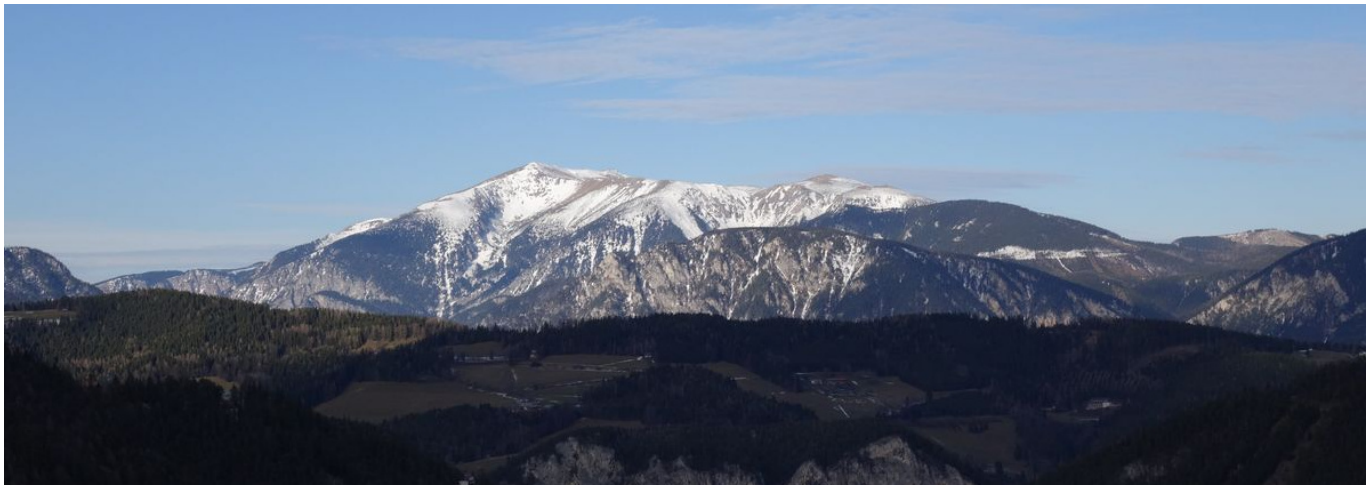

EuBIC Winter School 2017

Programme

10th – 13th January 2017, Sporthotel Semmering, Austria

https://www.fh-ooe.at/eubic-ws17

Contents

Day 1 – 10. January 2017

Educational Day

Time	Title		
09:10 - 09:30	Opening Welcome & Introduction		
09:30 - 09:45	Coffee Break		
09:45 - 12:00	Developing Spectral Libraries with Progenesis QI for Proteomics, Part I	ELIXIR and de.NBI Hackathon	Introduction to Proteomics Data Analysis, Part I
12:00 - 13:00	Lunch Break		
13:00 - 14:45	Developing Spectral Libraries with Progenesis QI for Proteomics, Part II	ELIXIR and de.NBI Hackathon	Introduction to Proteomics Data Analysis, Part II
14:45 - 15:00	Coffee Break		
15:00 - 19:00	Highlights of the Latest Developments of the Proteome Discoverer Framework	ELIXIR and de.NBI Hackathon	Introduction to Proteomics Data Analysis, Part III
19:00	Dinner		

Day 2 – 11. January 2017

Bioinformatics Challenges in Identification & Quantification

Time	Title	Speaker(s)
09:00 - 09:05	Morning Welcome and Announcements	
09:05 - 09:55	Constructing community knowledge for peptide identification and quantification	Nuno Bandeira
09:55 - 10:15	Coffee Break	
10:15 - 11:05	Broadening the mission: quality assessment for quantitative and label-dependent MS	David Tabb
11:05 - 12:00	Poster Flash Talks of Young Investigators Part I	
12:00 - 13:00	Lunch Break	
13:00 - 13:50	A discovery portal for bioinformatics resources	Elixir
13:50 - 14:20	Developing structural interactomics and its application in cell biology	Fan Liu
14:20 - 14:40	Coffee Break	
14:40 - 18:40	Workshop Session	see workshop descriptions
18:40 - 19:40	Dinner	
19:40	Poster Session & Come Together	

Parallel Workshops

Fan Liu	Structural Interactomics by Cross-linking Mass Spectrometry
Nuno Bandeira	Global Big Data Challenge
Qiagen	Ingenuity Pathway Analysis
David Tabb	Proteomics without a Genome: Leveraging RNA-Seq from Non-Model Organisms

Day 3 – 12. January 2017

Result Interpretation

Time	Title	Speaker(s)
09:00 - 09:05	Morning Welcome and Announcements	
09:05 - 09:55	Open Data	Lennart Martens
09:55 - 10:15	Coffee Break	
10:15 - 11:05	Issues of LC-MS Quantification	Oliver Kohlbacher
11:05 - 12:00	Poster Flash Talks of Young Investigators Part II	
12:00 - 13:00	Lunch Break	
13:00 - 13:50	Reactome: A Curated Knowledgebase of Biomolecular Pathways	Antonio Fabregat Mundo
13:50 - 14:10	Coffee Break	
14:10 - 18:10	Workshop Session	see workshop descriptions
18:10 - 19:10	Dinner	
from 19:10	Social Event	

Parallel Workshops

Jürgen Cox	Perseus & MaxQuant
Oliver Kohlbacher	Automated Processing of Quantitative Proteomics Data with OpenMS
Antonio Fabregat Mundo	Accessing Reactome Services and Integrating its Widgets

Day 4 – 13. January 2017

From Proteomics to Multi-Omics in Bioinformatics

Time	Title	Speaker(s)
09:00 - 09:05	Morning Welcome and Announcements	
09:05 - 09:55	The MaxQuant and Perseus Computational Platforms for Comprehensive Analysis of Large-scale (Prote)Omics Data	Jürgen Cox
09:55 - 10:30	Coffee Break & Poster Session	
10:30 - 11:25	Integrating Data and Software using Ontologies	Magnus Palmblad
11:25 - 12:15	Best Flash Talk and Best Poster Award	
12:15 - 14:00	Lunch, Outlook and Workshop Closing	

Sponsors


elixir
DENMARK


QIAGEN


Waters
THE SCIENCE OF WHAT'S POSSIBLE.®


ThermoFisher
SCIENTIFIC
The world leader in serving science


sapere aude
fhoö
UNIVERSITY
OF APPLIED SCIENCES
UPPER AUSTRIA


UNIVERSITAS BERGENSIS
UNIVERSITY OF BERGEN


SDU
UNIVERSITY OF
SOUTHERN DENMARK


mpc
MEDIZINISCHESPROTEOMCENTER


iPBS

Talk Abstracts

Day 2 – 11. January 2017

Constructing community knowledge for peptide identification and quantification

Nuno Bandeira

Over 95% of proteomics mass spectrometry data in the public domain does not have proper peptide or protein identifications; the situation is even worse when it comes to quantification. The bioinformatics challenges before us thus begin with i) systematic reanalysis of all public data for discovery of new protein variants and post-translational modifications and ii) principled aggregation of billions of search results into a community-wide spectral library reusable for peptide identification and quantification in both DDA and DIA experiments. We will present recent developments in these areas and discuss ways to tackle these challenges by incorporating contributions from multiple research groups.

Broadening the mission: quality assessment for quantitative and label-dependent mass spectrometry

David Tabb

In April of 2016, the HUPO-PSI created a new working group devoted to developing infrastructure for quality control (QC) of proteomic and metabolomic data sets. A quick survey of the QC literature would suggest that label-free, data-dependent LC-MS/MS has drawn almost attention in this field. A closer look, though, finds interesting QC advances and opportunities in other types of experiments. This talk will examine QC implementations in Spectrum Mill, SProCoP, and MSstats for isobaric labels like iTRAQ as well as targeted quantitation experiments. Finally we will look at some potential avenues for assessing quality in Data-Independent Acquisition and MALDI experiment designs.

A discovery portal for bioinformatics resources

ELIXIR

ELIXIR (1) the infrastructure for biological information, is building a discovery portal (2) for bioinformatics resources world-wide, built upon a distributed, community curation of a wide range of key resources, including analytical tools and data services. At the heart of this effort is a registry of essential scientific and technical information, conforming to a consistent syntactic and semantic standard. Convenient tools help an end-user to find, understand and compare the resources they need, and to use and connect them. The registry will also make software more citable, as PubMed does for scientific publications. ELIXIR is supporting resource providers, including research infrastructures, institutes, projects and individuals, to contribute to an emerging global distributed curation effort, that will ensure comprehensive content and high quality annotations, both of which are essential for the sustainable impact of the registry in the community. The registry will, in due course, expose the results of tool benchmarking and service monitoring to provide the end-user with a robust, scientifically relevant measure of quality and performance. Furthermore, integration of the registry with key workbench environments will assist the users with managing their day-to-day workflows. To get involved in this effort please contact the registry coordinators (3).
(1) `http://www.elixir-europe.org/`
(2) `https://bio.tools`
(3) registry@elixir-dk.org

Developing structural interactomics and its application in cell biology

Fan Liu

In the last decade, chemical cross-linking combined with mass spectrometry (XL-MS) has become an increasingly powerful approach to probe the in-solution structures of proteins/protein assemblies. In principle, XL-MS is able to profile the structure of individual proteins, topological maps of protein assemblies and protein interaction networks. However, its current scope is mainly limited to the characterization of endogenously purified or in vitro reconstituted protein assemblies. Therefore, to comprehensively understand the protein interactome, the XL-MS must move beyond simple protein complexes to the proteome-wide

analysis of interaction patterns in vivo. Accordingly, innovative methods are urgently needed to overcome the technical limitations in the field.

Here, we describe a major leap forward in cross-linking mass spectrometry demonstrating a new integrated workflow that robustly identifies cross-links from proteome samples. Our approach is based on the application of a cleavable cross-linker, sequential CID and ETD MS2 acquisitions of peptide fragmentation spectra of cross-linked peptides, and a dedicated search engine termed XlinkX. We applied this novel XL-MS strategy to several highly complex samples, including E. coli lysate, HeLa lysate and intact mitochondria. Through thousands to ten thousands of cross-links, were obtained crucial insights into the interaction patterns, the binding interfaces and the molecular organization of the protein components, allowing us to gain a deeper understanding of various functional processes of the cell.

Day 3 – 12. January 2017

Open Data

Lennart Martens

To hell with the problems with getting data open (that battle has been won) - why are we not doing more with the data that are there yet, and how do we fix this.

Issues of LC-MS Quantification

Oliver Kohlbacher

Reactome: A curated knowledgebase of biomolecular pathways

Antonio Fabregat Mundo

Reactome (`http://www.reactome.org`) is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways, aiming to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge and to support basic research, genome analysis, modelling, systems biology and education. Pathways are built from connected "reactions" that encompass many types of biochemical events. Reactions are derived from literature and must cite a publication that experimentally validates them. Pathways are authored by expert biologists and peer reviewed before incorporation into the database. In its latest release (v58), Reactome includes 10,168 reactions covering 10,212 human gene products and supported by 24,968 literature references. Users can search for proteins or compounds and see details of the complexes, reactions and pathways they participate in. Pathway diagrams allow users to examine the molecular events that constitute the steps in pathways and to view details of the proteins, complexes and compounds involved. Different forms of pathways analysis can be performed with the Reactome analysis tools. Users can submit a list of identifiers for overrepresentation analysis or submit quantitative datasets, such as microarray data, for expression analysis. Results of these analyses are overlaid onto the Pathways Overview and Diagram Viewer for easy navigation and interpretation. Interaction data from multiple resources can be used to expand pathways. Interactors from IntAct are included by default in the search feature and can be taken into account in the analysis service. Finally, pathways or all Reactome content can be downloaded in many formats including TSV, CSV, PDF, SBML, BioPax and PSI-MITAB.

Day 4 – 13. January 2017

The MaxQuant and Perseus Computational Platforms for Comprehensive Analysis of Large-scale (Prote)Omics Data

Jürgen Cox

Currently, a main bottleneck in proteomics is the downstream biological analysis of highly multivariate quantitative protein abundance data. It will be shown how the Perseus software supports researchers in

interpreting protein quantification, interaction and posttranslational modification data. A comprehensive portfolio of statistical tools for high-dimensional omics data analysis is contained covering normalization, pattern recognition, time series analysis, cross-omics comparisons and multiple hypothesis testing. A machine learning module supports classification and validation of patient groups for diagnosis and prognosis, also detecting predictive protein signatures. Central to Perseus is a user-friendly, interactive workflow environment providing complete documentation of computational methods used in a publication. All activities in Perseus are realized as plugins and users can extend the software by programming their own, which can be shared through a plugin store. Perseus combines a powerful arsenal of algorithms with intuitive usability by biomedical domain experts, making it suitable for interdisciplinary analysis of complex large datasets.

Integrating Data and Software using Ontologies

Magnus Palmblad

In a general sense, we can structure data in two ways: top-down (a priori, supervised) or bottom-up (a posteriori, unsupervised). In the former, we decide and impose the categories and hierarchies on the data. In the latter, we let the data speak for itself and use machine learning or statistical methods to generate a data-dependent structure.

