Multi-level proteome characterization using high resolution mass spectrometry

Author(s):
Rosenberg, George A.

Publication Date:
2016

Permanent Link:

https://doi.org/10.3929/ethz-a-010742448 →

DISS. ETH NO. 23616

Multi-level proteome characterization using high resolution mass spectrometry

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

George A. Rosenberger

MSc, ETH Zurich
born on July 24, 1986
citizen of Dübendorf ZH, Switzerland

accepted on the recommendation of

Prof. Dr. Ruedi Aebersold
Dr. Lars Malmström
Prof. Dr. Nenad Ban
Prof. Dr. Wolf-Dietrich Hardt

2016

Summary

Systems biology represents a new paradigm to elucidate biological mechanisms, pathways or phenotypes. Instead of focusing on the characterization of individual, isolated units, the systems approach acknowledges the complexity of biology in a holistic manner. Proteins are major functional components of molecular systems and because they are involved in almost all biological processes, detailed understanding is a requirement to enable systems biology. Proteomics is the discipline aiming for the characterization of related sets of proteins of an organism or system with mass spectrometry being currently the main method for identifying and quantifying proteins at large scale. In the most prominent implementation, bottom-up proteomics, proteomes of interest are enzymatically digested into peptides and separated using liquid chromatography (LC) followed by analysis using tandem mass spectrometry (MS/MS). Several technological advances over the past decades have improved the accessibility of proteins for mass spectrometry, increased the throughput both in terms of number of analytes as well as samples and have introduced new experimental strategies to investigate specific aspects of a proteome. Most recently, data-independent acquisition (DIA) increased the continuity and density of mass spectrometry data for bottom-up proteomics, enabling measurement of a large fraction of the peptides of a sample to acquire complex, convoluted data sets. This increased depth of coverage can be beneficial in many contexts, particularly for the consistent quantification over many samples.

This thesis describes the development of algorithms and resources to improve DIA-based peptide-, protein- and interaction-level analysis and demonstrates how the application of these methods to human and microbial samples enables proteome characterization from different perspectives.

To detect and quantify peptides measured by LC-MS/MS employing DIA, we developed OpenSWATH. Our algorithm provides a complete workflow for the fully automated analysis of DIA data using the SWATH-MS targeted data extraction approach. By validation using a synthetic gold standard data set, we demonstrate that OpenSWATH achieves high quantitative accuracy and sensitivity, even for complex proteome samples. The application to a study of *S. pyogenes* differential protein expression shows that our workflow enables reproducible and accurate quantification of 70% of the expressed proteome in a single LC-MS/MS measurement.

Based on this foundation and to address the challenge of detection and quantification of peptides carrying post-translational modifications (PTMs), we developed IPF (Infer-

ence of PeptidoForms). This algorithm integrates experimentally observed evidence on different levels of LC-MS/MS data to assign the most likely peptidoform to candidate peptide signals and to control the error-rate by an integrative approach. We demonstrate the performance of our algorithm by the application to a human twin study of modified blood plasma proteins. IPF provided accurate and reproducible quantification of modified peptides or peptidoforms across more than 200 complex samples. This enabled us to study the heritable and environmental effects on modified plasma proteins for the first time.

To improve inference and quantification on the protein-level, we provide a combined assay library as a resource to mediate comparability between heterogeneous human samples and to increase the number of targeted analytes for more than 10,000 human proteins. The assay library has been applied to a wide range of research and clinical studies. However, the large number of queries that it enables on single measurements or large-scale data sets raises new statistical challenges. For this reason, we propose statistical metrics that enable experimental context-specific analyses that are optimally suited to particular research hypotheses.

Bottom-up proteomics quantifies peptides, but for many biological research questions inferred protein abundances could provide additional useful information. To address this need, we developed aLFQ (absolute Label-Free Quantification), a computational toolkit providing algorithms for quantitative protein inference, absolute label-free abundance estimation and quantification error estimation.

