08101 Executive Summary and Abstracts Collection
Computational Proteomics
— Dagstuhl Seminar —

C. Huber[1], O. Kohlbacher[2], M. Linial[3], K. Marcus[4] and K. Reinert[5]

[1] Saarland University, DE
christian.huber3@sbg.ac.at

[2] Tübingen University, DE
kohlbach@informatik.uni-tuebingen.de

[3] Hebrew University, Jerusalem, IL
michall@cc.huji.ac.il

[4] Ruhr-University Bochum, DE
Katrin.Marcus@rub.de

[5] Free University Berlin, DE
reinert@inf.fu-berlin.de

Abstract. The second Dagstuhl Seminar on *Computational Proteomics* took place from March 3rd to 7th, 2008 in Schloss Dagstuhl–Leibniz Center for Informatics. This highly international meeting brought together researchers from computer science and from proteomics to discuss the state of the art and future developments at the interface between experiment and theory. This interdisciplinary exchange covered a wide range of topics, from new experimental methods resulting in more complex data we will have to expect in the future to purely theoretical studies of what level of experimental accuracy is required in order to solve certain problems. A particular focus was also on the application side, where the participants discussed more complex experimental methodologies that are enabled by more sophisticated computational techniques. Quantitative aspects of protein expression analysis as well as posttranslational modifications in the context of disease development and diagnosis were discussed. The seminar sparked a number of new ideas and collaborations and has resulted in several joint grant applications and paper submissions.

This paper describes the seminar topics, its goals and results. The executive summary is followed by the abstracts of the presentations given. Links to extended abstracts or full papers are provided, if available.

Keywords. bioinformatics, biomedicine, proteomics, analytical chemistry.

Dagstuhl Seminar Proceedings 08101
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2008/1784

1 Executive summary

1.1 Goals and structure of the workshop

The field of Computational Proteomics has grown rapidly and gained a lot of momentum over the last years. Computational Proteomics was previously a field with a small and specialized community. Over the last few years, however, it has been recognized by experimental groups, that the analysis of the increasingly complex proteomics studies has become intractable without efficient algorithms implemented in easy-to-use tools. Conversely, the computer science and bioinformatics communities started to realize the wealth of interesting problems in this area. Both sides are thus eager to come together and work on these problems. To initiate this close collaboration of researchers from different fields (biology, medical sciences, biochemistry, analytical chemistry, bioinformatics, and computer science), we need opportunities to bring both communities together in an inspiring and relaxed atmosphere. Past experience from our 2005 seminar has shown that Dagstuhl is an ideal place for this.[6]

Currently, computational proteomics faces a number of challenges. The increasing speed and accuracy of the new instrument generation yields datasets that are up to an order of magnitude larger than datasets seen a few years ago. This implies new algorithmic techniques and data analysis capabilities. The computer science problems to be tackled range from data management, over optimization problems, to machine learning. On the experimental side, the development of high mass accuracy and better separation techniques pose new challenges.

The seminar brought together a mixed audience from proteomics and bioinformatics: At the beginning we took a pool showing that there were 10 people who declared themselves as “wet lab” and 26 as “computer science”. Moreover, there were 21 people “holding a PhD degree” and 15 “working on it”. We are happy to see that the communities really grow together.

The talks and discussions were arranged on a daily basis featuring related topics:

- **Systems Biology and novel experimental techniques.**

  Important issues in the emerging field of systems biology were discussed including the elucidation of signaling pathways and networks and its requirements for the implementation of data analysis. There is an ongoing trend to integrate all available sources of data. New and emerging experimental as well as computational techniques were described to bring both communities to the state of the art.

[6] The web page of the Dagstuhl Seminar “Computational Proteomics” held in 2005 is http://www.dagstuhl.de/05471. The present one is http://www.dagstuhl.de/08101.

R. Zubarev presented an interpretation procedure that includes a database of genes, molecules and their interactions. His approach combines expression proteomics data with gene mapping, transcription factor analysis and keynode analysis.

M. Lappe's contribution raised a lively discussion about the relation of proteomics and systems biology.

H. Hermjakob developed a perspective for sharing of all publicly available proteomics data, why this is in fact necessary for validation, and what the obstacles for realization are. Will experimentalists appreciate direct benefits from providing open access to their data, or does it have to become a mandatory condition for publishing?

H. Schlüter reported his experiences from deciphering the physiological roles of proteases analyzed by LC-MS based methods using annotated protein databases. He gave examples of problems arising from synonyms, inaccurate assignment of functions and exact chemical composition, as well as missing links towards the underlying experimental data.

K. Melchior compared the established bottom up approach for proteome analyses with an alternative method in which intact proteins are separated before digestion. The new method was shown to yield a higher sequence coverage.

T. A. Hansen presented tools for the analysis of large quantitative phosphoproteomics datasets using kinase recognition motifs and gene ontology terms.

- **Quantitative analysis.**

  The second day had a focus on the efficient treatment of algorithmic problems for the quantitative analysis of peptides and proteins, in particular the LC-MS map alignment problem, as well as new lab techniques for differential analysis.

  C. Gröpl presented ongoing work from a comparison of computational tools for the map alignment problem. He raised a discussion about the relative importance of retention time correction compared to one-to-one assignment on the level of peptidic features.

  O. Schulz-Trieglaff went on with a comparison of computational tools for the feature detection step in LC-MS datasets. He developed guidelines how these algorithms could be benchmarked and their results validated.

  J. Cottrell explained the approach taken by Matrix Science Ltd., who are in the process of implementing support for a wide range of quantitation methods in Mascot.

  K. Podwojski has developed a new map alignment algorithm which avoids certain pitfalls using powerful statistical methods.

J. Vandekerckhove gave an overview of combined fractional diagonal chromatography, a versatile two-step chromatographic approach that can be applied to a large variety of peptide subsets. He also pointed out key features of ms_lims, a mass-spectrometry oriented laboratory information management system (LIMS) developed in their lab.

M. Askenazi presented mzAPI, a universal application programming interface that abstracts the layout of multiple file formats.

R. Bischoff is investigating disease-related changes in body fluids using mass spectrometry and liquid chromatography. He presented recent work on the alignment of complex, highly variable data sets in the retention time dimension. The influence of pre-analytical parameters on the overall proteomic profile was also discussed.

R. Zahedi reported from a recent analysis of the human platelet phosphoproteome. Interesting insights were achieved using a two-pronged strategy based on the enrichment of phosphopeptides by immobilized metal-ionaffinity chromatography (IMAC) and strong cation-exchange chromatography (SCX), coupled with subsequent analysis by nano-liquid chromatography tandem mass spectrometry (nano-LC-MS/MS) or precursor ion scanning.

O. Kohlbacher gave an overview of TOPP, the OpenMS proteomics pipeline. TOPP is based on OpenMS, an open-source software framework written in C++. TOPP can facilitate many computational tasks from a proteomics pipeline. The design of TOPP/OpenMS was made such that programming new components is easy, without sacrificing performance.

– **Identification.**

Currently mass spectrometry is the workhorse for peptide and protein identification. Starting from peptide fragment spectra it is common to identify the underlying peptides either by using fast searches in peptide or transcript databases, or by de novo prediction. We had a whole day of the seminar devoted to identification.

Future approaches will collect molecular information other than molecular mass or fragmentation patterns, such as retention time, isoelectric point, peptide detectability. This will further increase the speed of identification and minimize false positives, but it also requires efficient algorithmic tools to acquire, combine and evaluate such data.

N. Pfeifer presented machine learning techniques for improving peptide identification. His results are based on a new kernel function for support vector machines. The method was applied to peptide detectability and retention time prediction.

A. Zerck addressed the selection of precursor ions in data dependent acquisition of tandem MS spectra within an offline MALDI setting. An iterative procedure to select the precursor ions based upon identification results from earlier steps compares favorably to the theoretically optimal solution computed by an integer linear program.