For most purposes of multi-omics integration of heterogeneous data, a strong case can be made for a supervised system for formal naming and definition of data categories (types), properties and relationships within our domain. We call such a system an ontology. In bioinformatics, we use different ontologies for different purposes. We describe functions or cellular localizations of proteins using the Gene Ontology (GO). To integrate anatomically defined data involving a single species, we use a species-specific anatomical ontology such as ZFA (for zebrafish) or MA (for mouse). To compare data across model systems and the human, we can use the generic vertebrate UBERON. To provide metadata describing how our mass spectra were acquired, we use the PSI-MS controlled vocabulary. And for categorizing bioinformatics software tools themselves, we use the EDAM ontology in bio.tools.

In this talk I will gently introduce these concepts and illustrate how we use ontologies in our everyday work.

Workshop Abstracts

Day 1 – 10. January 2017

Introduction to Proteomics Data Analysis

Harald Barsnes and Marc Vaudel

Mass spectrometry based proteomic experiments generate ever larger datasets and, as a consequence, complex data interpretation challenges. In this course, the concepts and methods required to tackle these challenges will be introduced. The course will focus on protein identification and take the participant from the handling of the raw data to the statistical analysis of the identification data. The course will rely exclusively on free and user friendly software, all of which can be directly applied in your lab upon your return from the Winter School.

ELIXIR and de.NBI Hackathon

Veit Schwämmle, Magnus Palmblad, Jon Ison, Niall Beard, Gerhard Mayer and Julian Uszkoreit

Do you want to learn about the current European initiatives for computational infrastructure and data standards in life sciences? We aim to make the computational proteomics community an integrative part of these recent developments. The hackathon, kindly sponsored by ELIXIR Denmark and in cooperation with de.NBI, is the first outreach of ELIXIR - the European Infrastructure for Biological Information - to the Proteomics Community. Current ELIXIR efforts for establishing sustainable infrastructures within computational biology will be presented. We will introduce and discuss integration of computational proteomics into ELIXIR and initiation of common projects in both research and training. The main focus is on the bio.tools registry for software annotation, workflow composition, standards for data formats and data uploads. The participants will be engaged in the discussions and task forces to deepen collaborations between ELIXIR, de.NBI and experts in computational proteomics.

Developing Spectral Libraries with Progenesis QI for Proteomics

Joel Rein and Horst Schreiner (Waters)

Learn how to use Progenesis QI for proteomics to Quantify and Identify the peptides and proteins that are significantly changing in your samples, and how you'll soon be able to develop your own customised spectral libraries to improve the speed and specificity of your MSe searches. Bring your own laptop and download the demo software and tutorial from `http://www.nonlinear.com/progenesis/qi-for-proteomics/download/`

Highlights of the Latest Developments of the Proteome Discoverer Framework

Bernard Delanghe and André Müller (Thermo Fisher Scientific)

An overview of the features in the current version, PD 2.1 will be presented with emphasis on the improvements in quantification. Furthermore we will highlight the upcoming new release, including Label Free Quantification and Cross-linking.

This will be an interactive presentation with time for questions and feedback.

Day 2 – 11. January 2017

Structural Interactomics by Cross-linking Mass Spectrometry

Fan Liu

Chemical cross-linking combined with mass spectrometry (XL-MS) has emerged as a powerful approach to investigate protein conformation as well as protein-protein interactions. Especially in recent year, this technique has moved rapidly towards the analysis of very large protein assemblies and heterogeneous mixtures of protein complexes. This course is designed for scientist who is interested in using this technique to probe the structure of various proteins/protein complexes and to discover novel protein-protein interactions. The course will cover most aspects of the XL-MS workflow, including sample preparation,

cross-link enrichment, MS data acquisition and cross-link data analysis. Furthermore, attendees will also be provided with practical training on XL-MS data analysis using standalone XlinkX and XlinkX PD node.

Global Big Data Challenge

Nuno Bandeira

This "global big data challenge" workshop will focus on three key aspects of contributing to the global proteomics knowledge base: i) using advanced algorithms for discovery and inspection of post-translational modifications and highly-modified peptides; ii) sharing search results with the community at large and reviewing results contributed by others; iii) reusing spectral libraries for peptide identification and detection in both DDA and DIA mass spectrometry data. The workshop will cover topics of relevance to both experimentalists (e.g., how to critically inspect search results) and bioinformaticians (e.g., how to share and compare results from new software tools).

IPA Workshop

Mario Ricketts and Andre Koper (QIAGEN)

Ingenuity® Pathway Analysis (IPA®) is a powerful analysis and search tool that helps researchers to uncover the biological significance of 'omics data and to answer critical questions related to their studies. Built on extensive and primarily manually curated scientific content from QIAGEN's distinctive Knowledge Base, IPA's content-aware and causal analytics assists with the identification of canonical pathways, phenotypic effects, networks of interacting molecules and putative upstream drivers that help users interpret various types of 'omics experiments, including measurements of differentially expressed or phosphorylated proteins. The integration of different 'omics data is a strong focus for IPA that allows users to visualize different molecular data together and to gather evidence for a more comprehensive interpretation of experimental data with just a few mouse clicks. IPA has been cited in over 16,000 scientific publications.

This workshop will help the attendees gain an overview of IPA's capabilities and to experience its graphical user interface and approach to the biological interpretation of proteomics and phospho-proteomics datasets with or without accompanying gene expression data. As part of the workshop, we will focus on creating and interpreting IPA analyses and demonstrating tools in IPA to create, visualize and analyze causal effects of automated and user-defined molecular networks. At the end of this 4 hour session, each attendee should be able to upload and analyze their data in IPA in a comprehensive manner. In addition, QIAGEN will provide 14 day IPA trials to every attendee of this IPA workshop at the EuBIC Winterschool 2017.

Proteomics without a Genome: Leveraging RNA-Seq from Non-Model Organisms

David Tabb

As proteomics spreads to an ever-larger number of applications, some researchers (particularly in agriculture) are hindered by the lack of a high-quality genome annotation for their species of interest. In this workshop, we will examine methods for preparing a draft proteome sequence database from transcriptomic data collected through high-throughput sequencing (RNA-Seq). These stages include the following:

1. checking the quality of the sequencing reads
2. assembling the transcripts de novo
3. translating the transcripts to amino acids
4. finding nearest annotated taxonomic neighbors
5. annotating sequences by reciprocal BLAST and InterPro

Day 3 – 12. January 2017

Introduction to the MaxQuant and Perseus software platforms

Jürgen Cox

Description: This workshop provides an introduction to the computational proteomics platform MaxQuant and the downstream bioinformatics platform Perseus. The first part provides theory and background information to the workflows and algorithms while the second part is hands-on and participants will be able to apply the tools to some real-world examples.

Automated Processing of Quantitative Proteomics Data with OpenMS

Oliver Kohlbacher and Julianus Pfeuffer

The workshops provides a brief introduction to OpenMS, an open-source software for computational proteomics and metabolomics. Participants will familiarize themselves with the underlying concepts and get to know a few key tools of the OpenMS tool collection. We will construct tailor-made automated data analysis workflows for database search, label-free quantification, data visualization and quality control in proteomics.

These workflows will be applied to selected example data sets. Particpants are encouraged to bring their own data and discuss the analyses required with instructors from the OpenMS team. All software used will be provided and can be installed on particpants' own computers.

Accessing Reactome Services and Integrating its Widgets

Antonio Fabregat Mundo

Reactome is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways. It aims to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modelling, systems biology and education. Thus, the mainstays of its software development are usability and responsiveness from the user's point of view, likewise modularity and reusability from the developer's side. Reactome offers web services and widgets to facilitate integration in third-party software. One service provides database access while the other performs overrepresentation and expression analysis as well as species comparison. Widgets for the Pathways Overview and Pathway Diagrams are provided for JavaScript and GWT. Both widgets overlay the results of the Analysis Service. Protein-protein or protein-chemical interactions can be used to extend pathways beyond Reactome's curated content. IntAct is the default resource but all other PSICQUIC databases can be selected and in addition, users can submit custom interactions. Interaction data from IntAct are also included in the Reactome main search and the Analysis Service, helping users identify pathways of interest. In summary, Reactome has facilitated data integration by providing easy-to-use services and reusable widgets. Several resources such as OpenTargets, ChEBI, BluePrint, PRIDE, PINT, and IP2 have already integrated these services and widgets.

Poster Flash Talks of Young Investigators

Day 2 – 11. January 2017

Ludger Goeminne	MSqRob: analysis of label-free proteomics data in an R/Shiny environment
Johannes Griss	The spectra-cluster toolsuite: Enhancing proteomics analysis through spectrum clustering
Emma Ricart Altimiras	A Bioinformatics Tool for Nonribosomal Peptides Identification by Tandem Mass Spectrometry
Sebastian Dorl	Identifying tandem mass spectra of phosphorylated peptides before database search using machine-learning
David Hollenstein	MasPy – mass spectrometry-based proteomics data analysis with python
Adriaan Sticker	Mass spectrometrists should search for all peptides, but assess only the ones they care about
Surya Gupta	An unbiased protein association study on the public human proteome reveals biological connections between co-occurring protein pairs
Christophe Bruley	Proline: a software environment for label-free quantification data analysis and exploration

Day 3 – 12. January 2017

Hugo López-Fernández	Mass-Up and Decision Peptide-Driven: two open-source applications for MALDI-TOF MS data analysis and protein quantification
Thilo Muth	Analyzing metaproteome samples on the go: the full-featured MPA portable software provides protein identification enriched with taxonomic and functional meta-information
Corinna Klein	Electronic sample management and archiving system for proteomics MS-data
Adithi Varadarajan	An integrated proteogenomics approach to discover the entire protein-coding potential of prokaryotic genomes
Roman Mylonas	MsViz, a zero learning curve graphical software tool for detailed manual validation and quantitation of post-translational modifications
Dominik Kopczynski	PeptideMapper: Efficient and Versatile Amino Acid Sequence and Tag Mapping
Felix Van der Jeugt	Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples

Poster Abstracts

1. Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples

Bart Mesuere and Felix Van der Jeugt

Unipept (http://unipept.ugent.be) integrates a fast index of tryptic peptides built from UniProt records with cleaned up information from the NCBI Taxonomy to allow for biodiversity analysis of metaproteome and metagenome samples. Unipept has the following features:

Tryptic Peptide Analysis - Fast mapping of a single tryptic peptide to a list of all UniProt records wherein the peptide was found along with a complete taxonomic lineage derived from the NCBI taxonomy.

Metaproteomics Analysis - Fast mapping of a list of tryptic peptides to their taxonomic lowest common ancestors (LCA). These LCAs will be bundled and displayed in an interactive treemap giving you insight into the biodiversity of your sample.

Unique Peptide Finder - Finds unique tryptic peptides for a species or genus of choice. The application is powered by the Uniprot proteomes, lets you add your own set of genomes and enables you to download lists of unique peptides, which can be used for targeted proteomics experiments.

Peptidome Clustering - Lets you select a set of genomes of your choice, and calculate similarities between them based on their tryptic peptidome. The result is shown as a similarity matrix and a phylogenetic tree.

API - Unipept offers most of its peptide analysis features as a web service. This enables the integration of Unipept functionality into other applications and the creation of batch processing scripts. Additionally a command line interface is available to access the Unipept web service.

The next step in the development of Unipept is to add support for the functional and comparative analysis of metaproteomics datasets.

2. A Bioinformatics Tool for Nonribosomal Peptides Identification by Tandem Mass Spectrometry

Emma Ricart Altimiras, Mickael Chevalier, Maude Pupin, Valerie Leclere, Christophe Flahaut, Markus Mueller and Frederique Lisacek

Nonribosomal Peptides (NRPs) are natural compounds enzymatically synthetized by microorganisms such as bacteria and fungi. These peptides have shown a wide range of biological properties such as antibiotics, antitumor or immunosuppressant, being of great importance to the pharmacological and agricultural industry. Due to its high sensitivity and accuracy, Mass Spectrometry (MS) is crucial for the identification of these biomolecules. However, their unusual chemical structures (cyclic, polycyclic, branched...) and the presence of highly modified non-proteogenic monomers complicate the interpretation of their MS/MS spectra. Tools for the identification of some simple NRPs already exist, but they do not cover all NRP specificities, lack flexibility, efficient scoring and statistical validation as well as user friendliness. Here we present a new bioinformatics tool to match predicted MS/MS spectra against their experimental counterparts, either exactly or in a modification tolerant way.

Norine, a database entirely dedicated to non-ribosomal peptides, is used to retrieve the molecular structure of the NRPs. Based on this information we have developed a fragmentation model to calculate the MS/MS fragments of each peptide and predict their theoretical spectrum. The model covers all the structures observed in NRPs (cyclic, multicyclic and branched) and includes the 500+ non-proteogenic monomers. All the known fragmentation characteristics of NRPs have been included and multiple ring breakages are taken into account in order to calculate the putative fragment masses. Additionally, a combinatorics algorithm has been developed in order to allow modification and adduct tolerant searches. Once a spectrum-peptide match (PSM) is confidently identified, it can be added to a spectral library with its corresponding annotations.