Size-exclusion chromatography (SEC) in combination with bottom-up proteomics is an emerging method for the characterization of protein-protein interactions or complexes. However, the employed data analysis strategies do not yet make specific use of the improved quantitative performance provided by DIA methods like SWATH-MS. With SECAT (Size-Exclusion Chromatography Algorithmic Toolkit), we propose a method to integrate prior knowledge of protein-protein interactions from databases and to query SEC-SWATH-MS data sets *in silico* in an integrated statistical framework, analogously to affinity purification mass spectrometry (AP-MS) pull-down experiments. Based on such pairwise interactions, a network-centric representation can be produced.

In conclusion, this thesis presents a computational framework to make use of high resolution DIA data for advancing the analysis of peptides, proteins and interactions. In combination, these methods enable the characterization of proteomes from different perspectives. We and others demonstrated the usefulness, scalability and performance of the algorithms by a wide range of research applications.

Zusammenfassung

Die Systembiologie repräsentiert ein neues Paradigma um biologische Mechanismen, molekulare Signalwege oder Phänotypen aufzuklären. Anstatt auf die Charakterisierung von individuellen, isolierten Einheiten zu fokussieren, berücksichtigt der systemische Ansatz die Komplexität der Biologie auf eine holistische Art und Weise. Proteine sind wichtige funktionale Komponenten in molekularen Systemen und weil sie in fast allen biologischen Prozessen involviert sind, ist ihr detailliertes Verständnis eine Voraussetzung um die Systembiologie zu ermöglichen. Die Proteomik verfolgt als Disziplin das Ziel, definierte Gruppen von Proteinen eines Organismus oder Systems zu charakterisieren. Die Massenspektrometrie ist die zurzeit hauptsächlich angewandte Methode für das Identifizieren und Quantifizieren von Proteinen mit hohem Durchsatz. In der prominentesten Implementierung, der „Bottom-Up“ Proteomik, werden Proteome enzymatisch in Peptide verdaut und mithilfe von Flüssigkeitschromatographie (liquid chromatography, LC) separiert, gefolgt von der Analyse mittels Tandem-Massenspektrometrie (MS/MS). Mehrere technologische Fortschritte haben über die vergangenen Jahrzehnte die Zugänglichkeit der Proteine für die Massenspektrometrie verbessert und den Durchsatz sowohl in der Anzahl der Analyten als auch der Proben verbessert. Zusätzlich wurden neue experimentelle Strategien entwickelt um spezifische Aspekte eines Proteoms zu untersuchen. Seit kurzem ermöglicht eine neue Methode der datenunabhängigen Messung (data-independent acquisition, DIA) verbesserte Kontinuität von massenspektrometrischen Daten für die „Bottom-Up“ Proteomik. Dies ermöglicht die Messung einer grösseren Anzahl von Peptiden in einer Probe und generiert komplexere Datensätze. Diese verbesserte Abdeckung kann für viele Anwendungen nützlich sein, vor allem für die kontinuierliche Quantifizierung von Peptiden in vielen Proben.

Diese Dissertation beschreibt die Entwicklung von Algorithmen und Ressourcen, um die DIA-basierte Analyse auf der Stufe von Peptiden, Proteinen und Interaktionen zu verbessern und sie demonstriert, wie die Anwendung dieser Methoden auf humane und mikrobielle Proben die Charakterisierung von Proteomen aus verschiedenen Perspektiven ermöglicht.

Wir haben OpenSWATH entwickelt, um Peptide zu detektieren und quantifizieren, welche mittels DIA-basierter LC-MS/MS gemessen wurden. Unser Algorithmus ermöglicht einen kompletten Workflow für die vollautomatisierte Analyse von DIA-Daten mittels dem SWATH-MS Ansatz für die gezielte Extraktion von Daten. Durch eine Validierung basierend auf einem synthetischen Gold-Standard-Datensatz demonstrieren wir, dass OpenSWATH hohe quantitative Genauigkeit und Sensitivität erzielt, sogar für komplexe

Proben von Proteomen. Die Anwendung auf eine Studie der differentiellen Proteinexpression in *S. pyogenes* zeigt, dass unser Workflow die reproduzierbare und genaue Quantifizierung für 70% des exprimierten Proteoms in einer einzelnen LC-MS/MS Messung ermöglicht.

