. The data is separated in retention time and mass over charge.

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.

OpenMS is an open source software framework implemented in C++. It provides effective data structures and algorithms for the analysis of multi-dimensional HPLC-MS data. Data processing is based on a hierarchical concept where each step reduces the amount of data by at least one order of magnitude. However, a reference to the data of earlier processing steps is kept for use in following steps. Two main data reduction steps are already implemented. The first is an enhanced wavelet-based peak picking algorithm, which includes baseline reduction and noise removal. The second is the aggregation of peaks into more complex features, e.g. isotopic patterns or collections of charge variants. The large amounts of data are managed by a database system. For import of mzData, the HUPO PSI standard [8], and various import/export formats are supported. Additional applications such as a data viewer and a workflow system for the analysis pipeline are under development. The framework has proven its usefulness for proteomics analyses of biomarkers in complex matrices in exemplary studies [9,10]. In the following we will first introduce the main components of OpenMS and close with a brief overview of ongoing work.

2 Components of OpenMS

2.1 Peak Picking

Signal processing includes baseline reduction, noise removal, and peak picking. For baseline filtering, we apply morphological methods and wavelet-based algorithms for smoothing. For peak picking, raw data points that belong to a peak have to be detected, and important features like their centroid position, height and area have to be determined (see Fig. 2 for an illustration of the peak picking process). To find the centroid positions we use the continuous wavelet transformation adapting the wavelet to the theoretical peak shape. The algorithm allows the clear separation even of largely overlapping peaks. The parameters of the best fitting asymmetric Lorentzian or hyperbolic-secant-squared functions yield valuable information about the original peak shape for further analysis. For more details see [11].

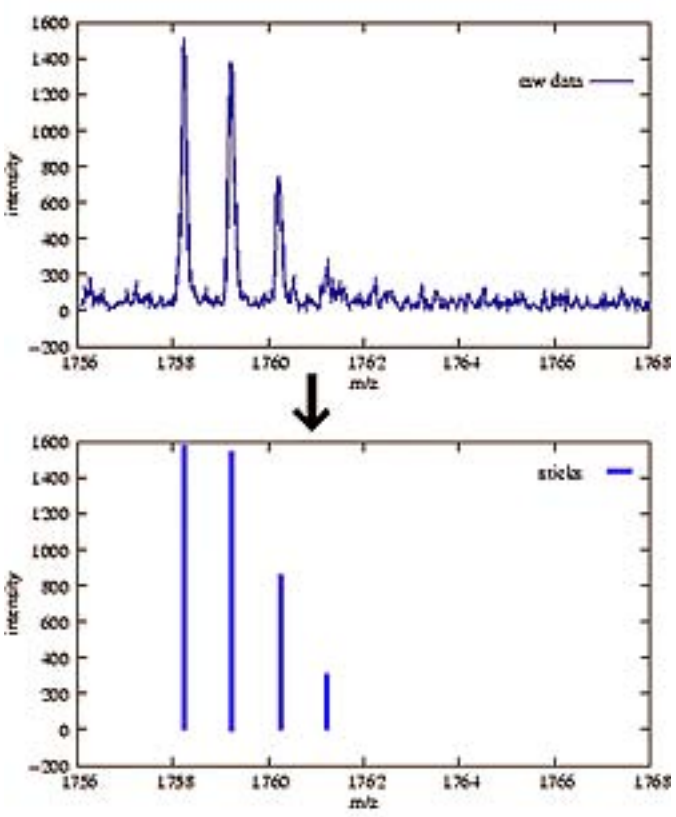

Fig. 2. Converting raw data into sticks.

2.2 Feature Finding

Feature Finding is a central concept of the OpenMS approach. We identify all peaks belonging to the same chemical species (for instance a peptide charge variant) in a multi-dimensional map, e.g. parts of several MS scans at different retention times. The data may consist of picked peaks or profile data. Sophisticated and efficient algorithms are then used to extract peptide or protein information from the LC/MS data by collecting isotopic patterns, elution profiles, charge variants, etc.