Our software is able to perform a configurable and complete computational fragmentation of NRPs, including those presenting complex structures containing multicycles and several branches. Preliminary tests with experimental MS/MS data from Fengycin show positive results: the tool is able to perform an efficient fragmentation and match high intensity peaks. Furthermore, this is the first NRP fragmentation tool that includes modification tolerant searches, which will be very useful for the identification of new peptides.

3. Hormonal Regulation of Lipolysis

Petra Krenn, Matthias Schittmayer and Ruth Birner-Gruenberger

Lipolysis, the hydrolysis of triglycerides, is a crucial process involved in maintaining energy homeostasis. Molecular events regulating lipolysis however are poorly understood. Adipocytes are the major site of lipid storage and mobilization. 3T3-L1 cells are the most widely used adipocyte model, but it is still challenging to achieve a high degree of differentiation. We differentiate the cells on glass plates, stimulate them to induce lipolysis and snap-freeze them to maintain the phosphorylation patterns. The complex interplay of the molecular posttranslational events involved in lipolysis requires a global untargeted analytical approach, which allows unbiased investigation of phosphorylation with temporal resolution and dissection of kinase-substrate relationships, for its elucidation. Thus we apply a global phosphoproteomics workflow to investigate these events. An enrichment specificity of about 95% is achieved using TiO2 affinity chromatography. Thousands of phosphopeptides are identified in a single experiment using a 2 hour gradient on a Bruker MaXis2 instrument. Label free quantification by MaxQuant and Perseus reveals changes in the phosphoproteome upon lipolytic stimulation over time.

The compiled data by us and others points towards a tight posttranslational regulation of lipid mobilization to enable efficient energy homeostasis while avoiding lipotoxicity. Multiple phosphorylation and dephosphorylation events on involved lipases and regulators as well as sensors for intracellular free fatty acid levels appear to allow their dynamic interplay by formation of different protein complexes to regulate the lipolytic output under basal and stimulated conditions. Moreover, other processes, like upstream and downstream signaling, cytoskeletal arrangements and membrane trafficking events involved in regulation of these lipolytic complexes, as well as connected metabolic pathways appear to be co-regulated.

4. Investigation of the effects of alkylation during sample preparation on proteomic data

Ksenia Kuznetsova, Elizaveta Solovieva, Maria Karpova, Dmitry Karpov, Mark Ivanov, Mikhail Gorshkov, Viktor Zgoda and Sergei Moshkovskii

In this work we present a detailed investigation of the alkylation step that usually take place during proteomic sample preparation. The reactions of reduction and alkylation are often used for elimination of the S-S interactions in cysteines for more efficient trypsin digestion of proteins for mass spectrometry in bottom-up proteomics.

Proteogenomic approach has been used together with the direct analysis of modifications with search engines. The search engines used were X!tandem with MPscore post-processing [Ivanov, Levitsky et al., 2013] and Andromeda embedded in MaxQuant package. First, the methionine to isothreonine conversion was noticed during proteogenomic processing of the data from NCI60 cell line panel (data taken from [Moghaddas Gholami, Hahne et al., 2013]) against a customized database including cancer mutations. The NCI60 data had been acquired after the in-gel way of sample preparation. Then, we tested the in-solution protocol for such Met to isoThr conversion events using 4 different alkylation agents: iodoacetamide (IAA), 4-vinylpyridine (4-VP), methyl methanesulfonate (MMTS) and chloroacetamide (CAA). Also, the efficiency of all 4 agents was estimated: the number of modified cysteines, identified peptides, PSMs and oxidized methionines. These experiments were held on two cell lines: HL-60 and HepG2.

Met to isoThr conversion was noted in the same place in 9 out of 10 cell lines and was shown not to be encoded genetically [Chernobrovkin, Kopylov et al., 2015]. Such phenomenon is important as it can be easily confused with a real Met to Thr substitution during proteogenomic investigations. This modification happens if the samples undertake in-gel processing and are treated with high concentration IAA. This effect was also shown on serum albumin treated in-solution with 50mM IAA, which is not normally used for in-solution digestion. In the data of HL-60 and HepG2 proteomes acquired after standard 10mM alkylation and in-solution processing there were no significant Met to isoThr conversion events.

Four alkylating agents show different efficiency, but the dependence of the number of modified cysteines on the peptides identified (i.e. on the overall success of the analysis) is not obvious. MMTS turned out to cause the smallest number of modified Cys in both cell lines, though, it led to the biggest number of identified peptides in HL-60 and the third biggest for HepG2. While the number of identified peptides after the 4-VP treatment have direct dependency on the incubation time, for MMTS in both cell lines and for IAA in HepG2 this relationship is inverse. Such surprising, at first glimpse, results can be explained

by a wider variety of factors effecting the reaction of trypsin digestion and, thus, the success of mass spectrometry analysis.

The importance of sample preparation effects investigation was first shown during peroteogenomic data processing and then resulted in a wider bioinformatic work on chemical modifications of amino acids in peptides. This topic requires future development, as it can be useful to achieve more significant and effective proteomic analysis.

5. Electronic sample management and archiving system for proteomics MS-data

Corinna Klein, Sascha Neinert, Peter Frommolt and Christian Frese

The electronic management and organization of meta information on experimental settings of proteomics projects and related samples is fundamental for proteomics core facilities serving as a high quality service within larger research institutions. Especially the combined and structured storage of mass spectrometry raw data together with its associated experimental meta information supports the sample processing and subsequent bioinformatic data analysis through facility staff.

We present a newly developed web portal, specifically adapted to the requirements of organizing and maintaining experimental meta data at a proteomics core facility. It is based on a back-end MySQL database, php-scripts, common javascript libraries, and css, meant to run locally on institute's hardware.

It serves as a project's archive ensuring to store a detailed electronic overview on active and completed projects for both sides - scientists and facility staff. The portal basically provides easy storage, access and retrieval of supporting sample and MS-measurement information not only for data analysis but also for publication, which in biomedical research usually happens much later after the acquisition of the underlying data. It is separated into a researcher and a facility domain, allowing common views on the data but covering different functionality. The first one consists of a password-protected research group-specific domain that provides a synopsis on the project's and sample's metadata as well as the analysis status. Specifically developed web forms are presented to ask for standardized, comprehensive background information on projects and samples from the users, preventing the lax completion of paper forms or the loss of them. Furthermore, the portal works as a data distribution source where scientists can download the bioinformatic analysis results of their data, which replaces the sending of large attachments via email.

The facility domain provides a complete overview on all projects and samples as well as on LC-MS parameter linked to each sample, e.g. the identity of the UPLC columns or the organization of samples provided in 96-well plates for injection into the LC-MSMS system. Thus, the portal serves as a convenient electronic interface between researchers and facility staff easing the daily work of proteomics core facilities.

6. An integrated proteogenomics approach to discover the entire protein-coding potential of prokaryotic genomes

Adithi Varadarajan, Ulrich Omasits and Christian Ahrens

Advances in sequencing and functional genomic technologies have led to the generation of vast amounts of genome, transcriptome and proteome data for several model organisms, providing the basis to gain better insights into their biology and function. In spite of these improvements and large datasets available, we still do not know the entire protein coding potential, not even that of key model organisms. During my PhD I will focus on proteogenomics, an approach that relies on high mass accuracy proteomics data to detect expression evidence for as of yet unannotated short ORFs in genome sequences [1].

In a first phase, I will apply an in house genome sequencing, assembly and annotation pipeline to obtain de novo assembled genomes for several prokaryotic model organisms. Using a combination of PacBio and Illumina MiSeq sequencing data, I will determine the complete genome sequences of Bartonella henselae as a model for an emerging pathogen, Caulobacter crescentus as a model to study basic biological processes, Staphylococcus aureus as medically relevant Gram positive bacterium, Listeria monocytogenes as important food-borne pathogen, and Bradyrhizobium diazoefficiens as model organism of agricultural relevance. In a second phase, I plan to further extend and apply our in house computational proteogenomics pipeline on high mass accuracy proteomics data obtained for these model organisms. This proteogenomics pipeline integrates multiple annotation sources and even a de novo assembled genome sequence to uncover condition specific total protein-coding potential. Using this approach together with an in house designed

integrated genome visualization, I aim to find proteomic expression evidence for already annotated ORFs, novel ORFs, pseudogenes, and correct protein coding termini for these ORFs.

My group has previously devised a generic iterative analysis-driven experimentation (ADE) strategy to describe complete, condition-specific expressed prokaryotic proteomes using Bartonella henselae as a model system [2]. Mining this high quality proteomics dataset along with the de novo assembled genome sequence provided evidence for the expression of a number of short, unannotated ORFs that were missed in the NCBI reference genome annotation. So far, we were able to uncover missed protein-coding ORFs, including short proteins, differentially regulated proteins, membrane-associated lipo-proteins and metabolic enzymes in each prokaryote we tested [3,4].

I will present the manifold implications of our broadly applicable integrated proteogenomics approach, up to the identification of coding differences from clinical genomes sequenced and assembled from long PacBio reads. I will also briefly discuss the next steps of my PhD. Keywords: Proteogenomics, Integrated genome visualization, De novo genome assembly, Novel short ORFs

References

[1]. Nesvizhskii, A. I., Nature Methods 2014, 11:1114–1125.
[2]. Omasits U et al., Genome Research 2013, 23:1916-1927.
[3]. Carlier A et al., Molecular Plant-Microbe Interactions 2013, 26:1325-1333.
[4]. Cuklina J et al., BMC Genomics 2016, 17:302.

7. Identifying tandem mass spectra of phosphorylated peptides before database search using machine-learning

Sebastian Dorl, Viktoria Dorfer, Karl Mechtler and Stephan Winkler

INTRODUCTION

Identification of post-translational modifications (PTMs), for example phosphorylation, is of high interest in proteomics research since modified proteins are often important for biological functionality. For the identification of modified peptides during tandem mass spectrometry, database search engines consider the selected PTMs for any of the spectra in a sample. Selecting many different PTMs together results in drastically increased search space, leading to longer search times and more false positive peptide identifications. To counteract this, we propose the use of machine-learning-trained models that can reliably classify those spectra which are highly likely to represent phosphorylated peptides before database search.

METHODS

Our goal is to limit database search of phosphate as a variable modification to only those spectra in a sample which are most likely to originate from phosphorylated peptides. For this purpose, we use a classification model that separates the raw MS/MS spectra into phosphorylated and non-phosphorylated before they are submitted to database search. After splitting the spectra accordingly, each batch is searched separately using MS Amanda (ProteomeDiscoverer 2.1, 5 ppm precursor mass tolerance, 20 ppm fragment mass tolerance, modifications: static carbamidomethylation of cysteine, variable oxidation of methionine and variable phosphorylation of serine or threonine, UniProt Swiss-Prot protein database). Then, results for the phosphorylated and non-phosphorylated batches are independently filtered to 1% FDR and combined.

The classification model uses a set of features that represent the occurrences and relative intensities of all neutral losses commonly observed in tandem mass spectra of phosphorylated peptides. The model was constructed with the random forest algorithm for supervised learning on a training data set consisting of 161,379 phosphorylated spectra and 164,859 non-phosphorylated spectra. We obtained the training spectra by combining data from several phosphopeptide-enrichment experiments publicly available in the PRIDE repository, including a variety of human cell lines and different kinds of mouse tissue samples (exclusively high-accuracy fragment ion spectra measured on QExactive instruments). The raw data was re-analyzed using MS Amanda and the resulting matches were filtered for high confidence identifications ($<$1% FDR using Percolator and $>$200 search engine score).

RESULTS

The classification model for fragment spectra of phosphorylated peptides achieved an average accuracy of 97.1% in 10-fold cross-validation experiments while correctly identifying an average of 96.4% of phosphorylated spectra. The model-assisted workflow was tested using data from an experiment on

mouse kidney samples which was completely separate from the training data pool. In this case, using the model-assisted workflow reduced the total search space of the database search by 45.6% while showing 99.4% as many peptide spectrum matches and 99.5% as many identified unique peptides as an equivalent standard workflow. Besides significantly reducing search time, removing non-phosphorylated spectra before database search also reduces the number of false positive identifications which can lead to increased identifications at the same FDR rate. Subsequently, we tested the split workflow in conditions where high numbers of false positive identifications are a major issue. This includes database searches with many possible variable modifications, searches in big uncurated databases, and database searches in phosphoproteome experiments without phosphopeptide enrichment.

8. Uncovering modified peptides in human sperm proteome

Thibault Robin, Lydie Lane, Frederique Lisacek and Markus Mueller

A significant number of unidentified tandem mass (MS/MS) spectra can be explained by peptides containing post-translational or chemical modifications. Open modification search (OMS) approaches allow identifying modifications without prior knowledge, which comes at the cost of reduced sensitivity compared to standard sequence searches due to the increased search space. To boost the identification of modified peptides, we use OMS as a first step and the most abundant modifications are then re-searched using standard sequence search tools, assuming that a modified peptide co-exists with its unmodified version in a sample.