M. Müller reported about ongoing work to set up a pipeline to accurately detect and quantify peptides and their posttranslational modifications (PTMs).

B. Küster gave a talk about computational prediction of proteotypic peptides for quantitative proteomics. Only a few so-called proteotypic peptides are repeatedly and consistently identified for any given protein present in a mixture. These can now be predicted with more than 85% cumulative accuracy.

S. Böcker analyzed a method for de novo identification of metabolites using tandem mass spectra. The molecular formula can be computed by reconstructing the most likely fragmentation tree. The method works surprisingly well.

N. Bandeira described recent applications of the spectral network approach in multistage mass spectrometry. A rigorous probabilistic framework results in accurate de novo peptide sequencing from multistage mass spectra and improved interpretation of spectral networks.

W. Lehmann discussed the versatility of information contained in small peptide fragments. In many cases, these are sufficient for assigning the location of a peptide within a known protein sequence. Modified residues give rise to specific reporter immonium ions. Quantitation evaluation of small fragment ions is more accurate compared to analysis of intact molecular ions due to the strongly reduced isotopic overlap.

K. Marcus discussed transpeptidation products that are present in virtually any proteome analysis by LC-MS/MS that uses trypsin for protein digestion, but rarely recognized due to available database search software. Evidence for the frequency of different types of side reactions was given, but today the full amount of peptides generated by transpeptidation is not clear.

- **Pathway analysis and biomarkers.**

  Proteomics data sets are large, multi-dimensional, and highly complex. Their visualization and analysis thus requires new approaches. At the same time, these new techniques allow interesting new inferences.

  J. Gobom described an approach to detect biomarkers for Alzheimer's disease in cerebrospinal fluid that uses immunoprecipitation in combination with mass spectrometry. He pointed out several shortcomings of current software.

  C. Stephan drew our attention to 'balancer' proteins, which are regulated up or down as a consequence of system stoichiometry, but not specific to a particular disease. They can be determined by comparing multiple proteomics studies, and are likely to be hubs of protein regulatory networks. He also presented an easy-to-use software tool to generate decoy databases for use with MS/MS search engines.

M. Linial presented PANDORA, a statistical visualization tool that allows fast and intuitive knowledge extraction from any set of identified peptides and proteins. The tool supports various annotation sources, including SwissProt and GO (gene ontology).

1.2 Participants and topics

Experts and young scientists working in computer science, (bio)informatics, proteomics, analytical chemistry, medical research came from Europe as well as the USA.

The following scientists participated in the seminar:

- Manor Askenazi, Sudarsky Center for Computational Biology (SCCB), Jerusalem, IL
- Nuno Bandeira, University of California, San Diego, USA
- Andreas Bertsch, Tübingen University, DE
- Chris Bielow, Free University Berlin, DE
- Rainer Bischoff, University of Groningen, NL
- Sebastian Böcker, Jena University, DE
- John Cottrell, Matrix Science Ltd., London, UK
- Jens Decker, Bruker Daltonik GmbH, Bremen, DE
- Johan Gobom, Sahlgrenska Academy, University of Gotenborg, SE
- Clemens Gröpl, Free University Berlin, DE
- Thomas Aarup Hansen, University of Southern Denmark, Odense, DK
- Henning Hermjakob, European Bioinformatics Institute (EBI), Cambridge, UK
- Christian Huber, Universität Salzburg University, AT
- Hans-Michael Kaltenbach, Institut Pasteur, Paris, FR
- Oliver Kohlbacher, Tübingen University, DE
- Sebastian Kühner, European Molecular Biology Laboratory (EMBL), Heidelberg, DE
- Bernhard Küster, Technical University Munich, DE
- Eva Lange, Beatson Institute for Cancer Research, Glasgow, UK
- Michael Lappe, Max Planck Institute for Molecular Genetics, Berlin, DE
- Wolf D. Lehmann, German Cancer Research Center (DKFZ), Heidelberg, DE
- Michal Linial, Sudarsky Center for Computational Biology (SCCB), Jerusalem, IL
- Katrin Marcus, Ruhr University Bochum, DE
- Katja Melchior, Saarland University, DE
- Markus Müller, Swiss Institute of Bioinformatics (SIB), CH
- Sven Nahnsen, Tübingen University, DE
- Anton Pervukhin, Jena University, DE

- Nico Pfeifer, Tübingen University, DE
- Katharina Podwojski, Dortmund University, DE
- Jörg Rahnenfuehrer, Dortmund University, DE
- Florian Rasche, Jena University, DE
- Knut Reinert, Free University Berlin, DE
- Hartmut Schlüter, Charité, Berlin, DE
- Ole Schulz-Trieglaff, Free University Berlin, DE
- Benno Schwikowski, Institut Pasteur, Paris, FR
- Albert Sickmann, Rudolf-Virchow-Zentrum, Würzburg, DE
- Christian Stephan, Ruhr University Bochum, DE
- Marc Sturm, Tübingen University, DE
- Joel Vandekerckhove, Ghent University and Flanders Institute for Biotechnology (VIB), BE
- Mathias Vandenbogaert, Institut Pasteur, Paris, FR
- René Zahedi, Rudolf-Virchow-Zentrum, Würzburg, DE
- Alexandra Zerck, Max Planck Institute for Molecular Genetics, Berlin, DE
- Roman Zubarev, University of Uppsala, SE

1.3 Detailed program

Monday 2008-03-03:
Introduction

Morning session: 09:00-12:10 (190 minutes) Chair: *Kohlbacher*

- *Organizers:*
  Welcome and introduction – the idea of Dagstuhl seminars (15).
- *All participants:*
  Introduction (60)
- *Break* (20)
- *Michael Lappe:*
  Systems Biology = Networks & Structures? (45)
- *Roman Zubarev:*
  Proteomics goes intelligent: identification of activated pathways from proteomics expression data (45)

Afternoon session: 14:15-17:55 (220 minutes) Chair: *Huber*

- *Henning Hermjakob:*
  Carrots and sticks (45)
- *Hartmut Schlüter:*
  Experiences of a 'protease hunting lab' with protein- and knowledge databases (45)
- *Break* (30)
- *Katja Melchior:*
  Comparison of an alternative approach for proteome research with the common bottom up method (35)
- *Thomas Aarup Hansen:*
  GOEater and KinaseEater: computational tools for meta-analysis of large-scale quantitative phosphoproteomic data sets (25)

Tuesday 2008-03-04:
Quantitative Proteomics

Morning session: 09:00-12:10 (190 minutes) Chair: *Reinert*

- *Clemens Gröpl:*
  LC-MS/MS data processing in OpenMS (30)
- *Ole Schulz-Trieglaff:*
  Benchmarking algorithms for label-free quantification (30)
- Break (20)
- *John Cottrell:*
  An integrated approach to protein quantitation software (45)
- *Katja Podwojski:*
  A retention-time alignment algorithm for LC/MS data (40)

Afternoon session: 14:15-17:55 (220 minutes) Chair: *Linial*

- *Joel Vandekerckhove:*
  COmbined FRActional DIagonal Chromatography (COFRADIC) (45)
- *Manor Askenazi:*
  mzAPI: Common APIs as a viable alternative to universal file formats in mass spectrometry (30)
- *Break* (30)
- *Rainer Bischoff:*
  Biomarker discovery by LC-MS: data processing and analysis (45)
- *René Zahedi:*
  Phosphoproteome of resting human platelets (30)

Wednesday 2008-03-05:
Discussion round and excursion

Morning session: 09:00-12:10 (190 minutes) Chair: *Schlüter*

- *Oliver Kohlbacher:*
  OpenMS and TOPP – Software for Computational Proteomics (45)
- Discussion round

Afternoon: Visit to Trier and wine tasting

Thursday 2008-03-06:

Protein/Peptide Identification

Morning session: 09:00-12:10 (190 minutes) Chair: *Marcus*

- *Nico Pfeifer:*
  Improving identification with new machine learning techniques (30)
- *Alexandra Zerck:*
  Data-dependent MS/MS analysis (25)
- *Break* (30)
- *Markus Müller:*
  A pipeline to detect and quantify posttranslational modifications (40)
- *Bernhard Küster:*
  Computational prediction of proteotypic peptides for quantitative proteomics (45)

Afternoon session: 14:15-17:55 (220 minutes) Chair: *Huber*

- *Sebastian Böcker:*
  De novo identification of metabolites by analyzing tandem mass spectra (40)
- *Nuno Bandeira:*
  Multi-spectra peptide sequencing and its applications to multistage mass spectrometry (45)
- *Break* (30)
- *Wolf Lehmann:*
  The diagnostic power of small peptide fragments for peptide ends, modifications, composition and quantification (45)
- *Katrin Marcus:*
  Tryptic transpeptidation products observed in proteome analysis by LC-MS/MS (45)

Friday 2008-03-07:

Pathway analysis and biomarkers

Morning session: 09:00-12:10 (190 minutes) Chair: *Kohlbacher*

- *Johan Gobom:*
  Finding CSF protein markers in Alzheimer's disease (45)
- *Christian Stephan:*
  Decoy database advantages and protein balancing for understanding the complexity of life (45)
- *Break* (30)
- *Michal Linial:*
  Family tree of all protein sequences: recovering hidden biological knowledge (45)
- *Closing discussion* (50)

Afternoon: End of seminar

1.4 Workshop conclusion

One of the reasons to organize this Dagstuhl seminar about *Computational Proteomics* was that computer scientists should understand how the data for proteome analysis are generated and what their implications are, while experimental scientists should know how the data can be evaluated and validated. The talks and (more importantly but less easily to document) the discussions showed that this goal was indeed achieved. The seminar set off new initiatives for collaboration between experimental and computational sciences that will result in novel, rugged and fully automatable total analysis systems for proteome characterization.

The outcome or the seminar was a better understanding of the expectations, needs, and possibilities both of experimental and bioinformatic tools for proteome analysis. Several new cooperations between different groups could be established leading to joint grant proposals involving several participants of the workshop (EU project NAPIRA (Reinert, rejected), DFG research unit Algorithms for Mass Spectrometry (Kohlbacher, rejected), BMBF MedSys (Sickmann, granted)) At least half a dozen joint manuscripts that were started in Dagstuhl are currently at least in preparation or under review.

The workshop on Computational Proteomics was a full success, as has been confirmed by its participants. Bringing together scientist from different communities – from computer science and life sciences – turned out to be fruitful indeed. Traditionally, proteomics and bioinformatics/computer science are mostly disjoint communities with separate meetings and conferences. The chance to get insights into the problems and challenges both of the experimental and computational world, the need to learn and understand the idiosyncratic ‘languages’ and ‘vocabulary’ of the different disciplines was well appreciated by the attendants.

Validation of proteomics data generation and evaluation was spotted as one of the most challenging issues in the application of proteomics as a technology for clinical diagnosis and monitoring. Participants from the two communities were exposed to new ideas, concepts, and techniques – both experimentally and computationally – they were not previously aware of. These ideas were then discussed over a glass of wine or two until late at night. The workshop produced a number of personal contacts which was positively remarked by the participants. In addition to the interaction and personal contacts of the attendants, the quiet atmosphere of the location also allowed ample time for developing new ideas for solving proteomic challenges.

In conclusion, the workshop was very successful. It sparked interesting discussions, research collaborations, several joint grant proposals (e.g. for the BMBF program QuantPro), and joint publications. Other publications sparked by the seminar will certainly follow in the near future. The seminar also initiated the implementation of a webpage for interchanging
proteomics data (www.computationalproteomics.net). The success of the sem-

inar and the positive feedback of the participants encourage us to organize a follow-up for this style of meeting.

Acknowledgment. We thank Clemens Gröpl and Chris Bielow for assistance with the preparation of this executive summary and formatting the abstract collection.

2 Abstracts of talks

2.1 mzAPI: Common APIs as a Viable Alternative to Universal File Formats in Mass Spectrometry

Manor Askenazi (SCCB - Sudarsky Center for Computational Biology, IL)

The adoption of common file formats in proteomics implies a concomitant abandonment of native data systems and their efficient support of fundamental data access patterns. Based on a comparison of data structure and access patterns across disparate scientific fields, we propose that a universal application programming interface (API) represents a more viable approach for mass spectrometry data access and portability. Here we define a minimal, redistributable API (mzAPI), incorporated into an open-source application (multiplierz), that leverages native features embedded in manufacturers' data systems and obviates the need for surrogate files. We show that mzAPI abstracts the layout of multiple native file formats and provides: 1.) integrity and re-use of proprietary data systems; 2.) rational integration of open-source and commercial tools; 3.) accessible and comprehensive customization for desktop users and data pipeline administrators.

Keywords: MzAPI, file formats, API, mass spectrometry

2.2 Multi-Spectra Peptide Sequencing and its Applications to Multistage Mass Spectrometry

Nuno Bandeira (University of California - San Diego, US)

Despite a recent surge of interest in database-independent peptide identifications, accurate de novo peptide sequencing remains an elusive goal. While the recently introduced spectral network approach resulted in an accurate peptide sequencing in low-complexity samples, its success depends on the chance presence of spectra from overlapping peptides. On the other hand, while multistage mass spectrometry (collecting multiple MS^3 spectra from each MS^2 spectrum) can be applied to all spectra in a complex sample, there are currently no software tools for de novo peptide sequencing by multistage mass spectrometry. We describe a rigorous probabilistic framework for analyzing spectra of overlapping peptides and show how to apply it for multistage mass spectrometry. Our software results in both accurate de novo peptide sequencing from multistage mass spectra (despite the inferior quality of MS^3 spectra) and improved interpretation of spectral networks. We further study the problem of de novo peptide sequencing with accurate parent mass (but inaccurate fragment masses), the protocol that may soon become the dominant mode of spectral acquisition. Most existing peptide sequencing algorithms (based on the spectrum graph approach) do not track the accurate parent mass and are thus not equipped for solving this problem. We describe a de novo peptide sequencing algorithm aimed at this experimental protocol and show that it improves the accuracy of both tandem and multistage mass spectrometry.

Keywords: Multistage mass spectrometry; de novo; peptide sequencing; accurate parent masses

Joint work of: Bandeira, Nuno; Olsen, Jesper V; Mann, Matthias; Pevzner, Pavel A

Full Paper: http://proteomics.bioprojects.org/

2.3 Biomarker Discovery by LC-MS: Data Processing and Analysis

Rainer Bischoff (University of Groningen, NL)

The lecture will focus on the processing and analysis of data generated by the analysis of body fluids (serum, urine) by LC-MS combined with dedicated sample preparation. Particular emphasis will be placed on alignment of complex, highly variable data sets in the retention time dimension using mass spectrometric information [1]. Another area of interest that will be touched upon relates to the assessment of the influence of pre-analytical parameters on the overall proteomic profile. This is work in progress and I am looking forward to discussions and ideas on this topic.

All our work has the goal of discovering disease-related changes in body fluids within a sea of unrelated variations due to analytical and/or biological variability. Disease areas of interest are cancer (cervical, prostate), neurological disorders (multiple sclerosis) and inflammatory disorders (chronic obstructive pulmonary disease (COPD)). The hyperlink below will lead you to publications from recent years from our group.

[1] Suits F, Lepre J, Du P, Bischoff R, Horvatovich P.: Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. Anal. Chem. 2008 May 1;80(9):3095-104. PMID: 18396914.

