ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DISS.ETH NO. 26968

High-resolution protein correlation profiling to resolve subcellular proteome organization

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZÜRICH
(Dr. sc. ETH Zürich)

presented by

Fabian Frommelt

Dipl.-Ing. in Biotechnologie, Universität für Bodenkultur, Wien
born on 12.11.1989
citizen of Nenzing, Austria

accepted on the recommendation of

Prof. Dr. Ruedi Aebersold
Dr. Matthias Gstaiger
Prof. Dr. Paola Picotti
Prof. Ph.D. Anne-Claude Gingras

October 2020

Summary

Systems biology is a holistic approach which studies the relations of the different layers of networks across all scales within any given biological system. Instead of following a reductionist approach by taking layers apart and study each single biomolecule, a system biology approach tries to identify the relationship existing within a system and to integrate these information to generate system-level blueprints of molecular mechanisms. These approaches are especially useful to understand how biomolecules organize in greatly specialized hierarchical networks, or pathways. This is achieved by perturbation of the cell in combination with high-resolution and high-throughput measurements. Thereby spatial and temporal insights into the architecture of networks and layers of networks (or "networks of networks") are gained and new hypothesis can be formulated. System biology thereby helps to link genetic perturbations back to its phenotypic manifestation. One of the most important biomolecule of this networks of networks in a cell are proteins, as they are involved in almost all cellular processes. Proteomics is the omics methods which studies proteins, whereas mass spectrometry (MS) based proteomics methods are the most widely applied techniques to identify and quantify proteins. Proteomics allows further to study the relation and organization of proteins into interaction networks. The interactome encompass all protein-protein interaction within a cell, and can be probed by affinity purification mass spectrometry (AP-MS). Higher order organization, such as protein complexes, can be investigated by combinations of biochemical fractionation with quantitative MS. While proteins interact at the molecular scale (i.e interactions), the cellular scale (i.e organelles) greatly contributes to cellular diversity and mediates homeostasis by specialization of organelles into cellular functions. By employing subcellular fractionation techniques combined with MS it is possible to shed light on the spatial distribution of the proteins across these organelles. These systems approaches to study the organization of the proteome were enabled by rapid developments in instrumentation and data analysis strategies. One of the developed techniques, data independent acquisition (DIA) was especially useful for systems approaches, as it enables increased continuity and completeness across large datasets. As protein correlation profiling (PCP) as tool to elucidate protein organization across cellular states, is becoming more widely used, developments focusing on fractionation techniques, throughput in sample preparation, and robust and high-throughput data acquisition schemes are needed. Increased throughput however results in highly-complex datasets, and thus the current data analysis strategies need to be improved.

This thesis describes biochemical and computational advances for the analysis of protein complexes from cellular extracts and subcellular enriched fractions by applying native fractionation techniques combined with high-resolution "bottom-up" mass spectrometry based proteomics.

To study protein-protein interactions and protein complexes on a system-wide level, native biochemical fractionation methods were established. These methods record the quantitative elution profile of each protein across a gradient, and by subsequent correlation based analysis, allow to identify protein complexes. The established methods are mostly based on combinations of size exclusion chromatography (SEC) with quantitative mass spectrometry approaches. In the first part of the thesis we present the ongoing efforts to increase throughput, sample quality, and yield within large-scale protein complex profiling experiments. We summarize the developed biochemical and mass spectrometer acquisition techniques which enables the study of hundreds of complexes across several cell-states and replicates at a throughput not yet reported in the field. These methodological developments are the basis for other projects presented in this thesis.

The subcellular organization of the human proteome is of great importance for cellular processes. However, despite many studies investigating the subcellular location of proteins, there is a clear lack of knowledge how the protein is organized within organelles. We developed and integrated workflow to study protein localization and protein associations in the form of protein complexes on a system-level, by investigating 12 organelles. We employed high-resolution separation techniques combined with robust data acquisition by SWATH-MS, which enabled together with a novel machine learning approach, the mapping of >4500 proteins to specific subcellular location. SEC-SWATH allowed us further to investigate the assembly state within the organelles. Together, this enabled us to outline a map of the organization of the human proteome on subcellular level.

Next, we implemented a method to experimentally identify all co-purified protein complexes within an AP-MS experiment. A single AP-MS yields binary protein-protein interactions (PPIs), and often multiple AP-MS experiments are combined to generate high-density interaction networks from these binary interaction networks. These networks are then used to derive the intrinsic modularity in the functional assemblies. We set out to directly probe all co-purified bait containing protein complexes of specific affinity purified samples. We combined an affinity purification with subsequent separation by blue native PAGE, a well known powerful tool for separating proteins and multimeric assemblies. To limit the samples losses and increase robustness, we developed a high-throughput sample preparation protocol, with similar or better yields than established methods. The data acquisition time was reduced, by employing short gradients, allowing us to measure a gel within one day. To ensure high-quality of quantitation, despite the 21 minutes gradient, we optimized the MS acquisition and developed a window scheme which allows to employ close to optimum duty cycle and fill time. We further introduced a novel data analysis strategy for protein-correlation data, which allowed us to retrieve not only the modularity, but also the PPIs and filter the interactions without the need for additional controls. The workflow is introduced on the example of the Prefoldin (PFD) and the R2TP/Prefoldin-like (R2TP/PFDL) protein complexes. We were able to resolve the complexes and including sub-modules. We further identified a yet not reported canonical PFD assemblies. The data allowed us to identify clients of the PFD and PFDL protein complex.