We apply this workflow to human sperm MS/MS data from the HPP project to identify new modified spectra and peptides in missing proteins, the existence of which is lacking experimental evidence. We first extracted the unidentified spectra based on the peptide-spectrum matches obtained from the initial HPP search. The identified spectra were grouped together to generate consensus spectra and create a target library. Our in-house tool MzMod was then used to match all the unidentified query spectra against the target library in OMS mode. The most abundant mass shifts found corresponding to known modifications were finally configured as variable modifications to perform a new search.

We identified 877 additional modified spectra in 131 of the 248 target missing proteins, corresponding to 302 unique peptides of which 110 were not found in the results of the HPP search. Interestingly, our approach also led to identify new peptides for 7 of the 53 missing proteins for which only a single peptide was found.

9. MsViz, a zero learning curve graphical software tool for detailed manual validation and quantitation of post-translational modifications

Roman Mylonas, Trinidad Martin, Alexandre Masselot, Patrice Waridel, Manfredo Quadroni and Ioannis Xenarios

Introduction:

Mass spectrometry has become the tool of choice for the identification, localization and quantitation of post-translational modifications. Recent technological developments allow proteome-wide studies of PTM's, especially phosphorylation, acetylation and ubiquitination. While data on PTM's on thousands of proteins can be obtained, fully deciphering the complexity and combinatorics of PTM patterns observed even on a single protein remains often a challenge. At the same time, functional investigation of PTMs on a protein of interest requires the exact location of PTM's in the sequence and the accurate quantitation of its changes across several conditions. Software tools for large scale analyses are highly effective but are mostly not conceived for interactive, in-depth evaluation of PTM's on a single protein of interest.

Methods:

Data-dependent LC-MS/MS data were acquired on Orbitrap mass spectrometers at high resolution in both MS1 and MS2. Data for protein identification were imported from MASCOT and MaxQuant database searches after conversion to MzIdent format. Raw files were also converted to MzML for import in the MsViz database. MsViz was developed as a client-server application using Scala and the Play framework. Data are stored in a MongoDB database allowing to easily scale up to large amounts of data. The frontend was developed using Javascript and HTML5 which communicate with the backend through a REST API. MsViz runs on a Java Virtual Machine and can therefore be installed on any common operating system.

Results:

MsViz can import database search results for several runs by parsing MASCOT or MaxQuant outputs. MS1 information is imported from mzML converted files. MsViz displays the sequence of a single protein of interest across several samples, and graphically shows the sequence coverage, the number and position of peptide spectrum matches (PSMs) and the location of all PTMs identified in all samples. Starting from this global view, the user can zoom in on selected stretches of sequence for which MsViz can display, in a single visualization, all the PSMs covering that sequence with the positions of PTMs. Individual PSMs can be chosen and, with a single click, both the relative MS2 spectrum and the extracted ion chromatogram (XIC) are retrieved. Both MS2 spectra and XIC traces can be shown simultaneously. Data for PSMs within the same run or across different runs can be freely selected for display and aligned for comparison. With a single mouse drag it is then possible to quantitate XIC peaks from the selected traces. Quantitative values thus obtained are easily exported to tables together with information on the PSM considered (sequence, mass, charge, RT, ID score, localization probabilities).

We tested and validated the software against quantitative values obtained by manually exploring data with native software (Xcalibur). We estimate that MsViz can decrease by 5- to 10-fold the time required for manually validating and quantitating PTMs and at the same time provides a much better overview of all PSM evidence covering modifications of interest. MsViz is web-browser based and thus OS-independent and requires virtually no training.

10. Proteomic and phospho-proteomic analysis Super-SILAC labeled human tumors using myProMS

Patrick Poullet, Alexandre Sta, Guillaume Arras, Stephane Liva, Damarys Loew and Emmanuel Barillot

Cell-wide profiling of differentially expressed proteins under many perturbation conditions, or between normal and disease states (e.g. cancer) is becoming a mandatory counterpart of genomic and transcriptomic approaches to capture the complexity of the biological processes at work. Mass spectrometry (MS) is a widely adopted technology in proteomics allowing the characterization of proteins and post-translational modifications (PTMs) from biological samples of interest. The Super- SILAC approach [1] is a powerful solution to quantify proteins and PTMs from tissue biopsies at a proteome scale. However, efficient data integration, annotation and processing are necessary to translate the complex data generated into meaningful biological information. To achieve this goal, we have been developing myProMS [2] (http://myproms-demo.curie.fr), a web-based tool that optimizes coherent management and analysis of large datasets by project's collaborators according to their expertise level.

Typically, peptide/protein identification data generated by commonly used database search engines (Mascot, Sequest, MaxQuant/Andromeda...) are imported into myProMS within a defined experimental context. When available, quantification data may also be imported. Alternatively, myProMS relies on the MassChroQ [3] software to perform peptide quantification. FDR and quality control thresholds can be applied before further data processing. Biological samples (biopsies, cell lines...) and associated annotations can be recorded and linked to the data to help biological interpretation following downstream statistical analyses. Dedicated algorithms are used to perform data normalization and differential analysis, focusing on whole proteins or PTMs (e.g. phosphorylation). Complex designs including fractions, technical/biological replicates and multiple biological states are handled. Moreover, exploratory (PCA, clustering...) and functional (Gene Ontology/pathway...) analysis tools are provided to investigate data structure and help biological interpretation. Results are displayed through interactive graphical interfaces to ease data mining and quality check.

Because missing values remain a major challenge in discovery proteomics, we have improved our workflow to better deal with this issue. In addition, we have implemented a statistical test called SSPA (State-Specific Protein Analysis) that relies on non-random missing values distribution across samples to extract additional knowledge such as potential biomarkers identification.

We will present a case study of 42 Super-SILAC labeled Medulloblastoma tumor biopsies corresponding to a dataset of over 1000 MS runs.

References:

1-Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nature Methods 2010, 7:383-5.

2-Poullet P, Carpentier S, Barillot E. myProMS, a web server for management and validation of mass spectrometry-based proteomic data. Proteomics 2007, 7:2553-6.

3-Valot B, Langella O, Nano E, Zivy M, MassChroQ: A versatile tool for mass spectrometry quantification. Proteomics 2011, 11:3572-7.

11. Quantitative assessment of histone deacetylase efficiency in a closed system with label-free mass spectrometry

Sander Willems, Filip Van Nieuwerburgh, Lieven Clement, Elisabeth Govaert, Laura De Clerck, Simon Daled, Maarten Dhaenens and Dieter Deforce

Introduction:

Histone proteins and DNA are the main components of chromatin. Their interaction is largely dictated by the many post-translation modifications (PTMs) present on histones. As a result, different combinations of histone PTMs have different epigenetic effects. One common type of histone PTMs is lysine acetylation. Such an acetylation can be added by histone acetyltransferases (HATs) or removed by histone deacetylases (HDACs). However, the efficiency of removing acetylations with HDACs is dependent on many factors, including recognition/binding-efficiency of these HDACs. As histones can be intensely modified, neighboring PTMs can thus affect the efficiency of HDACs to remove acetylations. In a closed system containing histones and HDACs, quantitative changes of acetylated histone peptides and their unmodified counterparts are expected to be inversely correlated over time. Here, we use this inverse correlation in a functional data analysis approach to assess the quantitative efficiency of HDACs.

Material and methods:

A single batch of extracted histone proteins was equally divided over several vials. An equal amount of HDACs was added to different vials at time points: 0 hours, 2 hours, 4 hours, 5 hours, 5:30 hours, 5:45 hours and 6 hours. After 6 hours, all samples were snap-frozen to block subsequent (bio-)chemical reactions, resulting in HDAC incubation times of 0 min, 15 min, 30 min, 1 hour, 2 hours, 4 hours and 6 hours. Hereafter, a standard histone MS preparation protocol was applied to all samples: chemical derivatization to block lysine residues, digestion of proteins into peptides using trypsin, a second round of propionylation to propionylate peptide N-termini, LC separation, and HDMSE data acquisition using a Synapt G2-Si.

All acquired data was aligned to obtain quantitative measurements for each peptide ion at each time point. Smoothed splines, i.e. continuous and derivable functions, were fitted for each peptide ion by penalized regression in R. These smoothed splines were then clustered using e.g. principal component analysis. Correlations were calculated between clusters and individual peptide ions.

In parallel, peptide ions were identified. Both identified and unidentified peptide ions were considered nodes in a peptide-ion-network. Different types of edges were set: a) Mass-edges between (un)identified peptide ions with mass shifts corresponding to known PTMs e.g. 42.010565 Da for an acetylation, b) Correlation-edges between (un)identified peptide ions with correlated smoothed splines and c) Identification-edges between identified peptides with equal backbones but different PTMs.

Results:

At the time of writing, three primary clusters were detected: Unaffected peptide ions, down-regulated peptide ions and up-regulated peptide ions. Some peptide ions from the up- and down-regulated clusters showed a high correlation that could be used to set correlation-edges in the peptide-ion-network. Combined with mass-edges and identification-edges, HDAC efficiency was indeed found to be variable at different modification sites.

12. Collagens in colorectal liver metastasis

Nick A. van Huizen, Robert R.J. Coebergh van den Braak, Michael Doukas, Lennard J.M. Dekker, Theo M. Luider and Jan N.M. Ijzermans

Introduction

In 2015 colon cancer was the second most identified cancer type in the Netherlands (15500 new diagnoses) with 30-40% chance of liver metastasis (colorectal liver metastasis, CRLM) and 30% 5-year survival rate for CRLM. At present, 6-month CT-scans and blood carcinoembryonic antigen (CEA) exams are performed during 5-year follow up.

We reported on a collagen alpha-1(I) natural occurring peptide (COL1A1-NOP) in urine as new biomarker, that in combination with CEA, increased the sensitivity and specificity to 85% and 84%, respectively. Still, these parameters do not allow a change in follow up programs, because the sensitivity and specificity are similar to the currently used techniques.

Aim

COL1A1-NOP in urine indicates that there might be a difference in collagen expression in CRLM compared to healthy liver tissue. Therefore the aim of this study was to determine expression differences of collagens in CRLM compared to healthy liver tissue.

Methods

Formalin fixed paraffin embedded tissue of CRLM and matching healthy adjacent tissue (control) of 30 patients was provided by the pathology department (Erasmus MC, Rotterdam, the Netherlands). Formalin cross-links were removed by incubation with TRIS, followed by reduction, alkylation, and trypsin digestion. Resulting peptides were identified by analysis with a nanoLC-ESI-Orbitrap Fusion mass spectrometer. Mass spectrometry data was searched using MASCOT against the UniProt/Swiss-Prot human database with hydroxylation of proline and lysine as variable modification and processed with Scaffold, and Progenesis QI. Significance of a protein was studied by ranking the individual peptides and summation of the ranks per sample, the summed ranks were further analysed. Student's t-test was used to determine if peptides or corresponding proteins were significantly different between normal and tumour tissue, a p-value below 0.05 was considered as significant.

Subsequently, a cross-validation with immunohistochemistry (IHC) was performed on one of the significant collagens. Stained tissue sections were scored by a pathologist.

Results

Permutation testing showed that the number of significant different peptides (28.1%) in the true data set was significantly higher than would be expected ($p<0.001$). Normal collagen tissue shows intra- and intercollagen alpha chain correlations as expected by literature.

In 18 out of 20 collagen alpha chains the number of peptides that were upregulated exceeded the number of peptides that were downregulated. Also, summed peptide ranks showed that 18 out of 20 collagen alpha chains were significantly upregulated.

Collagen alpha-1(XII) (COL12A1) was significantly upregulated in the mass spectrometry data, cross-validation with IHC also showed COL12A1 to be present in 28 CRLM tissue sections (93%), and in none of the control tissue. No significant correlation was observed between IHC and mass spectrometry data.

Conclusion

Our study showed that collagens are strongly upregulated in CRLM tissue in comparison to control tissue. Which is likely due to tissue remodelling by the proliferation of metastasising colon cancer.

In addition, the mass spectrometry data of COL12A1 was successfully cross-validated with IHC. COL12A1 was previously described in literature as a marker of tumour cells along the desmoplastic invasion front in colorectal cancer. Further research is required to study the role of COL12A1 in CRLM.

13. Shotgun proteomics data from laser capture microdissected blood-brain barrier

Marina Zajec, Dana Mustafa, Diana Nijholt, Lennard Dekker, Johan Kros and Theo Luider

Aim

The blood-brain barrier (BBB) serves as defense system for the brain controlling the passage of molecules (e.g. nutrients). Additionally, BBB can regulate passage of drugs thereby complicating treatment of various brain diseases (brain tumors, neurodegenerative disorders and cerebrovascular diseases). BBB consists of brain endothelial cells and their tight junctions with specific pinocytic properties, basal membrane which is produced by pericytes and astrocytes, microglia and possibly other cells.

Aim of this study was identification of proteins that are specific for basal membranes of the BBB. As control samples we collected pial vessels (PV, situated in the subarachnoid cavity, surrounded by cerebrospinal fluid) as they do not qualify as BBB, and surrounding tissue (ST, brain tissue surrounding the vessels).