Keywords: Biomarker, proteomics, time alignment, pre-analytical parameters, COW, CODA

Joint work of: Horvatovich, Peter; Christin, Christin; Govorukhina, Natalia; Kemperman, Ramses; Reijmers, Theo; Suits, Frank; Bischoff, Rainer

Full Paper: http://www.rug.nl/farmacie/onderzoek/basiseenheden/bioanalyseentoxicologie/publications

2.4 Towards de novo identification of metabolites by analyzing tandem mass spectra

Sebastian Böcker (Universität Jena, DE)

Mass spectrometry is among the most widely used technologies in proteomics and metabolomics. For metabolites, de novo interpretation of spectra is even more important than for protein data, because metabolite spectra databases cover only a small fraction of naturally occurring metabolites. In this work, we analyze a method for fully automated de novo identification of metabolites from tandem mass spectra. Mass spectrometry data is usually assumed to be insufficient for identification of molecular structures, so we want to estimate the molecular formula of the unknown metabolite, a crucial step for its identification. This is achieved by calculating the possible formulas of the fragment peaks and then reconstructing the most likely fragmentation tree from this information. We present tests on real mass spectra showing that our algorithms solve the reconstruction problem suitably fast and provide excellent results: For all 32 test compounds the correct solution was among the top five suggestions, for 26 compounds the first suggestion of the exact algorithm was correct.

[1] Sebastian Böcker and Florian Rasche: Towards de novo identification of metabolites by analyzing tandem mass spectra Bioinformatics, 24:I49-I55, 2008.
Proc. of European Conference on Computational Biology (ECCB 2008).
doi:10.1093/bioinformatics/btn270

Keywords: Tandem mass spectrometry, metabolomics, de novo interpretation

Joint work of: Böcker, Sebastian; Rasche, Florian

Full Paper: http://bio.informatik.uni-jena.de/downloads/BoeckerRasche_TowardsDeNovoIdentification_Bioinf_2008.pdf

2.5 An Integrated Approach to Protein Quantitation Software

John Cottrell (Matrix Science Ltd., GB)

We are in the process of implementing support for a wide range of quantitation methods in Mascot.

Some methods can be implemented within the Mascot search engine, and reported alongside the search results, because all the required information is contained within the peak list. Examples in this category are iTRAQ and emPAI. Other methods, such as SILAC, ICPL, ICAT, 18O, and metabolic labelling, require access to the raw data, so that signals can be identified and integrated across multiple scans. In our approach, these methods require the raw data to be processed by Mascot Distiller. After the database search, the results are returned to Mascot Distiller for the generation of a quantitation report.

Particular attention has been paid to the following aspects:

(i) Maintaining a simple user interface by encapsulating the parameters required to define each quantitation experiment into named quantitation methods.

(ii) An unusual degree of flexibility in the way a quantitation method is defined. The number of components, the nature of the components, (the presence of an isotope tag, the elemental composition of the peptide, a reporter ion mass value, or whatever), which ratios are to be reported, etc.

(iii) Generic 'impurity' corrections for isotope distribution overlap and incomplete enrichment

(iv) Accurate peak areas by fitting calculated isotope distributions to the experimental spectra

(v) Support for essentially all of the binary file formats used in protein mass spectrometry, together with mzXML

(vi) The correct application of appropriate statistical methods

We will give examples of processing data from a variety of quantitation experiments and discuss some of the challenges that lie ahead.

Keywords: Quantitation database search

Joint work of: Cottrell, John; Creasy, David

2.6 Finding CSF protein markers in Alzheimer's Disease

Johan Gobom (Sahlgrenska Academy, University of Gotenborg, SE)

There is a need for new biomarkers as diagnostic tools for Alzheimer's disease. Immunoprecipitation (IP) in combination with MS enables the detection of low-abundant proteins in CSF, inaccessible to generic proteomic approaches. IP-MS analysis of known Alzheimer-related proteins, previously analyzed by immunoassays, reveals many modified forms, which may reflect disease processes. To make full use of such data new bioinformatics tools are required.

Joint work of: Gobom, Johan; Portelius, Erik; Zetterberg, Henrik; Blennow, Kai

2.7 LC-MS/MS data processing in OpenMS

Clemens Gröpl (FU Berlin, DE)

Two basic steps in the analysis of liquid chromatography-mass spectrometry data sets are the detection of two-dimensional features and the alignment of features across data sets, e.g. to be used for differential quantification.

We will focus on the map alignment problem. We will explain the main ideas from of our algorithm, which is based on pose clustering, a powerful method from computational geometry. We will also discuss the trade-offs involved in the proper choice of a family of warping functions for the correction of retention times. We are currently working on a comparison of map alignment programs on benchmark data sets. We found that the methods which are being used to find a detailed one-to-one assignment on the level of individual features are at least as important as the correction of retention times alone.

All algorithms are implemented in the open-source software framework OpenMS.

Keywords: Liquid chromatography, mass spectrometry, map alignment, data processing.

Full Paper: http://www.biomedcentral.com/1471-2105/9/375

Full Paper: http://dx.doi.org/10.1186/1471-2105-9-163

Full Paper: http://dx.doi.org/10.1093/bioinformatics/btm209

See also: http://www.openms.de

2.8 GOEater and KinaseEater: Computational tools for meta-analysis of large-scale quantitative phosphoproteomic datasets

Thomas Aarup Hansen (University of Southern Denmark - Odense, DK)

Functional phosphoproteomics is a rapidly growing field, mainly due to the development of phosphopeptide-selective sample preparation methods and sensitive and specific high-performance mass spectrometry techniques. One of the current bottlenecks in phosphoproteomics is data analysis subsequent to mass spectrometry experiments. In particular, the validations of phosphorylation site assignments made by MS/MS and automated database search engines and the subsequent extraction of biologically meaningful information from quantitative mass spectrometry experiments are time-consuming and challenging tasks. We have developed two computational tools that allow extraction and meta-analysis of large scale phosphoproteomics datasets based on (i) detection and statistical analysis of kinase recognition motifs present in the dataset of observed phosphorylation sites, and (ii) extraction of Gene Ontology terms associated with highly regulated phosphorylation events. These tools enable the identification and classification of kinases and biological processes involved in cellular mitogen activated processes as demonstrated by analysis of a large-scale quantitative phosphoproteomic dataset obtained by analysis of the yeast pheromone response.

Keywords: Phosphoproteomics, data analysis, neural networks, gene ontology

Joint work of: Ingrell, Christian Ravnsborg, Hansen, Thomas Aarup, Jensen, Ole Nörregaard

2.9 Of carrots and sticks

Henning Hermjakob (EBI - Cambridge, GB)

The landscape of publicly available proteomics data is still deeply fractured; small, green valleys of publicly available and useful data are still surrounded by mountains of inaccessible data and deserts of accessible, but useless data.

I'd like to initiate a discussion on possible carrots and sticks to encourage data producers to make their data publicly available. Other areas of molecular biology have mostly used the big stick of requiring a database accession number prior to publication of genomic or protein structural data. However, currently this stick is little more than a twig in proteomics, a modest encouragement rather than a mandatory condition for publishing. Are there carrots we can use instead, directly benefits for experimentalists providing open access to their data? And how are we extending this approach in the future? Will we need bundles of sticks and carrots to ensure public access to the data from large, multi-omics studies?

[1] Proteomics data validation: why all must provide data. Martens L, Hermjakob H. Mol Biosyst [2007 Aug (3)]:518-522

Keywords: Databases, Proteomics Standards, Data availability

Joint work of: Hermjakob, Henning

Full Paper: http://www.ebi.ac.uk/citexplore/citationDetails.do?externalId=17639125&dataSource=MED

2.10 OpenMS and TOPP: rapid prototyping of pipelines and software systems

Oliver Kohlbacher

Analysis of large- and small-scale proteomics datasets is a major hurdle for many labs. Bringing computational tools to the lab bench is thus an important task for everyone in computational proteomics. OpenMS is a software framework in C++ for rapid software prototyping. It provides core data structures and a collection of algorithms for the analysis of MS(/MS) and LC-MS(/MS) data. While OpenMS is intended for software developers, TOPP – The OpenMS Proteomics Pipeline – is intended for users on the application side. TOPP is a collection of stand-alone tools to construct data analysis pipelines.