Basierend auf dieser Grundlage und um die Detektion und Quantifizierung von Peptiden mit posttranslationalen Modifikationen (PTM) zu verbessern, haben wir IPF (Inference of PeptidoForms) entwickelt. Dieser Algorithmus integriert experimentell beobachtete Evidenz auf verschiedenen Ebenen von LC-MS/MS Daten, um die wahrscheinlichste Peptidoform möglichen Peptidsignalen zuzuordnen und die Fehlerrate über einen integrativen Ansatz zu regulieren. Wir demonstrieren die Leistungsfähigkeit unseres Algorithmus durch die Anwendung auf eine humane Zwillingsstudie von modifizierten Blutplasmaproteinen. IPF ermöglicht die genaue und reproduzierbare Quantifizierung von Peptiden in mehr als 200 komplexen Proben. Dies ermöglichte es uns zum ersten Mal, die vererbbaren und umweltassoziierten Einflüsse auf modifizierte Plasmaproteine zu studieren.

Um die Inferenz und Quantifizierung auf der Ebene der Proteine zu verbessern, stellen wir eine kombinierte Assay-Sammlung als Ressource zur Verfügung, welche die Vergleichbarkeit von heterogenen Proben ermöglicht und die mögliche Anzahl der abgefragten Analyten für mehr als 10'000 humane Proteine vergrössert. Die Assay-Sammlung wurde bereits auf ein breites Spektrum von Projekten in der biologischen Forschung angewandt. Dabei stellt die grosse Anzahl von Abfragen von einzelnen oder mehreren Messungen, die dadurch ermöglicht werden, neue statistische Herausforderungen. Aus diesem Grund schlagen wir statistische Metriken vor, welche eine experiment- und kontextspezifische Analyse ermöglichen, die optimal auf eine spezifische Hypothese zugeschnitten ist.

Die „Bottom-Up" Proteomik quantifiziert Peptide, wobei abgeleitete Proteinmengen für viele biologische Fragestellungen zusätzliche nützliche Informationen liefern könnten. Um diese Nachfrage zu bedienen, haben wir aLFQ (absolute Label-Free Quantification) entwickelt; ein Toolkit, welches Algorithmen für die quantitative Proteininferenz, die absolute Bestimmung von Proteinmengen und die Schätzung des Quantifizierungsfehlers bereitstellt.

Grössenausschlusschromatographie (size-exclusion chromatography, SEC) ist in Kombination mit der „Bottom-Up" Proteomik eine aufstrebende Methode für die Charakterisierung von Protein-Protein-Interaktionen oder Komplexen. Die zurzeit verwendeten Strategien zur Datenanalyse nutzen jedoch die verbesserte quantitative Leistungsfähigkeit, ermöglicht durch DIA Methoden wie zum Beispiel SWATH-MS, noch nicht vollständig. Mit SECAT (Size-Exclusion Chromatography Algorithmic Toolkit) schlagen wir eine Methode zur Integration des gesammelten Wissens über Protein-Protein Interaktionen aus Datenbanken zur Abfrage von SEC-SWATH-MS Datensätzen vor. Analog zur Affinitätsreinigung kombiniert mit Massenspektrometrie (affinity purification mass spectrometry, AP-MS) können die Daten als *in silico* „Pull-Down"-Experiment in einem integrierten statistischen Framework interpretiert werden. Basierend auf solchen paarweisen Interaktionen erzeugt diese Methode eine netzwerkzentrische Repräsentation.

Zusammenfassend präsentiert diese Dissertation ein rechnergestütztes Framework, welches hochauflösende DIA-Daten verwendet, um Fortschritte für die Analyse von Peptiden, Proteinen und Interaktionen zu erzielen. In Kombination ermöglichen diese Methoden die Charakterisierung von Proteomen aus verschiedenen Perspektiven. Wir und andere Gruppen haben die Nützlichkeit, Skalierbarkeit und Leistungsfähigkeit unserer Algorithmen für ein breites Spektrum von Fragestellungen demonstriert.