We used a combination of LCM and nanoLC-Orbitrap mass spectrometry to study BBB using normal human formalin-fixed, paraffin-embedded (FFPE) tissue. Laser capture microdissection (LCM) offers the

possibility to collect fine structures from tissue sections. LCM relies on a laser beam cutting tissue area of interest and produces high-quality, enriched tissue samples.

Methods

Paraffin blocks of the frontal cortex of seven autopsy brains were acquired from the archives of the Department of Pathology, Erasmus MC (Rotterdam, Netherlands). Three types of samples were laser microdissected (Zeiss PALM MicroBeam IV microscope) from tissue sections (BBB, PV, ST) and a total of 21 samples was collected.

LCM samples were reduced and alkylated followed by trypsin digestion and measured on the Orbitrap Fusion. Obtained mass spectrometry data were analyzed with Mascot Daemon (Matrix Science) and Scaffold (Proteome Software). Quantitative analysis was performed using ANOVA, p-values were corrected for multiple testing using Benjamini-Hochberg correction, p-value of <0.05 was considered as significant. Protein lists were searched for statistically significant proteins identified exclusively in BBB.

Results

Fourteen proteins were exclusively detected in BBB and two were chosen for immunohistochemical (IHC) validation. Tripeptidyl peptidase 1 (TPP1) was identified in most samples (n=6) and it is not mentioned in literature to be associated with BBB. P-glycoprotein (ABCB1), (n=4), a highly researched transporter, is not known to be exclusively present in BBB.

Due to small sample number, and therefore low power of statistics, we also performed black and white analysis, looking at proteins detected only in BBB but not supported by statistics. Membrane protein MLC1 is a protein located in astrocytes and offers the opportunity to distinguish between BBB and PV. It was identified in one BBB sample by two unique peptides. IHC validation showed remarkably strong staining for MLC1 in astrocytic end feet surrounding the BBB, and was not observed in PV.

Conclusion

We have identified BBB-specific proteins by using microdissection and mass spectrometry. Future studies should focus on the lower abundant proteins, detected by improved sample collection methods, data analysis and faster ways for validating proteomics results.

14. MasPy – mass spectrometry-based proteomics data analysis with python

David Hollenstein and Jakob Hollenstein

Modern liquid chromatography-mass spectrometry (LC-MS)-based proteomics experiments generate huge amounts of raw data. Analysis of such raw data-sets often comprises complex analytical workflows and depends on high-throughput bioinformatic tools. Various software packages have been developed to automatically analyze LC-MS data, as for example MaxQuant, Proteome Discoverer, OpenMS or Crux. These suites feature advanced algorithms for efficient and reliable identification and quantification of peptides and proteins. Furthermore there is also a large number of smaller programs and tools that have been developed to perform only one or a few specialized tasks of a given data processing workflow. However, many of the established software packages and tools have been developed with very specific analytical problems in mind. Therefore it may not always be possible to apply the software to non-standard data or to extend it to meet special requirements of data analysis. This might be especially difficult for researchers who have not been involved in the development of the software, as proper code documentation is a time intensive task and often neglected during the typically fast-paced development of scientific software.

To overcome some of these limitations, we have developed MasPy, an open source Python library with the aim to provide simple, convenient and versatile access to proteomics data. To achieve this it features an internal data representation that facilitates comfortable use of this data. MasPy is intended as a tool for researchers that allows the combination of different software to generate customized and scriptable data processing workflows and to enable interactive data analysis with Python. Thereby external tools are either executed separately or if possible controlled via an API with Python. Our aim is to seamlessly integrate MasPy into the prevailing software environment. MasPy fully supports reading and writing of the proteomics standards initiative (PSI) format mzML, which is a file format for representation of raw mass spectrometry results. We are also planning to include support for other data formats developed by PSI. Data processing algorithms implemented in MasPy are currently focused on the integration of different data types produced by external tools and include for example functions to match peptide spectrum matches

to LC-MS peptide features, perform a retention time alignment of LC-MS runs or group corresponding features between multiple LC-MS runs.

MasPy opens up numerous possibilities for data analysis, as for example the quantitative analysis of cross-linking experiments using isotope-labeled cross- linkers or label free quantification between LC-MS runs. We have also applied it to compare the performance of different data analysis software or to investigate the efficiency of sample preparation methods. Moreover, inspired by the freeware program "RawMeat", we have used MasPy to develop a Python based graphical user interface for viewing quality control reports of mass spectrometer files in the mzML format. We believe that MasPy has the potential to become a valuable tool for bioinformaticians and advanced mass spectrometry users by providing a framework that simplifies many steps needed for accessing and processing data, thereby allowing researchers to focus on their task at hand.

15. Mapping variant and novel peptides to reference proteins

Georg J. Pirklbauer, Christoph N. Schlaffner and Jyoti S. Choudhary

A key goal of genetic research is the identification of variants underlying specific phenotypes. These missense mutations can have significant impact on the structure and functions of proteins in a cell. Genomics has proven successful in identifying somatic variants on a large scale. However, advances in mass spectrometry instrumentation now enable proteomics to capture almost complete proteomes. The integration of these with underlying genomics datasets facilitate the identification of alterations not captured in reference protein databases.

Proteogenomics uses customized protein sequence databases incorporating genetic alterations and gene predictions for database searching to allow the identification of variant peptides. Furthermore, de novo identification of variants through spectrum clustering also results in peptide sequences not present in a reference database. To provide annotators with evidence for expressed variants and prevent identification of false novel protein coding loci it is crucial to map identified variant peptides to reference proteins. Imperfect string matching is a nontrivial problem and numerous solutions have been proposed for text mining, spanning from alignment algorithms to suffix trees. Time-consuming preprocessing steps, however, make them impractical for proteogenomics applications. Here we describe a fast method to identify all occurrences of peptides with up to two mismatches in a reference protein database thorough k-mer indexing the database and peptide length dependent generation of all mismatch combinations within a peptide. Our algorithm is incorporated in the genome mapping tool PoGo.

To demonstrate the effectiveness of our method, we performed imperfect matching of 230,000 unique high confidence peptide sequences from three large scale human tissue proteome datasets against the translation of annotated protein coding transcripts from GENCODE (v20) resolve high complexity regions in the protein coding genome. Our tool exhibited superior performance on benchmark against another tool, PGx, and was able to identify two additional blocks of 8 amino acids in repeat region of the SPRR3 gene. These were validated through peptides identified with the exact sequence. Furthermore, our method was able to map peptides across species between human and mouse. These data show that our method is effective in identifying parent reference proteins for variant peptides even across the phylogenetic tree. The genomics efforts of sequencing different strains of the same species and missing visualization tools pose the challenge mapping identified peptides onto the reference. Our algorithm has been developed with this in mind: we anticipate it will have a central utility for interpretation of trans-strains and species datasets as well as studies on personal variation and precision medicine.

16. Extending the coverage of immune signaling pathways in response to Salmonella infection, by analyzing efficiently enriched phosphopeptide samples before and after phosphatase treatment

Roland Dreier, Erik Ahrné, Petr Broz and Alexander Schmidt

Introduction

Innate immune detection of intracellular Gram-negative bacteria (e.g. Salmonella) is largely based on the recognition of cytosolic bacterial LPS (Lipopolysaccharide).

In murine macrophages, cytosolic LPS can activate an inflammasome including caspase-11. Since phosphorylation is a main regulator of many signaling cascades, we analyzed the global cellular phosphorylation changes within the first 8 hours upon Salmonella infection to characterize which innate immune signaling cascade are important for caspase-11 activation and establish their temporal dynamics.

Methods

We extracted and digested proteins of primary wt/B6 bone marrow derived macrophages (BMDM) infected with Δ orgA Salmonella Typhimurium for 0.5, 1, 2, 4 and 8 hours and subjected them to phosphopeptide enrichment using TiO2. To extend phosphorylation site coverage we analyzed our efficiently enriched phosphopeptide samples before and after phosphatase treatment. Next, we employed a random forest classification algorithm to estimate the False Discovery Rate amongst the novel phosphopeptide candidates found in the phosphatase treated samples. This method allowed us to identify a set of phosphorylation specific peptide features.

All samples were analyzed by 1D-LC-MS using a Q Exactive-HF hybrid quadrupole-Orbitrap. Label-free quantification was performed using the Progenesis QI software followed by statistical analysis using the SafeQuant R package.

Conclusion

Employing our standard phosphoproteomic analysis workflow, around 5000 phosphopeptides could be quantified across all 6 time points with more than 700 showing a significant change in at least one time point. The corresponding phosphoproteins cover many known immune signaling pathways, including TLR, IFN, NF-kB, NOD1/2 or MAPK signaling. However, the coverage of several pathways was incomplete and known phosphoproteins (e.g. Jak/Stat) were not found in the initial dataset. When extending our workflow to include a parallel phosphatase treatment step the coverage of activated signaling pathways improved dramatically. Interestingly, we observed that the removal of phosphorylation groups, increased the MS signal response by 5-10-fold, ultimately allowing us to double the number of quantified phosphopeptides.

17. Proteome and transcriptome analysis enabled the identification of the differences in the innate immune response towards an adjuvant and a complete vaccine formulation

Sietske Kooijman, Jolanda Brummelman, Fabio Marino, Gideon Kersten, Bernard Metz, Albert Heck, Jeroen Pennings, Cecile van Els, Elly van Riet and Hugo Meiring

Introduction/ objectives:

Aluminum based adjuvants are the most widely used adjuvants in human vaccines, however the role of the adjuvant in the induction of the innate immune is still elusive. In this study we compare the innate immune response directed to a complete vaccine formulation (infanrix GSK® or aP) with the response to the adjuvant (Al(OH)3) alone, by a system-wide analysis in monocytes to determine what the influence is of an adjuvant on the innate immune response induced by vaccines.

Materials and methods:

Monocytes were isolated from human blood donors and cultured with Al(OH)3, Al(OH)3 containing vaccine aP or in culture medium for 24 and 48 hours. Cell surface markers were analyzed using flow cytometry and supernatants were used for ELISA. In addition, the cells were used for targeted transcriptome analysis on 84 genes from the innate and adaptive immune system as well as a comprehensive proteome analysis using TMT(6) isobaric labeling. Pathway analysis was performed on proteome and transcriptome data sets to annotate the relevant pathways involved in the immune response.

Results:

Flow cytometry data revealed a very similar cell surface marker profile on the selected markers (CD40, CD80, CD83, CD86) however, whole proteome data revealed differences in expression of other markers e.g. CD9, CD44 and CD99, CD101 and CD180. Transcriptome data revealed a major difference in Interferon expression and in IL-2 and IL-10 expression, IL-2 was upregulated after Al(OH)3 stimulation while IL-10, was upregulated after aP stimulation. It is known that IL-10 has an inhibitory effect on the formation of IL-2 and IFNγ. IL-2 and IFNγ were not upregulated after aP stimulation.

Proteome data confirmed the formation of interferon gene transcripts after 24 hours of Al(OH)3 stimulation, since proteins downstream of type I and type II interferons were formed aP stimulation did not induce these proteins after 24 hours and some after 48 hours. With respect to proteome data from cells stimulated for 24 hours, aP stimulation resulted in over representation of 4 KEGG pathways e.g:

endocytosis, and complement cascade and no GO terms. In contrast Al(OH)3 stimulation resulted in an overrepresentation of 5 KEGG pathways including antigen processing and presentation and multiple GO terms including exocytosis. Other GO terms were related to homeostatic processes.

After 48 hours, aP stimulation resulted in the enrichement of more immunogenic GO tems "antigen processing and presentation"and "response to molecule of bacterial origin" while this was not the case for Al(OH)3 . Noteworthy after 48 hours of aP stimulation the response to IFNγ was also upregulated, which fits the downregulation of IL-10 at transcript level.

Discussion/conclusion

By combining proteome analysis, targeted transcriptome analysis and flow cytometry we were able to analyze the different pathways in monocytes from healthy blood donors after stimulation with Al(OH)3 alone and an Al(OH)3 containing vaccine AP. The results indicate that the innate immune response induced by this adjuvant is significantly altered by the other vaccine components and that some responses are delayed or inhibited in the Al(OH)3 containing vaccine e.g the production and secretion of interferons.

19. Mass spectrometrists should search for all peptides, but assess only the ones they care about

Adriaan Sticker, Lennart Martens and Lieven Clement

Reliable peptide identification is key to every mass spectrometry-based shotgun proteomics workflow. The growing concern on reproducibility triggered leading journals to require that all peptide-to-spectrum matches (PSMs) are reported along with their false discovery rate (FDR).

In many experiments, however, researchers want to focus on proteins of particular pathways, or few organisms in a metaproteomics sample. Hence, a large fraction of the reliable PSMs are deemed irrelevant for their scientific hypothesis and are often discarded (classical method). Within this context Noble (2015, Nature Methods ,pages) urged researchers to remove irrelevant peptides from the database prior to searching to improve statistical power.

We argue that his method produces suboptimal peptide identifications and that its false discovery rate (FDR) control is problematic. Moreover, we also show that the FDR of the classical method, that discards the irrelevant peptides post FDR calculation, is wrong. Instead, we propose that searching for all expected peptides, and removing irrelevant peptides prior to FDR calculation results in a more reliable FDR control. Unfortunately, the uncertainty on the FDR estimation increases for smaller subsets.