[1] Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Marc Sturm: TOPP - the OpenMS proteomics pipeline, Bioinformatics 2007 23(2):e191-e197; doi:10.1093/bioinformatics/btl299.

Joint work of: Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Marc Sturm

Full Paper: http://dx.doi.org/doi:10.1093/bioinformatics/btl299

See also: http://www.openms.de

2.11 Computational prediction of proteotypic peptides for quantitative proteomics

Bernhard Küster (Interdisciplinary Protein Analysis Group, Technical University Munich, DE)

Mass spectrometry based quantitative proteomics has become an important component of biological and clinical research. Although such analyses typically assume that a protein's peptide fragments are observed with equal likelihood, only a few so-called 'proteotypic' peptides are repeatedly and consistently identified for any given protein present in a mixture. Using >600,000 peptide identifications generated by four proteomic platforms, we empirically identified >16,000 proteotypic peptides for 4,030 distinct yeast proteins. Characteristic physicochemical properties of these peptides were used to develop a computational tool that can predict proteotypic peptides for any protein from any organism, for a given platform, with >85% cumulative accuracy. Possible applications of proteotypic peptides include validation of protein identifications, absolute quantification of proteins, annotation of coding sequences in genomes, and characterization of the physical principles governing key elements of mass spectrometric workflows (e.g., digestion, chromatography, ionization and fragmentation).

[1] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. Nat Biotechnol. 2007 Jan;25(1):125-31. doi:10.1038/nbt1275

Full Paper: http://www.nature.com/nbt/journal/v25/n1/abs/nbt1275.html

See also: http://www.wzw.tum.de/bioanalytik/

2.12 Systems Biology = Networks & Structures?

Michael Lappe (MPI für Molekulare Genetik, DE)

Proteomics forms a significant part of the technological basis of what might be called ”Systems Biology” today. However, there is considerable confusion about what Systems Biology actually entails. A common denominator seems to be the holistic view of the (awesome!) complexity of biological systems ”putting everything back together again” . The description of Protein-Protein Interaction (PPI) Networks on a large scale had a huge impact in pushing ”network thinking” forward in molecular biology, considering phenotypic effects as emergent properties of the underlying molecular networks. As valuable as PPI Networks are as an integrative platform of a plethora of other biological data (like genome-wide sequence data), their incompleteness and intrinsic error rates make structural and functional inference difficult to obtain. Here I propose to turn to protein structures as a source for biological network data. Structure databases like the PDB are among the oldest and most reliable resources in bioinformatics. Representing structures as networks of interacting amino acids (so-called Residue Interaction Graphs or RIGs), generates networks similar in architecture to large-scale PPI networks. In this context, we can now attempt to quantify the effect of network perturbations on the resulting structure as an emergent property (= phenotype). The general insights into biological networks and the resulting algorithms might prove to be applicable across disciplines to future proteomics data.

2.13 The Diagnostic Power of Small Peptide Fragments for Peptide Ends, Modifications, Composition and Quantification

Wolf D. Lehmann (DKFZ - Heidelberg, DE)

The versatility of information contained in small peptide fragments is discussed:

Peptide ends: y1 and b2 ions are present in most CID spectra of peptides with average to high abundance. Using proteases with C-terminal specificity (e.g. trypsin, chymotrypsin, Lys-C), the variation of y1 ion is highly restriced facilitating their recognition. The variability of b2 ions is restricted to about 400 cases. In analogy, using a protease with N-terminal specificity, such as AspN, the variety of b2 ions is restricted to about 40 (cleavage at D and E). In addition AspN b2 ions contain a minimal sequence information including the direction. Often, b2/b3 and y1/y2 ion pairs are present. These provide a di- or tripeptide sequence information including the sequence direction. This information is in general unique for assigning the location of a peptide within a known protein sequence. It is proposed to rank these key fragments with a higher score compared to internal sequence ions, since their occurrence shows a minor (and predictable) dependence on the peptide sequence. N-terminal acetylation and myristoylation leads to b1 ions, which specifically identify the N-terminus. In MS/MS spectra with high signal-to-noise ratio internal dipeptide ions frequently can be observed, which support peptide identifications.

Peptide composition and modification : Immonium ions clearly indicate the presence of certain amino acids, such as P, V, I/L, H, F, Y, pY, W (sometimes also D, E, Q, N). They are particularly abundant for residues located at position 1-3 from the N-terminus and positions 1 and 2 from the C-terminus. Residues R, K, H, and W at the N-terminal position give rise to b1 ions. Immonium ions are particularly meaningful in LC-MS/MS analyses, where a high purity of the precursor ions can be achieved. A subgroup of modified residues gives rise to modification-specific reporter ions, e.g. pY, acetyl-K, carbamidomethyl-C)

Peptide quantification: Quantification methods relying on isotopic labeling of a peptide end result in labeled pairs of peptide-end specific ions. Proteolytic 18O labeling is an example. Quantitative evaluation of small fragment ions is more accurate compared to analysis of intact molecular ions due to the strongly reduced isotopic overlap.

Keywords: Collision-induced dissociation, low-mass fragment ions, peptide modification, b2 ions

See also: Hung CW, Schlosser A, Wei J, Lehmann WD. Anal Bioanal Chem 2007, 389, 1003-1016.

2.14 Family tree of all protein sequences: Recovering hidden biological knowledge

Michal Linial (SCCB - Sudarsky Center for Computational Biology, IL)

Mapping of remote evolutionary links is a classic computational problem of much interest. Relating protein families allows for functional and structural inference on uncharacterized families. However, since many sequences have diverged beyond reliable alignment, these are too remote to identify by conventional methods. I will present a method to systematically identify remote evolutionary relations between protein families, leveraging a novel evolutionary tree of all protein sequences and families. The tree is a hierarchical tree that includes over 2 million protein sequences. It allows tracing very faint links, owing to the robustness of considering the entire volume of pairwise sequence similarities at construction. I will present the power and limitation of navigating through such a tree and a method that systematically scans the tree for evolutionary breakpoint in putative ancient superfamilies.

I will discuss the limitation of knowledge extractions tools in the end of a large scale proteomics experiments. I will introduce PANDORA–a statistical visualization tool that allows fast and intuitive knowledge extraction from any set of identified peptides and proteins. This research is carried out by the ProtoNet team at the Hebrew University of Jerusalem with the kind support of the SCCB: the Sudarsky Center for Computational Biology.

[1] Noam Kaplan, Avishay Vaaknin and Michal Linial: PANDORA: keyword-based analysis of protein sets by integration of annotation sources. Nucleic Acids Research, 2003, Vol. 31, No. 19 5617-5626.

Keywords: Protein families, classification, remote homologues, annotations, webtool

Joint work of: Linial, Michal; The ProtoNet Team

Full Paper: http://nar.oxfordjournals.org/cgi/content/full/31/19/5617

See also: http://www.pandora.cs.huji.ac.il/

2.15 Tryptic transpeptidation products observed in proteome analysis by LC-MS/MS

Katrin Marcus (Ruhr-Universität Bochum, DE)

Commonly, prior to mass spectrometry based analysis of proteins or protein mixtures, the proteins are subjected to specific enzymatic proteolysis. For this purpose trypsin is most frequently used. However, the process of proteolysis is not unflawed. For example, some side activities of trypsin are known and have already been described in the literature (e.g., chymotryptic activity). Here, we describe the occurrence of transpeptidated peptides during standard proteome analysis using two-dimensional polyacrylamide gel electrophoresis followed by mass spectrometric protein identification. Different types of transpeptidated peptides have been detected. The most frequently observed transpeptidation reaction is N-terminal addition of arginine or lysine to peptides. Furthermore, addition of two amino acids to the N-terminus of a peptide has also been detected. Another transpeptidation that we observed, is combination of two peptides, which were originally located in different regions of the analyzed protein. Currently, the full amount of peptides generated by transpeptidation is not clear. However, it should be recognized that protein information is presently lost as these effects are not detectable with available database search software.