By analyzing the measured compounds we assign quality values to each dimension of description of such a *feature*. They are used in the final statistical analysis as well as to evaluate the overall quality of LC/MS measurements. The goal here to be as generic as possible, allowing us to formulate many biochemical questions as multi-dimensional search problems and solve them using similar algorithms and data structures. Fig. 3 shows the fit of a two-dimensional distribution to the measured data. A detailed description of the feature finding algorithm has been published elsewhere [9].

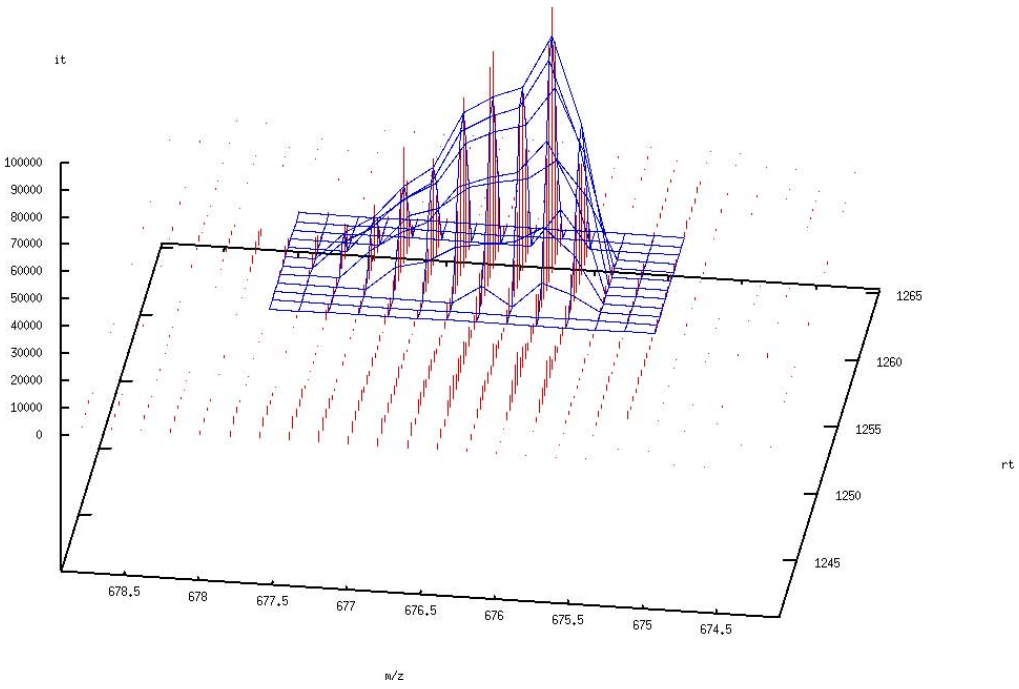

Fig. 3. Fit of the model to the raw data.

2.3 Map Mapping

Map mapping is the process in which maps of peaks or features from different measurements are superimposed for comparison. The first step is to find a transformation that moves features from one map close to corresponding features in the other one. A standard geometric hashing approach is sufficient in most cases; more complex (alignment-based) calibration schemes have also been implemented. In the second step we determine a combinatorial matching and extract groups of corresponding features for differential analysis. In our concept we can then track ambiguities and inconsistencies detected back to an earlier stage of the data reduction pipeline.