Therefore, we develop a more stable method for FDR estimation in subsets by exploiting the information on all PSMs. We provide diagnostic plots to check the assumptions needed to use this additional information. We promote the use of our method within the proteomics community by providing a user friendly tool.

20. Alpha-, beta- and gamma-synuclein quantification in cerebrospinal fluid by multiple reaction monitoring reveals increased concentrations in Alzheimer's and Creutzfeldt-Jakob disease but no alteration in synucleinopathies

Patrick Oeckl, Fabian Metzger, Magdalena Nagl, Christine A.F. von Arnim, Steffen Halbgebauer, Petra Steinacker, Albert C. Ludolph and Markus Otto

Alpha-synuclein (alphaSyn) is a major constituent of proteinaceous aggregates in neurodegenerative diseases such as Parkinson's disease (PD) and a potential biomarker candidate for diagnosis and treatment effects. However, studies about alphaSyn in cerebrospinal fluid (CSF) in diseases are inconsistent and mainly based on immunological assays. Quantitative information about beta-synuclein (betaSyn) and gamma-synuclein (gammaSyn) in CSF is not available.

Here, we present an alternative method for the simultaneous quantification of alphaSyn, betaSyn and gammaSyn in CSF by multiple reaction monitoring (MRM) with a high sequence coverage (70%) of alphaSyn to validate previous, ELISA-based results and characterize synucleins in CSF in more detail.

The MRM has high sensitivity in the low pg/mL range (3-30pg/mL full-length alphaSyn) using 200μL CSF. A high portion of CSF alphaSyn is present in the N-terminally acetylated form and the concentration of unmodified peptides in the non-amyloid component region is about 40% lower than in the N-terminal region. Synuclein concentrations show a high correlation with each other in CSF (r$>$0.80) and

in contrast to alphaSyn and gammaSyn, betaSyn is not affected by blood contamination. CSF alphaSyn, betaSyn and gammaSyn concentrations were increased in Alzheimer's and Creutzfeldt-Jakob disease but not altered in PD, PD dementia (PDD), Lewy body dementia and atypical parkinsonian syndromes. The ratio betaSyn/alphaSyn was increased in PDD (1.49±0.38, $p<0.05$) compared with PD (1.11±0.26) and controls (1.15±0.28). BetaSyn shows a high correlation with CSF tau concentrations ($r=0.86$, $p<0.0001$, $n=125$) In conclusion, we could not confirm previous observations of reduced alphaSyn in PD and our results indicate that CSF synuclein concentrations are rather general markers of synaptic degeneration than specific for synucleinopathies. Betasyn is an attractive biomarker candidate that might be used as an alternative to or in combination with tau in AD and CJD diagnosis and in combination with alphaSyn it is a biomarker candidate for PDD.

21. Functional Analysis of the Conserved Interactome of NUP98-Fusion Proteins in Acute Myeloid Leukemia

Stefan Terlecki-Zaniewicz, Thomas Eder, Johannes Schmöllerl and Florian Grebien

Leukemia summarizes a group of hematopoietic malignancies that are characterized by malignant cells crowding out healthy blood cells. While the etiology of leukemia appears to be multifactorial, the final common pathway leads to DNA damage, resulting in the introduction of oncogenic mutations. Chromosomal rearrangements in cancer can lead to the fusion of two gene loci, resulting in the production of a pathogenic fusion protein. In leukemia, a particular high number of fusion oncogenes has been identified.

Fusion proteins involving the NUP98 gene (Nucleoporin 98) are found in 2% of acute myeloid leukemia (AML) patients. The NUP98 protein itself is part of the nuclear pore complex (NPC) and is found on both nucleoplasmic and cytoplasmic sides of the NPC, where it participates in the mediation of transport of macromolecules through the NPC.

The NUP98 multi-partner translocation family (MPTF) in leukemia features >25 different fusion proteins, all harbouring the N-terminal part of NUP98 fused to distinct C-terminal fusion partners. Previous studies and data from our group indicate that different NUP98-fusions cause similar AML phenotypes in humans and mouse models. Thus, we postulate that NUP98 fusion proteins share common molecular mechanisms to modulate important oncogenic pathways. The aim of our work is to identify critical common effectors of the NUP98 MPTF among the protein interactomes of five NUP98-fusion proteins by using affinity purification coupled to mass spectrometry (AP-MS) followed by functional investigation of candidates.

Doxycycline (Dox)-inducible, Strep-HA-tagged variants of selected NUP98 fusion proteins (NUP98-HOXA9, -JARID1A, -DDX10, -NSD1 and -PSIP1) were cloned into retroviral vectors. Human AML cells were transduced with the constructs and selected for stable transgene integration. Dox-mediated transgene induction was reported by expression of GFP. Routinely, 70%-90% GFP+ cells could be detected 24 hours after Dox treatment.

We are using these cell line models to characterize protein-protein interactions of wild-type NUP98 vs. exogenous NUP98-fusion proteins. While protein complexes around exogenously expressed NUP98-fusion proteins will be purified through affinity reagents directed against the Strep/HA tag, endogenous NUP98 complexes will be purified by antibodies recognizing the NUP98 N-terminus. The concept of using two different enrichment approaches allows a more stringent, but also more versatile way of data analysis. Parental AML cells and cells expressing an N-terminal NUP98 breakpoint construct will serve as controls.

We have been successful in optimising enrichment of protein complexes around tagged NUP98 fusion proteins. Samples of large-scale affinity purifications are currently used to perform MS experiments and first results comparing StrepTactin and anti-NUP98 antibody immunoprecipitation of control cells are currently being processed.

Resulting data will be subjected to detailed bioinformatics analysis to generate high confidence interaction maps for each NUP98-fusion protein and to identify overlaps in the interactomes of the selected proteins. After identification of common interactors, we will perform a CRISPR/Cas9-mediated loss-of-function screen to identify essential genes for maintaining leukemogenic potential.

Altogether, this study will provide novel insights into protein-protein-interactions that mediate NUP98-fusion protein-dependent leukemic transformation. It will thereby improve the understanding of oncogenic

fusion proteins to pave the way for new therapeutic targets and individualised treatment.

22. MSqRob: analysis of label-free proteomics data in an R/Shiny environment

Ludger Goeminne, Kris Gevaert and Lieven Clement

Data analysis of bottom-up mass spectrometry-based proteomics experiments is often very challenging . Proteins are almost never completely covered and peptides that are identified in a particular sample are often missing in another sample and vice versa. Common workflows adopt software tools that have point-and-click graphical user interfaces, but often only provide inefficient or even inappropriate statistical inference. Moreover, they are often based on protein level abundance values obtained after summarization of peptide-level data. We have shown that summarization-based pipelines perform suboptimal compared to peptide-based models in terms of sensitivity, specificity, accuracy, and precision. Indeed, peptide-based models account for peptide-specific effects due to different ionization efficiencies as well as differences in accuracy due to a different number of peptides. Here we present MSqRob, an R package that improves existing peptide-level methods by three modular extensions: (1) ridge regression, (2) empirical Bayes variance estimation and (3) M-estimation with Huber weights. MSqRob provides state-of-the-art statistical inference for label-free proteomics experiments with simple and complex designs: MSqRob can cope with multifactorial, block, repeated measures and time series designs, which cannot be analyzed properly in existing proteomics data analysis software. Finally, we also developed a Shiny graphical user interface for MSqRob, which makes the analysis very user-friendly and requires no statistical programming experience. Our presentation will focus on the use of the Shiny MSqRob GUI for data exploration, normalization, fold change estimation and statistical inference. The GUI also provides graphical point-and-click tools to visualize the results in tables and volcano plots, and to evaluate the underlying normalized peptide-level data of the discoveries.

23. Mapping Clinical Omics Data to the Reactome Pathway Database

Luis Francisco Hernández Sánchez, Antonio Fabregat Mundo, Pål R. Njølstad, Henning Hermjakob and Marc Vaudel

Matching experimental data to biomedical knowledge databases can allow finding mechanisms possibly responsible for diseases. It is challenging to decipher the complex omics datasets recorded under different conditions, patients, time points. Mapping clinical data to biological pathways holds the promise to mine large datasets intuitively and possibly uncover trends otherwise hidden in the noise.

However, one of the main difficulties in such pathway analyses is that much of the clinical data is lost in the process due to lack of mapping tools considering any type of omics data. One reference resource for pathway analysis is Reactome, a free, open source, curated and peer reviewed database of biological reactions. It can be readily queried with omics datasets, and we are improving its features by extending the matching the clinical data to the biological pathways. Not only will the gene names be used, but also mutations or post translational modifications such as phosphorylation. Ultimately, any omics dataset will be mapped directly to the functional knowledgebases allowing the functional interpretation by researchers and clinicians.

24. PeptideMapper: Efficient and Versatile Amino Acid Sequence and Tag Mapping

Dominik Kopczynski, Marc Vaudel, Harald Barsnes, Pål R. Njølstad, Albert Sickmann and Robert Ahrends

The mapping of amino acid sequences is an essential task in bioinformatics. Notably, the mapping of peptide sequences on a proteome is required for the post-processing of proteomics results. However, this step can quickly become a bottleneck when working with extensive numbers of peptides or large protein sequence databases. Here, we present PeptideMapper, a novel amino acid sequence mapper for both peptide sequences and de novo sequencing identification results. By taking advantage of the latest advances in pattern matching, PeptideMapper achieves unprecedented performance (i.e. up to 500x faster than state-of-the-art software) in terms of memory footprint and execution speed, with regards to both the indexing and the querying of protein sequence databases. PeptideMapper is implemented in the open source Java CompOmics framework under the permissive Apache 2.0 license https://github.com/compomics/compomics-utilities.

25. MS-Angel: a solution for real-time proteomics workflows

Julie Poisat, David Bouyssié, Alexandre Burel, Christophe Bruley, Sarah Cianférani, Jerome Garin and Odile Burlet-Schiltz

Today search engines are a major tool in proteomics for the identification of proteins through the mass spectrometry-based analysis of complex peptide mixtures. Open source tools now offer algorithms that are equivalent or superior to the commercial ones regarding the quality of the results. However they are often harder to use, many of them relying on a command line interface for their execution. Recently, the SearchGUI [1] tool was developed to ease the submission of files to several open source search engines, providing a simple and user-friendly graphical interface. Nevertheless, unlike Mascot Daemon, it is not able to monitor the creation of new acquisition files in order to handle them in background without any intervention of the user. Considering the increasing number of instruments in the laboratories and the throughput of data they produce, this kind of real-time processing could really simplify and shorten the execution of complex proteomics workflows.

In this context, we created MS-Angel, an open source software based on the client-server model. This tool handles, with a minimal mediation of the user, the automated detection of newly created raw files and their conversion into peak lists, as well as the submission of the resulting MS/MS spectra to one or more search engine(s). These operations can be easily chained by the user through the design of workflows. In this way, MS-Angel broadens the benefits of Mascot Daemon to open source search engines, and maintains the submission of peak lists to the Mascot server.

Moreover, MS-Angel has been developed in order to be coupled to the Proline software suite (http://proline.profi free and open source. This one is dedicated to the processing of mass spectrometry data, especially the target/decoy validation of identification results and the label-free quantitation by spectral counting or by analysis of the LC-MS signal. Using MS-Angel workflows, the user can then schedule the import of the MS/MS search result files into Proline as soon as they are created; then validate and quantify the data all together.

MS-Angel is thus a new free solution, fully automated for the management of raw files and the submission of searches in databases using MS and MS/MS data. Combined to the Proline software [2], it offers a complete and easy-to-use pipeline for the analysis of LC-MS/MS data.

It will be publicly available with the next release of Proline (v2.0), scheduled for the beginning of 2017. The development of a standalone version coming without Proline - thus supporting only the conversion and submission to search engines steps - is in progress.

[1] Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A., Martens, L. (2011). SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics, 11(5):996-9. doi: 10.1002/pmic.201000595

[2] Proline: http://proline.profiproteomics.fr/

26. An unbiased protein association study on the public human proteome reveals biological connections between co-occurring protein pairs

Surya Gupta, Kenneth Verheggen, Lennart Martens and Jan Tavernier

Objective:

Mass-spectrometry based proteomics produces large amounts of data. While typically acquired to answer specific biological questions, these data can also be reused in orthogonal ways to reveal biological knowledge1. Here we present a novel method to unravel biological importance of protein co-occurrence across all human proteomics data in PRIDE2.

Method:

Mass-spectrometry data was obtained from the PRIDE database. To calculate the weight of protein co-occurrence across experiments, we used statistical Jaccard similarity. Protein pairs with a similarity above 0.4 were mapped to four knowledgebases; Reactome, Ensembl, IntAct, and CORUM, to assign potential biological relevance. Moreover, using published articles and String database, we were able to determine the possible biological connection between unannotated protein pairs with no known biological correspondence.