[1] Schaefer H., Chamrad D.C., Marcus K., Reidegeld K.A., Blüggel M., Meyer H.E., Tryptic transpeptidation products observed in proteome analysis by LC-MS/MS, Proteomics, 2005 (5), 846-852

Joint work of: Schaefer H.; Chamrad D.C.; Marcus K.; Reidegeld K.A.; Blüggel M.; Meyer H.E.

Full Paper: http://www3.interscience.wiley.com/journal/110429091/abstract

See also: http://www.proteomic-basics.eu/

2.16 Comparison of an alternative approach for proteome research with the common bottom up method

Katja Melchior (Universität des Saarlandes, DE)

Aim of the study: The goal of the investigation was to find an alternative method to the established bottom up approach for proteome analyses employing multidimensional chromatography. Upon separation intact proteins in the first dimension, the new method should offer better sequence coverage for the individual proteins and thus more information about protein variants.

Methods: Intact proteins from tissue extract were separated in first dimension for the semi top down method by IP-RP-HPLC in a 100 x 4.6 mm i.d. monolithic column and were then digested. For bottom up approach the whole proteome was digested prior to fractionation and the peptides were separated with SCX chromatography and also collected in fractions of two minutes. The obtained peptides of both methods were separated using IP-RP-μHPLC in a 60 x 0.1 mm i.d. monolithic capillary column. The column effluate was mixed with alpha-cyano-4-hydroxycinnamic acid and automatically spotted onto a stainless steel target. Identification of the proteins was performed offline by matrix assisted laser desorption/ionization tandem mass spectrometry (MALDI TOF/TOF).

Results: Employing the developed and optimized technology a total of 1642 proteins were identified in the proteome of glioblastoma multiforme tissue, of which 1004 were represented by more than one peptide. The identifications were compared to results obtained by the classical bottom-up approach and to results from serological screening of tumor antigens by serological identification of antigens by recombinant expression screening (SEREX). Furthermore 12% more proteins with a sequence coverage >10% were identified and 1% more with a sequence coverage >50% in comparison to the classical bottom up approach.

Novelty of the approach: For clinical studies, the advantage of IP-RP-HPLC separation of intact proteins in the first dimension rests within the possibility to focus on potential tumor markers discovered in selected fractions without the need to analyze the whole set of fractions. Moreover, identified tumor antigens may be isolated micropreparatively for further biological and structural investigation.

Keywords: Multidimensional chromatography, MALDI-TOF/TOF, proteomics

Joint work of: Melchior, Katja; Heisel, Sabrina; Meese, Eckart; Keller, Andreas; Lenhof, Hans Peter; Tholey, Andreas; Huber, Christian G.

2.17 A pipeline to detect and quantify posttranslational modifications

Markus Müller (Swiss Institute of Bioinformatics, CH)

Posttranslational modifications (PTM's) are of paramount importance in biological research. They might also be significant for clinical diagnostics since it is known that certain drugs or diseases induce PTM's, which can change the function of the affected proteins. At the Swiss Institute of Bioinformatics we are developing a pipeline to accurately detect and quantify peptides and their PTM's. The pipeline will integrate existing MS2 identification software such as Phenyx, XTandem, and Inspect, sequence tag extraction tools (Popitam) as well as quantification tools (Superhirn and MSight). Also, programs for spectrum library creation based on clustering and search are under development and will be integrated. The combination of these MS1 and MS2 tools together with other predictors for peptide detectability, pI and retention time for both modified and unmodified peptides, should provide a deeper insight into proteomic samples.

Keywords: Proteomics, Pipeline, PTM, Data Integration

Joint work of: Müller, Markus; Erik Ahrne; Laurant Geisser; Patricia Hernandez; Frederique Lisacek

2.18 Improving Identification with new Machine Learning Techniques

Nico Pfeifer (Universität Tübingen, DE)

Background: High-throughput peptide and protein identification technologies have benefited tremendously from strategies based on tandem mass spectrometry (MS/MS) in combination with database searching algorithms. A major problem with existing methods lies within the significant number of false positive and false negative annotations. So far, standard algorithms for protein identification do not use the information gained from separation processes usually involved in peptide analysis, such as retention time information, which are readily available from chromatographic separation of the sample. Identification can thus be improved by comparing measured retention times to predicted retention times. Current prediction models are derived from a set of measured test analytes but they usually require large amounts of training data.

Results: We introduce a new kernel function which can be applied in combination with support vector machines to a wide range of computational proteomics problems. We show the performance of this new approach by applying it to the prediction of peptide adsorption/elution behavior in strong anion-exchange solid-phase extraction (SAX-SPE) and ion-pair reversed-phase high-performance liquid chromatography (IP-RP-HPLC). Furthermore, the predicted retention times are used to improve spectrum identifications by a p-value-based filtering approach. The approach was tested on a number of different datasets and shows excellent performance while requiring only very small training sets (about 40 peptides instead of thousands). Using the retention time predictor in our retention time filter improves the fraction of correctly identified peptide mass spectra significantly.

Conclusions: The proposed kernel function is well-suited for the prediction of chromatographic separation in computational proteomics and requires only a limited amount of training data. The performance of this new method is demonstrated by applying it to peptide retention time prediction in IP-RP-HPLC and prediction of peptide sample fractionation in SAX-SPE. Finally, we incorporate the predicted chromatographic behavior in a p-value based filter to improve peptide identifications based on liquid chromatography-tandem mass spectrometry.

Keywords: Identification, Computational Proteomics, Machine Learning, Kernels, Support Vector Machines

Full Paper: http://dx.doi.org/10.1186/1471-2105-8-468

2.19 A retention-time alignment algorithm for LC/MS data

Katharina Podwojski (Universität Dortmund, DE)

Liquid chromatography coupled to mass spectrometry (LC/MS) has advanced to a leading technology for the analysis of complex protein mixtures. Typical quantitative proteomic studies aim at detecting differently expressed peptides between different proteomes. Prior to the final analysis of differently expressed peptides a multi-step preprocessing procedure is necessary. This is very critical as mistakes made during this part of the analysis may have heavy effects on the final detection of differences.

Especially the combination of several LC/MS maps is a crucial step in a typical analysis workflow. Nonlinear shifts in retention-time between LC/MS maps make this a nontrivial task. We have developed a statistical two-step algorithm for the retention-time alignment of different LC/MS maps. First a clustering procedure detects well-behaved features. Afterwards these are used to calculate a non-linear deviation curve for each map. This procedure showed a good alignment of most features both in simulated and real data. Thus we were able to correctly bin together most features after our alignment.

Keywords: LC/MS, Retention-Time, Alignment

Joint work of: Podwojski, Katharina; Fritsch, Arno

2.20 Experiences of a 'protease hunting lab' with protein- and knowledge databases

Hartmut Schlüter (Charité - Berlin, DE)

Many proteases are key players in a wide range of biological processes such as the release of peptide hormones, nutrient acquisition, cell growth, differentiation, antigen processing and protein turnover. It is becoming more and more obvious that the abnormal functioning of some proteases may lie behind several types of diseases, including Alzheimer's disease, cancer [1] and inflammation. The MEROPS database, which is specialized in proteases, lists about 570 known and putative genes encoding proteases in homo sapiens (19th of February 2008). The proteolytic activities of approximately 170 of these genes (30 %) is not yet validated. The number of those proteases with no endogenous substrate assigned is even much higher. In these cases the physiological role is completely in the dark. However, for the identification of the physiological roles of proteases the knowledge of the endogenous substrate(s) is fundamental.