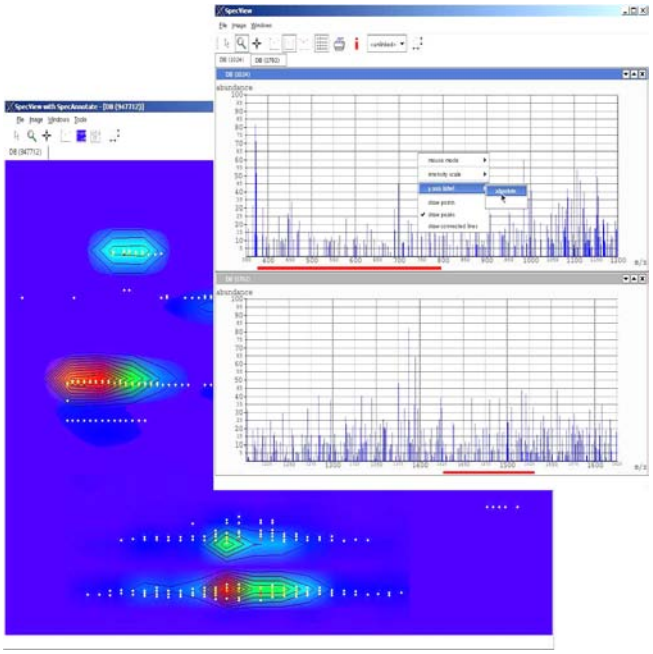

Fig. 4. Screenshot of SpecView.

2.4 Viewer

The MS data viewer *SpecView* comes with OpenMS and is an example for how classes of the OpenMS framework can be used to build up powerful applications. SpecView can visualize 1D and 2D MS spectra both from files or a database. The representation of the spectra is configurable and images can be saved or printed directly from the viewer. Fig. 4 shows two instances of SpecView. The one in the front holds two 1D spectra. In the background a 2D spectrum is displayed with a colored surface, contour lines and peaks drawn as colored dots.

2.5 Database, XML, Workflow

Large-scale laboratories need their data to be annotated, analyzed and stored. The huge amount of data makes database support an essential feature of nearly every proteomics application. OpenMS comes with database support through the QT SQL module, which allows the use of a variety of relational and object relational databases (PostgreSQL, MySQL, DB2 and many more). The database model of OpenMS is being adapted to the MIAPE and PSI-OM standard. Besides the database several file formats for loading and storing data persistently are supported. Examples of these formats are AndiMS, DTA, mzXML, Mascot result files and Sequest result files. A workflow system for combining several applications into an analysis pipeline for MS data is in progress.

3 Conlusion and Outlook

We have presented a C++ framework for the quantitative analysis of HPLC/MS data which is still under very active development. Nevertheless, many components are already functional and can be used to easily develop powerfull analysis applications. Our aim is to add more functionality to OpenMS and then conduct various analyses on HPLC-MS data from fields such as diagnostics or systems biology. OpenMS should allow the researchers to quickly prototype analysis pipelines, allow for more complex experimental setups, and reduce the dependence on vendor-supplied analysis software.

References

1. McDonald, W., Yates, J.r.: Shotgun proteomics and biomarker discovery. Dis. Markers **18** (2002) 99–105
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. Orchard, S., Hermjakob, H., Julian, R.J., Runte, K., Sherman, D., Wojcik, J., Zhu, W., Apweiler, R.: Common interchange standards for proteomics data: Public availability of tools and schema. Proteomics **4** (2004) 490–491
9. Gröpl, C., Lange, E., Reinert, K., Kohlbacher, O., Sturm, M., Huber, C., Mayr, B., Klein, C.: Algorithms for the automated absolute quantification of diagnostic markers in complex proteomics samples. In Berthold, M., Glen, R., Diederichs, K., Kohlbacher, O., Fischer, I., eds.: Proceedings of the 1st Symposium on Computational Life Sciences (CLS 2005). Volume 3695 of Lecture Notes in Bioinformatics (LNBI)., Springer (2005) 151–161
10. Mayr, B.M., Kohlbacher, O., Reinert, K., Sturm, M., Gröpl, C., Lange, E., Klein, C., Huber, C.G.: Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. J. Proteome Res. **5** (2006) 414–421
11. Lange, E., Gröpl, C., Reinert, K., Kohlbacher, O., Hildebrandt, A.: High accuracy peak-picking of proteomics data using wavelet techniques. In: Proceedings of the 11th Pacific Symposium on Biocomputing (PSB-06). (2006) 243–254