Results:

Of the 2325 protein pairs that pass the Jaccard similarity threshold, we have successfully been able to map 81% of pairs. 68% protein pairs were mapped with four existing biological knowledgebases; Reactome, IntAct, CORUM, and Ensembl, and 13% with possible biological connection using Gene Ontology (GO) terms. For 68% of annotated pairs, 66% protein pairs were found to involve in similar biological pathways using Reactome, 1% were found to either have direct protein-protein interaction using IntAct or were found in a complex together using CORUM, and in remaining 1% protein pairs, both protein were found to be paralog of each other according to Ensembl. Furthermore, we also found 13% protein pairs with similar GO Biological Process or GO Molecular Function.

Conclusion:

We have successfully been able to map the majority of strongly co-occurring protein pairs with existing biological knowledge. Our approach shows that by re-using publically available data in a fully orthogonal way, effectively treating these data as a proteome-wide association study, we can extract various biologically meaningful patterns. In future we will use these patterns to investigate as-yet unknown biology and to predict protein pairs that have a high probability for biological significance.

Reference:

(1) Vaudel, M.; Verheggen, K.; Csordas, A.; Raeder, H.; Berven, F. S.; Martens, L.; Vizcaíno, J. A.; Barsnes, H. Exploring the potential of public proteomics data. Proteomics 2016, 16 (2), 214–225. (2) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: the proteomics identifications database. Proteomics 2005, 5 (13), 3537–3545.

27. Advancing affinity-based protein profiling: The photome

Matthias Stahl, Philipp Kleiner, Wolfgang Heydenreuter, Vadim S. Korotkov and Stephan A. Sieber

Photocrosslinkers are important tools for chemical biology research which mediate irreversible linkages between proteins and ligands upon UV irradiation. A prominent application is affinity-based protein profiling (AfBPP), in which small molecules are equipped with a photocrosslinker in order to facilitate irreversible binding to their protein target and subsequent indentification thereof. One class of photocrosslinkers includes the diazirines, which have recently experienced a renaissance in application. Despite this popularity, very little is known about photocrosslinker-associated off-target binding. Thus, control reactions such as competition with the parent, unmodified ligand are recommended; however, these experiments are challenging due to the irreversible nature of the photoprobe-protein interaction while the competing parent compound binds reversibly. In addition, a minimal photoprobe lacking the ligand specific structure can be applied as a control, but results heavily depend on the detailed structure of the control compound.

In order to obtain a comprehensive picture focused on the needs of accurate target identification, we recorded a full inventory of photocrosslinker off-targets (photome) in human cell lines via quantitative high-resolution proteomics.

In detail, we synthesized a set of four minimal diazirine photocrosslinkers bearing a core diazirine photocrosslinker motif attached to various functional groups. We then applied these to mass spectrometry-based target identification experiments. Peptides were identified and proteins quantified using MaxQuant. Similar to signal-to-noise improvement algorithms in image processing, we assumed that diazirine-related off-targets will be enriched in all of the samples, irrespective of probe structure. However, targets related to the structure of the whole photocrosslinker probe will only appear in individual experiments. We therefore overlaid the quantitative information of each experiment and, consequently, identified a set of highly enriched protein off-targets that can solely be attributed to the diazirine moiety, the only structural motif common to all probes. The hits were classified according to the strength of their enrichment and can easily be mapped on other AfBPP experiments to identify putative off-targets.

In a proof-of-principle study, H8, a protein kinase A inhibitor, was equipped with a diazirine moiety and applied to AfBPP. Alignment of the results with our photo-background data set revealed unprecedented insight into its in situ proteome targets. Taken together, our findings built from proteomics and bioinformatics analyses guide the identification of biologically relevant binders in photoprobe experiments.

28. mzDB tools: creation and processing of mzDB files

Alexandre Burel, David Bouyssié, Christophe Bruley, Christine Carapito, Sarah Cianferani, Jerome Garin and Odile Burlet-Schiltz

In the last decade, mass spectrometers' performances have significantly improved in speed and precision. These benefits came with a size issue, MS analyses containing more and more data and therefore being heavier and heavier. A single MS acquisition can weight several GB. Another problem is that each vendors file has its own proprietary format. Ideally, MS data sets should be stored in a generic and open source format. This format should produce files of reasonable size and data have to be rapidly accessible for processing.

The mzDB format has been presented in 2014 as a solution to these three points:

(i) It is a free and open source generic format for MS data sets.

(ii) The format relies on the SQLite software library, it produces binary files instead of text or XML files. A strict conversion will have a similar size as the raw input file.

(iii) The SQLite format allows a 3D indexing on retention time, m/z and precursor m/z. MS data can be extracted very quickly with simple SQL queries.

The Proteomics French Infrastructure ProFI (http://www.profiproteomics.fr/) is supporting the mzDB format and uses it in the Proline software suite to run extra fast label-free quantitation. In order to use the mzDB format in its workflow, the ProFI team is also developing the following solutions:

- raw2mzDB is a converter able to produce mzDB files from Thermo, Bruker and AB Sciex analyses, and also from mzML generic files. MS data can be stored in their original format or turned from profile to centroid or fitted mode. MS analyses in data dependent acquisition (DDA) as well as data independent acquisition (DIA) are accepted.

- mzScope is an mzDB viewer that allows the user to view the content of an mzDB file, create an mgf peaklist from it, and run different processes such as extracting features from MS2 spectra. It is also possible to display chromatograms from multiple mzDB files in a single graph. Visualizing spectra and XICs is extremely fast and user-friendly thanks to the mzDB format. mzScope is now a part of the Proline suite.

Both solutions are open source and available to the community (http://proline.profiproteomics.fr).

29. Analyzing repeated measures designs in label-free proteomics

Lieven Clement, Ludger Goeminne and Kris Gevaert

Repeated measures designs are experimental designs in which different observations are obtained on the same experimental unit, e.g. measuring the same plants over time, assessing the proteome of knockout and wild type mice before and after administering a compound, ... The advantage of these designs is the increased statistical power for within subject treatment effects because the between-subject variability is eliminated from the estimation.

In repeated measures designs, however, the data are correlated, i.e. the data from the same experimental unit are typically more similar than data between experimental units. Unfortunately, existing label-free proteomics data analysis workflows cannot address correlation structures, properly. This results in a power loss when assessing treatment effects within experimental units (e.g. compound effects) and/or an improper control of false positives when assessing treatment effects between experimental units (e.g. knockout vs wild type).

In this contribution, we present how repeated measures designs can be analyzed properly with MSqRob, an all-round R package for mass spectrometry-based proteomics data analysis with simple and complex experimental designs. MSqRob builds upon the mixed model framework to model peptide level data while acknowledging the correlation structure of the data at different hierarchical levels, i.e. correlation between peptide intensities from the same protein at the level of (a) the technical replicate, (b) biological replicate, (c) repeated measures of an experimental unit, ...

MSqRob allows the user to model the data with fixed and random effects, enabling a proper assessment of the treatment effects in studies with multiple technical repeats and/or repeated measures designs.

With its flexible model structure, it can also correct for peptide specific effects while providing robust inference on a protein level. We illustrate our method on a case study with technical repeats and on a study with a repeated measures design that assesses early and late responses of inflammatory stimuli in knockout and wild type mice.

30. Proline: a software environment for label-free quantification data analysis and exploration.

David Bouyssié, Anne-Marie Hesse, Emmanuelle Mouton-Barbosa, Alexandre Burel, Christine Carapito, Véronique Dupierris, Charlotte Macron, Jean-Philippe Menetrey, Aymen Romdhani, Jérémie Turbet, Odile Burlet-Schiltz, Sarah Cianferani, Jerome Garin and Christophe Bruley

Label-free quantification based on the extraction of MS1 signal is an analysis method commonly employed especially to analyze and compare a large number of samples. The accuracy and reliability of quantitative measurement are critical factors in the implementation of these approaches while the signal processing duration can itself become problematic when the number of LC-MS/MS analyses becomes large. Apart from the intrinsic quality of the results, another important issue to successfully achieve a mass spectrometry-based quantitative experiment is the examination of the results by the user. However the availability to review and correct manually the data is not always provided by existing tools. Indeed, because of the complexity of proteomics data and the increasing size of data generated from large scale studies, data examination poses its own challenges. The first is the relationships and connections between the different information (spectra, peptide-spectrum matches, peptides, protein sets, proteins, etc.) which is not well taken into account by the solutions based on flat-file following a tabular structure. The second challenge is the development of an affordable user interface based on efficient navigation and visualization tools. Proline software tries to address those issues by providing in a single software environment:

- A fast and efficient extraction algorithm of MS signal allowing quantification of a large number of samples
- A robust and consistent quantification method minimizing variance of ions abundance measurements and reducing the number of missing values.
- A datastore making input raw data, intermediate and final results available for user
- A user friendly graphical interface facilitating navigation, examination and understanding of results

The accuracy and variance of the peptide measurements has been evaluated on a controlled data set in which relative abundances of proteins are known. Metrics such as peptide abundance variance and specificity and sensitivity of differential analysis have been calculated. Finally, the software allows quantification from identifications coming from various identification engines (Mascot, X! Tandem and OMSSA).

Acknowledgments

This work was partially funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir" call)

31. Mass-Up and Decision Peptide-Driven: two open-source applications for MALDI-TOF MS data analysis and protein quantification

Hugo López-Fernández, Daniel González-Peña, Miguel Reboiro-Jato, Florentino Fdez-Riverola, José L. Capelo-Martínez and Hugo M. Santos

Introduction: Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers (MALDI-TOF MS) has become popular during the last years thanks to its high speed and sensitivity for detecting peptides and proteins, allowing large groups of samples to be rapidly analyzed in a row. In this scenario, MALDI-TOF MS based proteomics is used to perform protein quantification as well as sample classification or clustering based in the presence or absence of peptides and proteins. Effective bioinformatics tools are key to perform these kind of analyses, since they are able to process the large amounts of raw data generated and apply the necessary algorithms to extract new knowledge and useful conclusions from a biological point of view.

The purpose of this work is to present to open-source, free-to-use desktop applications for the analysis of MALDI-TOF-MS data and protein quantification using 18O inverse labeling: Mass-Up (http://sing.ei.uvigo.es/mup/) and Decision Peptide-Driven (http://sing.ei.uvigo.es/DPD/), respectively.

Results: The identification and quantification of the protein contents of biological samples plays a crucial role in biological and biomedical research. The Decision Peptide-Driven (DPD) [1] tool implements an application for protein quantification based on an 18O direct and inverse labeling protocol followed with MALDI-TOF-MS analysis. DPD compares the results of the direct and inverse experiments in order to identify reproducible peptides (i.e. those that have similar direct and inverse ratios) that can be used

to subsequent and accurate protein absolute quantification. Then, it calculates the O16/O18 ratio from a mixture of unknown amount of protein and an internal standard labeled with O18 in order to consequently determine the unknown amount of protein.

In addition to protein quantification, there is also interest in the ability of differentiate samples belonging to different conditions (i.e. phenotypes, strains or diseases) by analyzing MALDI-TOF MS spectra. Mass-Up [2] is desktop multiplatform application specifically created to perform these types of analyses using MALDI-TOF MS data. It supports full raw data preprocessing using state-of-the-art libraries (e.g. MALDIquant) as well as different analysis including (i) quality control, (ii) biomarker discovery, (iii) clustering, (iv) three-dimensional PCA visualization and (v) classification.

Conclusions: Mass-Up and DPD are completely free, distributed under license GPLv3, and provide a friendly graphical user interface designed to avoid the need for advanced informatics skills to use them. Consequently, biologists and chemists can use them to analyze their experimental data without the intervention of a bioinformatics expert.

References:

[1] H.M. Santos; M. Reboiro-Jato; D. Glez-Peña; M.S. Diniz; F. Fdez-Riverola; R. Carvalho; C. Lodeiro; J.L. Capelo (2010) Decision Peptide-Driven: a free software tool for accurate protein quantification using gel electrophoresis and matrix assisted laser desorption ionization time of flight mass spectrometry. Talanta. 82(4):1412-1420. ISSN: 0039-9140. DOI: 10.1016/j.talanta.2010.07.007

[2] H. López-Fdez; H.M. Santos; J.L. Capelo; F. Fdez-Riverola; D. Glez-Peña; M. Reboiro-Jato (2015) Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics. 16:318. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0752-4

32. Analysis of expression profiles and regulatory data integration in the context of radiation research

Sherry Freiesleben, Markus Wolfien, Sonja Strunz, George Iliakis, Ralf Kriehuber and Olaf Wolkenhauer

Ionizing radiation (IR) is a known carcinogen and biological tissues exposed to IR undergo a series of highly orchestrated events leading to tissue damage or even tissue death. It is however unclear how low and high radiation doses differentially affect molecular pathways. The purpose of this study is to identify gene signatures associated to low or high radiation doses as well as to expose regulatory mechanisms influenced by low or high radiation doses. With the increased use of high-throughput technologies and bioinformatics approaches, a trend towards creating biochemical networks by integrating different "omic" technologies has emerged. We are currently analyzing and integrating microarray expression data of blood-derived tissues exposed to IR as well as other publically available data in order to gain a deeper understanding of the effects of IR on biochemical networks. It is anticipated that this study will elucidate differences regarding mechanisms of action behind low and high levels of IR and that it will further demonstrate the need of multi-omics bioinformatics analyses.