Therefore in the first step on the road towards the decipherment of the role of a protease we are concentrating on the assignment of its endogenous substrate. One strategy for the identification of unknown endogenous substrates of peptidases with known identity includes the immobilization of protein fractions, containing the potential substrate, and their subsequent incubation with the known peptidases. By analysis of the peptides in the incubation mixtures, the parental protein can be identified [2]. The second approach starts with looking for known proteolytic reaction products either by analysis of knowledge databases or by looking at degradomes by multidimensional LC-MS approaches. By the knowledge of the amino acid sequences of the proteolytic products and the cleavage sites of their parental proteins, probes can be synthesized for the detection of the protease, which are responsible for the generation of the proteolytic products. The detection of the defined proteolytic activity is performed with the mass spectrometry based enzyme screening (MES) method [3]. The identification of an unknown protease with defined catalytic activities includes its chromatographic purification guided by the MES protease assay and the subsequent mass spectrometric analysis of the tryptic peptides of the purified active protein fraction [4]. The analysis of the mass spectrometric data against a protein database usually yields several proteins. Protein- and knowledge database searching answers the question, if any of these proteins is assigned as a protease. This way, a suggestion for the identity of the protease is yielded. The identity of the potential protease is validated by additional experiments via recombinant expression of the identified protein and comparison of the proteolytic properties of the recombinant protein with the purified active fraction.

Problems occurred on the search of protein databases and knowledge databases for the functional properties of the proteins, which were identified in the purified active fractions. These problems arise from synonyms, inaccuracies according the assignment of function and exact chemical composition of proteins and missing links towards the experimental data, as will be shown by examples.

References:

[1.] Villanueva J, Martorella AJ, Lawlor K, Philip J, Fleisher M, Robbins RJ, Tempst P (2006) Serum Peptidome Patterns That Distinguish Metastatic Thyroid Carcinoma from Cancer-free Controls Are Unbiased by Gender and Age. Mol. Cell. Proteomics 5:1840-1852.

[2.] Schlüter H, Rykl J, Thiemann J, Kurzawski S, Gobom J, Tepel M, Zidek W, Linscheid M (2007) Mass spectrometry-assisted protease substrate screening. Anal. Chem. 79:1251-1255.

[3.] Schlüter H, Jankowski J, Rykl J, Thiemann J, Belgardt S, Zidek W, Wittmann B, Pohl T (2003) Detection of protease activities with the mass spectrometry assisted enzyme screening (MES) system. Anal. Bioanal. Chem. 377:1102-1107.

[4.] Rykl J, Thiemann J, Kurzawski S, Pohl T, Gobom J, Zidek W, Schlüter H (2006) Renal cathepsin G and angiotensin II generation. J. Hypertens 24:1797-1807.

Keywords: Human proteases, degradome, protein identification, protein function, protein database, knowledge database

Joint work of: Schlüter, Hartmut; Trusch, Maria

2.21 Benchmarking of Algorithms for label-free Quantification

Ole Schulz-Trieglaff (FU Berlin, DE)

Liquid chromatography (LC) coupled to mass spectrometry (MS) is already well established for the identification of proteins in complex mixtures. The quantification of proteins in different samples is often considered as the next step in proteomics experiments leading to a comparison of protein expression in different proteomes.

Accordingly, computational methods for quantitative proteomics have also moved into the focus of the Bioinformatics community. Numerous software tools and algorithms are available for this task. Most of them were developed for data from a specific mass spectrometer and it is not clear to what extent they can be applied to data generated from other machines.

We briefly review the most common computational methods for a label-free quantification (i.e. without isotopic labelling) and give some guidelines how these algorithms could be benchmarked and their results validated.

Keywords: Liquid chromatography mass spectrometry, proteomics, quantitative

Joint work of: Ole Schulz-Trieglaff, Nico Pfeifer, Clemens Gröpl

2.22 Decoy Database Advantages and Protein Balancing for understanding the complexity of life

Christian Stephan (Ruhr-Universität Bochum, DE)

Decoy Database Builder: One of the major challenges for large scale proteomics research is the quality evaluation of results. Protein identification from complex biological samples or experimental setups is often a manual and subjective task which lacks profound statistical evaluation. This is not feasible for high-throughput proteomic experiments which result in large datasets of thousands of peptides and proteins and their corresponding mass spectra. To improve the quality, reliability and comparability of scientific results, an estimation of the rate of erroneously identified proteins is advisable. Moreover, scientific journals increasingly stipulate that articles containing considerable MS data should be subject to stringent statistical evaluation. We present a newly developed easy-to-use software tool enabling quality evaluation by generating composite target-decoy databases usable with all relevant protein search engines. This tool, when used in conjunction with relevant statistical quality criteria, enables to reliably determine peptides and proteins of high quality, even for nonexperienced users (e.g. laboratory staff, researchers without programming knowledge). Different strategies for building decoy databases are implemented and the resulting databases are characterized and compared. The quality of protein identification in high-throughput proteomics is usually measured by the false positive rate (FPR), but it is shown that the false discovery rate (FDR) delivers a more meaningful, robust and comparable value.

Balancer Proteins: Systems biology is one of the key fields to gain a proper understanding concerning the dynamic processes taking place in living cells and tissues. The integration of all available scientific areas including proteomics, genomics, molecular biology and the overall bracket bioinformatics will allow achieving essential insight about how cell activities are regulated. Therefore, numerous studies identified several hundred of regulated proteins or genes; only a few of them seem to be disease-specific regulated proteins or genes. By comparing these studies it will become obvious that most of the reported proteins/genes are not specific for the designed study because they are found in many totally different studies. Recently, evidences were reported that the maintenance of the overall protein amount plays a crucial role in the functionality of a given cell. The loss or the change in the amount of essential proteins will lead to a disturbance of the protein stoichiometry and to a system malfunction that is partly compensated by other, disease-unspecific proteins. These proteins can be regarded as 'balancer' proteins, also showing significant regulation, but being erroneously identified as specific-regulated. The amount of these balancing proteins could be adopted by increasing or decreasing the expression rate in a defined manner, compensating the changed overall protein amount in these cells as well as in locally organized topological cell compartments. However, the triggered effect could have a high influence by increasing or decreasing the amount of some 'effector' related proteins, controlled by protein regulatory networks which underlie these mechanisms of 'resilience' biology. The 'balancer' proteins essentially contribute to the overall network entropy and can be named hubs of these protein regulatory networks.

2.23 COmbined FRActional DIagonal Chromatography (COFRADIC)

Joel Vandekerckhove (Gent University + VIB - Gent, BE)

To reduce the complexity of peptide mixtures while retaining the representative character of their parent proteins, we follow strategies of targeted peptide isolation and identification. One of the first examples of such a strategy was the ICAT technology introduced by the group of Aebersold (1). Here, cysteine peptides were targeted, while other peptides were discarded. We have introduced the principle of diagonal chromatography to sort subsets of peptides from highly complex mixtures derived from total lysates. The procedure consists of three steps. First, peptides are separated by conventional RP-HPLC (called the primary run) and collected in fractions. Second, in each of the fractions a subset of peptides is altered either chemically or enzymatically. Third, each treated fraction is rerun a second time using identical separation conditions (the secondary run).Unaltered peptides elute within the same time interval as previously, while altered peptides shift from their original position and can be recognized, collected and identified by MS/MS-based analyses. By combining several primary fractions prior to the secondary runs, we reduce the overall number of HPLC runs. This procedure can be executed in a fully automated manner and is called COmbined FRActional DIagonal Chromatography (COFRADIC).

Due to the broad choice in chemistries, the COFRADIC approach is highly versatile and can be applied to select for a large variety of peptide subsets and even for individual peptides. Examples have been published for the isolation of methionine or cysteine-containing peptides, phosphopeptides, N-glycopeptides, sialylated N-glycopeptides and peptides covering the extreme N-termini of proteins (for reviews, see (2-4)).

This approach, which is essentially gel-free or peptide-centric, poses special requirements on the supporting bioinformatics platforms. Below we mention some of these challenges.