33. Exosome proteomics reveal a malignant memory profile in patients with urothelial bladder cancer

Michael Mints, Stefanie Hiltbrunner, Maria Eldh, Robert Rosenblatt, Benny Holmström, Farhood Alamdari, Markus Johansson, Johan Hansson, Jonas Vasko, Ola Winqvist, Amir Sherif and Susanne Gabrielsson

Introduction: Urothelial bladder cancer (UBC) patients have a poor prognosis, with approximately 50% 5-year survival after radical cystectomy (RC). Circa 20% of lymph-node negative patients relapse after RC, and this group benefits from neoadjuvant chemotherapy (NAC). These facts strongly suggest early micro-metastatic dissemination in UBC. Exosomes, nano-sized membrane vesicles that serve as means of communication between cells, are produced by both cancer and immune cells, and are believed to impact carcinogenesis and metastasis. They are found in all bodily fluids and a promising, non-invasive source of biomarkers. We characterised the exosomal proteome in bladder tissue and urine from UBC patients in order to understand the role of exosomes in bladder carcinogenesis and to find therapeutic and diagnostic biomarkers.

Materials and Methods: Exosomes were isolated by 2h ultracentrifugation from the supernatant of tissue explants and urine from bladders and ureters, all removed at cystectomy from nine UBC patients

who underwent transurethral resection (TUR-B) followed by NAC and RC. FACS was used to verify that the separated vesicles were in fact exosomes. The proteome was analysed by LC-MS/MS, and associations of protein expression with clinical parameters and functional signalling pathways were studied using principal component analysis (PCA) and network set enrichment analysis (NSEA).

Results: In total, 1443 proteins were identified with at least two peptides – 741 in urine and 302 in tissue with 400 found in both sets. On PCA, proteins from tissue exosomes could be separated based on whether they came from the tumour site or normal tissue, while urinary exosomal proteins clustered based on whether they came from bladder urine – in contact with the tumour site – or ureter urine, which had no tumour contact. The signalling pathways associated with bladder urine represented platelet activation, glycolysis, growth and immune signalling, while complement activation was enriched in exosomes from the tumour site. Two proteins were significantly up-regulated in tumour site tissue, while 17 were up-regulated and one down-regulated in bladder urine compared to control exosomes.

Conclusions: We show that despite macroscopic ablation through TUR-B and NAC, exosomes from the tumour site or from urine in contact with the tumour site are altered compared to control tissue and urine. The enriched pathways in bladder urine suggest constant secretion of metastasis-driving exosomes from the bladder even after tumour resection. This can be described as malignant tissue memory, and means that the time to radical surgery must be minimised. The proteins up-regulated in bladder urine will be explored as metastasis biomarkers and targets for immunotherapy.

35. Analyzing metaproteome samples on the go: the full-featured MPA portable software provides protein identification enriched with taxonomic and functional meta-information.

Thilo Muth, Robert Heyer, Dirk Benndorf, Erdmann Rapp, Udo Reichl, Lennart Martens and Bernhard Renard

The key role of microbial consortia has recently gained increased attention due to promising findings on their functional repertoire in the human intestinal tract. Complex microbial communities fulfill essential host-related functions regarding nutrient uptake, digestion and immune response. Metaproteomics, the mass spectrometry-based analysis of multi-species proteins from microbial samples, aims to elucidate the functional potential and taxonomic origin of such microbial consortia. This proteomic technique is also employed for rapidly detecting pathogens and studying their host-adaptation mechanisms. While throughput and resolution of the instrumentation have evolved enormously in the last decade, the analysis and interpretation of upcoming data still remains challenging – mainly due to lacking adapted solutions from the field of proteome bioinformatics. In addition, many laboratories have limited access to extensive computational infrastructure and expertise.

To overcome these shortcomings, we here present the freely available MetaProteomeAnalyzer (MPA) Portable software. MPA Portable is a light-weight and full-featured application which can be used for processing and analyzing metaproteomics data. In contrast to the original server-based MPA software [1], the presented tool requires no further installation steps and is independent of any SQL database system. Besides previously included algorithms, the novel development now also supports the state-of-the-art database search engine MS-GF+ [2]. While the results of multiple search engines can still be combined to increase the overall protein identification yield, an additional two-step search workflow [3] is implemented as optional feature. The objective is to increase the sensitivity for providing sufficient analysis resolution in further post-processing steps, such as protein grouping as well as automated sequence annotation at the taxonomic and functional level.

We tested our proposed software package on various data sets from single species and microbial community samples. The results show that both the use of modern search algorithms and two-step searching significantly increased the overall identification yield. Accounting for protein databases of different sizes, we found that an increased taxonomic coverage and depth can be achieved when excluding irrelevant sequences prior to the search process. Further, benchmarking the performance of the software with a sample of known species composition [4] indicated two important findings: first, using ground truth data, the proposed software reliably assigned peptides to their taxonomic origin, and second, the proportion of incorrect species assignments could be minimized when applying a taxon-specific significance threshold.

Eventually, the availability of MPA Portable should compensate for the rising interest in the research field of microbial community proteomics and its strong demand on user-friendly, yet robust solutions for

comprehensive data analysis.

References

[1] Muth et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res, 2015, 14(3):1557-65.

[2] Kim and Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Communications, 2014, 5:5277.

[3] Jagtap et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics, 2013, 13(8):1352-7.

[4] Tanca et al. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. PLoS One, 2013, 8(12):e82981.

36. The spectra-cluster toolsuite: Enhancing proteomics analysis through spectrum clustering

Johannes Griss and Juan Antonio Vizcaíno

Shotgun proteomics experiments generate continuously growing amounts of data with considerable redundancies between experiments. In the past, several research groups used spectrum clustering to merge highly similar spectra before identifying them to reduce the computational effort of identifying the same peptide multiple times. Additionally, we and other research groups showed that spectrum clustering can identify unidentified spectra by grouping them with identified ones, sometimes even from unrelated experiments. Finally, with the new PRIDE Cluster resource we were able to recognize millions of consistently unidentified spectra in the PRIDE Archive repository for MS/MS based proteomics data. Even though we and others showed that spectrum clustering can enhance proteomics data analysis, we are unaware of any software able to analyse spectrum clustering results and support these approaches.

Here we present our newly released spectra-cluster toolsuite (https://spectra-cluster.github.io) which is built around our spectra-cluster algorithm. Its final goal is to make spectrum clustering based data analysis available to any researcher without requiring in-depth bioinformatics knowledge.

Most importantly, the toolsuite provides graphical user interfaces to the spectra-cluster algorithm itself and contains detailed documentation on how to prepare data for clustering. Next, it contains a set of tools that take advantage of the fact that clustering does not require any identification data. One of these approaches is focused on quality control were mass spectrometry runs or experiments can be compared using the generated clusters as features for a principal component analysis. Thereby, outliers runs can quickly be identified.

If identification data is already present, the clustering results can be used to identify unidentified spectra if these were clustered with identified ones. Even though these inferred identifications do not increase the number of identified peptides or proteins, it can considerably improve the quality of label-free quantitation results. Therefore, the toolsuite also contains ready-to-use R scripts to perform label-free quantitation supporting a wide range of spectral-counting based methods. Additionally, we provide a consensus spectrum exporter that can be used to export, for example, the consensus spectra of all unidentified cluster with a certain size. These can then be targeted using computationally more expensive methods.

To highlight how spectrum clustering in general and the spectra-cluster toolchain specifically can improve proteomics research we present a dataset on tumour heterogeneity in human melanoma. There, spectrum clustering not only considerably improved label-free quantitation results but also helped us to identify hundreds of unidentified spectra to target for further analysis. This example highlights that spectrum clustering and the spectra-cluster toolchain can efficiently enhance proteomics data analysis.

37. An open-source software for fast and automatic processing of 2D-gel and MALDI-based mass spectrometry protein data

José Araújo, Hugo López-Fernández, Daniel Glez-Peña, Miguel Reboiro-Jato, Florentino Fdez-Riverola and José Capelo-Martínez

Introduction: 2D-gel electrophoresis and MALDI-TOF-MS (mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers) are usually employed together in experimental workflows. While 2D-gel electrophoresis allows obtaining a set of differentially expressed spots,

MALDI-TOF-MS allows identifying the proteins associated with such spots. Such processes require a lot of data processing steps in order: (i) to analyze 2D-gels across samples to obtain the differentially expressed spots using Progenesis SameSpots software, (ii) to excise such spots and to treat them for protein identification and (III) to bind those protein identifications to the 2D-gel spots, among other tasks. Such procedure is usually performed in an intensive handling and time consuming way. Moreover, doing this repetitive process in a non-automated manner is error-prone, threatening study reliability and reproducibility.

In order to fast and automatically perform these tasks we are developing the S2P software application (http://sing.ei.uvigo.es/s2p/), to help researchers to overcome this tedious but necessary data processing. The purpose of this work is to present this open-source software.

Results: S2P (Spot to Proteomics) is an AIBench based desktop multiplatform application specifically created to perform fast processing of 2D-gel and MALDI-mass spectrometry protein identification-based data. Currently, S2P allows researchers: (i) to collect Progenesis SameSpots reports into a full table where all samples can be compared and analyzed, (ii) to link spots to protein identifications obtained with Mascot, and (iii) to view and export data in different ways (such as heatmaps or Venn Diagrams).

For instance, processing data from an experiment involving fifteen samples takes more than two weeks when done manually by laboratory staff without S2P; now, using S2P the same data processing plus additional data analysis features can be done in less than five minutes.

Conclusions: S2P (http://sing.ei.uvigo.es/s2p/) is completely free, distributed under license GPLv3, and provide a friendly graphical user interface designed to allow researchers saving time in data processing tasks related to 2D and MALDI mass spectrometry protein identification-based data.

38. The study of the proteins of cestoda Triaenophorus sp. at different stages of the parasite life cycle.

Albina Kochneva, Ekaterina Borvinskaya, Darya Bedulina, Anton Gurkov and Boris Baduev

We compare the soluble proteins of parasitic tapeworms Triaenophorus nodulosus (Cestoda) at different stages of their life cycle. Members of Cestoda family are unique organisms, which are obligate parasites with a complex life cycle that includes the life stages at very different habitats. Thus, cestodes can be considered as an appropriate model for the study of molecular adaptations to the changes of environmental conditions. Complex life cycle of parasite includes: first intermediate host of T.nodulosus is planktonic crustaceans (Copepoda), then a secondary host is different freshwater fish (target organ is liver), and finally, the definitive host – pike (target organ is intestine). The proteins of T.nodulosus collected from the second and final hosts were separated by a method of single- and two-dimensional electrophoresis. By the program of image processing (a custom-made addon for ImageJ Software), we have revealed 22 proteins spots with different optical densities on gel images ($p < 0.05$) in worms at different developmental stages. In addition, a comparison was made for the parasites at one developmental stage (called - plerocercoid) obtained from different host species (perca Perca fluviatilis, burbot Lota lota, ruffe Gymnocephalus cernuus). We found significant differences in the protein concentration for three spots. Thus, it was revealed the target proteins, which might be involved in the molecular mechanisms of adaptation of parasite (including different developmental stages) in relation to specific "environment" of hosts. Further identification of peptide mass mapping by MALDI-TOF (under analysis) will determine the role of these proteins in the metabolism of parasite as well as in the parasite-host relationships. This research supported by the State budgetary theme No0221-2014-0003 and Russian Foundation for Basic Research No16-34-50090

39. Variance-sensitive clustering: a shortcut to data interpretation

Veit Schwämmle and Ole N. Jensen

Grouping of similar expression profiles facilitates the recognition of the underlying patterns in multi-dimensional data sets. This clustering process becomes highly relevant when searching for common expression profiles in large-scale experiments with thousands of measurements at different conditions. Optimal clustering results yield groups of co-regulated features, thereby revealing detailed information about the relevant involved biological processes in the system. Modern experimental designs include replicated measurements of each feature within the same condition to statistically assess their differences.

Currently, the noisy character of the experimental data, being different for each feature, impedes optimal cluster procedures when neglecting feature-specific variances leading to a large number of false positives in the cluster groups, ergo preventing confident in-depth biological interpretation using the full information content of the data sets. We present a novel tool that accounts for feature-specific variance based on a algorithm we derived from the fuzzy c-means clustering method. The software unifies statistical testing with pattern recognition, and provides a simple and straightforward method to obtain structural feature groups for subsequent data interpretation. Thorough performance evaluation using artificial and experimental data sets shows i) better recognition of underlying patterns and better filtering of incorrectly assigned cluster members; ii) prior data filtering leads to wrong cluster number estimations in noisy data sets; iii) higher enrichment of features with common biological function in clusters for transcriptomics and proteomics data sets; iv) mapping of low-level data (probes, peptides) to main features (transcripts, proteins) can be avoided by direct application the clustering to low-level data. Hence, by considering replicated measurements of the same feature, our tool provides a short-cut for in-depth insights into the data structure without relying on mostly arbitrary prior filtering and averaging. Access VSClust at computproteomics.bmb.sdu.dk/Apps/VSClust.