- As we sort individual peptides as signatures for their parent proteins, it is mandatory to dispose of robust identification software. We assist the MASCOT search algorithm by an additional filtering algorithm that removes spectra that contribute little to protein identification. This typically removes more than half of all spectra and has pointed to high quality, though unidentified MS/MS spectra as originating from peptides carrying unanticipated modifications (5).

- Another algorithm clusters and can merge MS/MS spectra after calculating spectral similarities and can thereby increase the confidence of identification (6).

- To further parse huge amounts of often non-accessed, though both important and interesting data object underlying peptide identifications, we built Mascot-Datfile which enabled us to analyze MASCOT result files in great detail (7).

- N-terminally located peptides are of particular nature, characterized by an open amino-terminal border and a predicted C-terminal arginine end. This often imposes constraints on commercial identification software. Therefore, we created DBToolkit (8) that processes protein sequence databases to databases holding peptides that are either N-terminally or C-terminally truncated or contain specific amino acids. These databases form then the search space for search engines such as MASCOT, and result in a higher scoring rate by readily detecting internal protease cleavage sites and are as such important tools for protease degradomics research (9).

- Data handling and data management are crucial when a community shares and uses data often coming from different projects and generated by different machines. To cope with this complexity, we have built a robust LIMS system that interconnects different projects and allows every member of the lab to plug in and retrieve data (http://genesis.ugent.be/ms_lims/).

- Our lab was also one of the first to realize the importance the general accessibility of the proteomics data. PRIDE, developed in collaboration with the European Bioinformatics Institute, offers a data upload facility and a documented application programming interface for direct access to a large variety of proteomics data (10).

- Probably the most challenging problem is linking the proteomics information with underlying biology. There the needs are both highly diverse and often linked to individual labs. For instance, we are currently creating a database of protein processing sites formed as the result of normal cellular physiology or that are typical for aberrant cellular behavior. In the same line, and based on our massive data set, we can start thinking on creating reliable predictors for protein processing sites.

References:

1. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17, 994-9.

2. Gevaert, K., Van Damme, P., Martens, L., and Vandekerckhove, J. (2005) Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics? Anal Biochem 345, 18-29.

3. Gevaert, K., Van Damme, P., GhesquiÃČÂÍre, B., and Vandekerckhove, J. (2006) Protein processing and other modifications analyzed by diagonal peptide chromatography. Biochim Biophys Acta 1764, 1801-10.

4. Gevaert, K., Impens, F., Van Damme, P., Ghesquiere, B., Hanoulle, X., and Vandekerckhove, J. (2007) Applications of diagonal chromatography for proteome-wide characterization of protein modifications and activity-based analyses. Febs J 274, 6277-89.

5. Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., and Eidhammer, I. (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. Proteomics 6, 2086-94.

6. Flikka, K., Meukens, J., Helsens, K., Vandekerckhove, J., Eidhammer, I., Gevaert, K., and Martens, L. (2007) Implementation and application of a versatile clustering tool for tandem mass spectrometry data. Proteomics 7, 3245-58.

7. Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2007) MascotDatfile: An open-source library to fully parse and analyse MASCOT MS/MS search results. Proteomics 7, 364-66.

8. Martens, L., Vandekerckhove, J., and Gevaert, K. (2005) DBToolkit: processing protein databases for peptide-centric proteomics. Bioinformatics 21, 3584-5.

9. Van Damme, P., Martens, L., Van Damme, J., Hugelier, K., Staes, A., Vandekerckhove, J., and Gevaert, K. (2005) Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. Nat Methods 2, 771-7.

10. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: the proteomics identifications database. Proteomics 5, 3537-45.

Keywords: COFRADIC, gel-free proteomics, peptide-identification, targeted analysis

2.24 Phosphoproteome of resting human platelets

René Zahedi (Rudolf-Virchow-Zentrum, DFG-Forschungszentrum für Experimentelle Biomedizin, Universität Würzburg, DE)

Platelets are small anucleate blood cells which circulate in the blood stream for about 8 days. Under physiological conditions they show almost no interaction with other blood components. However upon encountering sites of endothelial lesion, they adhere to exposed components of the subendothelial matrix, are activated and aggregate. Thereby, they build a stable thrombus which seals the wound and prevents the organism from uncontrolled blood loss. This process is termed hemostasis. Unfortunately, the same underlying mechanisms can also lead to pathological thrombosis which in the end might result in vessel occlusion and thus myocardial infarction or stroke. Since cardiovascular diseases are the major cause of death in industrialised countries, a lot of effort has been made to refine our understanding of platelet activation.

The almost complete absence of protein synthesis renders platelets ideal targets for proteomic studies. Therefore, in a comprehensive study we characterize the platelet proteome on various levels: global, PTM-directed, organellar and membranous. So far, we have identified more than 1500 proteins, more than 1000 phosphorylation sites and more than 350 glycosylation sites from human platelets. A targeted analysis of platelet derived plasma membranes furthermore yielded the identification of more than 500 distinct proteins. Extending the conducted experiments to the quantification of PTM and protein abundance upon specific stimulation of isolated platelets will allow novel hypotheses about platelet function which will be further addressed by additional biochemical studies.

Thus, our knowledge of platelet function will be refined on the molecular basis and the implementation of generated and validated data will lead us one step further into platelet systems biology.

Joint work of: RenÃČÂľ Zahedi, Urs Lewandrowski, Albert Sickmann

Full Paper: http://pubs.acs.org/doi/abs/10.1021/pr0704130

See also: http://www.protein-ms.de

See also: http://www.proteomics-workshop.de/

2.25 Data-dependent MS/MS analysis

Alexandra Zerck (MPI für Molekulare Genetik, DE)

When analysing complex samples by LC-MS/MS it is often not possible to acquire MS/MS spectra of all peptide signals. But especially when one protein is present in high abundance this protein is often identified by a large number of peptides. This makes it difficult to identify lower abundant proteins.

Through a directed selection of precursor ions we can speed up protein identification, especially for lower abundant proteins.

We present an iterative procedure to select the precursor ions based on the identification results from earlier iteration steps. Using this data dependent precursor ion selection we save time in the analysis, optimize the sample usage and decrease the redundancy of the data. To evaluate our online heuristic approach we also propose a theoretical formulation of the problem as an Integer Linear Program (ILP).

2.26 Proteomics goes intelligent: Identification of activated pathways from proteomics expression data

Roman Zubarev (University of Uppsala, SE)

Rapid and easy identification of pathways activated in a cell at given experimental conditions is a very important task solving which could provide new diagnostic biomarkers and candidates for drug discovery. In the absence of a priori data on which pathway may be activated, expression proteomics results are an ideal input for ‘blind’ pathway identification. These data consisting of IDs and relative abundance changes of proteins measured at the level of whole proteomes can be mapped onto known pathways in an attempt to identify which pathway accommodates the biggest amount of abundance changes. However, direct mapping (also known as enrichment analysis) seldom produces correct answers, because many proteins whose abundances change relate only distantly to the originally stimulated pathway. Correct pathway identification requires knowledge of the ‘hidden layer’ of cell regulation.

We have found that activated pathways can be correctly identified from expression proteomics data using an interpretation procedure that is based on a database of genes, molecules, and their interactions. The procedure combines the results of straight gene mapping and transcription factor analysis, and in both cases the mapping is preceded by a special routine known as keynode analysis.

Here we describe the above procedure and validate it using original as well as published proteomics data. We also illustrate how useful information can be obtained in real-life biological research using the proposed approach. For best results, proteomics experiment should be performed as a single LC/MS/MS run. Identification and quantification of several hundred proteins is sufficient for reliable pathway analysis.

Keywords: Pathway analysis; protein quantitation; expression proteomics

Joint work of: Zubarev, Roman; Fung, Eva; Savitski, Mikhail; Kel, Alexander; Wingender, Edgar; Kel-Margoulis, Olga

Participants of Dagstuhl Seminar 08101 on *Computational Proteomics* held on 3rd to 7th March 2008.