In collaborative efforts we employed AP-MS to study mutant transcription factor interactions in cancer disease models, thereby identifying potential therapeutic targets. Further, we elucidated with a multi-layered proteomics and phosphoproteomics approach in a cellular model potential effects of cancer mutations on the associations and assembly state of a kinase complex. In two further projects we studied the effects of perturbations on the assembly state by employing global protein complex analysis across cellular states. These studies showed, that SEC-SWATH enabled the study of proteome reorganization in mitotic cells and to reveal the organization in the proteome upon dosage compensation in mouse embryonic stem cells.

In conclusion, this thesis presents biochemical and computational methods to analyse complexes on subcellular level. This was achieved by combining fractionation techniques with DIA mass spectrometry. These approaches were employed to draw a spatially resolved map of the human protein complex landscape and thereby enabling new insights into the functional organization of the proteome. Together in collaborations, we demonstrate the applicability of interaction proteomics, protein complex profiling for differential assembly states analysis and the usefulness of combining affinity purification with protein co-elution profiling.

Zusammenfassung

Die Systembiologie verwendet einen allumfassenden Ansatz, um die Zusammenhänge der verschiedenen Netzwerke und ihrer molekularen Bestandteile über alle hierarchischen Ebenen eines bestimmten biologischen Systems zu untersuchen. Anstatt dem traditionellen reduktionistischen Ansatz zu folgen, bei welchem Netzwerke oder Reaktionswege in einzelne molekularen Komponenten zerlegt werden um diese dann einzeln für sich betrachtet einer genauen Analyse zu unterziehen um die biochemische und molekulare Funktionsweise zu verstehen, versucht ein systembiologischer Ansatz, die in einem System vorhandenen wechselseitigen Abhängigkeiten zu identifizieren, und diese Information in Blaupausen molekularer Reaktionsmechanismen auf Systemebene zu integrieren. Diese Vorgehensweise ist besonders dann nützlich, wenn es darum geht zu verstehen wie sich Biomoleküle in hochspezialisierten hierarchischen Netzwerken und Reaktionswegen organisieren. Eine der nützlichsten Methoden um die Abhängigkeiten der Biomoleküle in einem Netzwerk zu entschlüsseln ist es, die Komponenten des Systems im Gleichgewichtszustand und in einem abgeänderten, von aussen beeinflussten Zustand, mittels hochauflösenden Messungen zu erfassen. Dadurch können räumliche und zeitliche Einblicke in die Netzwerkarchitektur und die inhärenten Abhängigkeiten zwischen Netzwerken (oder "Netzwerke von Netzwerken") gewonnen und neue biologische Hypothesen formuliert werden. Durch diese Verknüpfung von unterschiedlichen Netzwerken ermöglicht die Systembiologie eine genetische Mutation und die direkten phänotypischen Manifestationen auf molekularer Ebene in Relation zu stellen. Eines der wichtigsten Biomoleküle, welches diese zellulären Netzwerke schliesst sind Proteine, da sie an fast allen zellulären Prozessen beteiligt sind. Die Proteomik untersucht diese Proteine, und die auf Massenspektrometrie (MS) basierende Proteomik- Methode ist bei weitem die am Häufigsten angewandte Methode, um die Identifizierung und Quantifizierung von Proteinen in einer Messung zu ermöglichen. Weiteres erlaubt die Proteomik, die wechselseitigen Abhängigkeiten und die Organisation von Proteinen in Interaktionsnetzwerken zu untersuchen und darzustellen. Alle Protein-Protein-Wechselwirkungen (Protein-Protein Interactions, PPIs) innerhalb einer Zelle werden als Interaktom bezeichnet, welches durch die InteraktionsproteomikmethodeAffinitätsreinigungs-Massenspektrometrie (AP-MS) untersucht wrid. Organisationen höherer Ordnung, wie Proteinkomplexe, können durch Kombination von biochemischer Fraktionierung mit quantitativer MS untersucht werden. Während Proteine auf molekularer Ebene interagieren, tragen auf zelluläre Ebene die Organellen stark zur zellulären Spezialisierung bei und kontrollieren Homöostase durch Spezialisierung von Organellen auf zelluläre Funktionen. Durch den Einsatz subzellulärer Fraktionierungstechniken in Kombination mit MS ist es möglich, die räumliche Verteilung der Proteine über diese Organellen zu untersuchen. Diese Systemansätze zur Untersuchung der Organisation des Proteoms wurden durch Entwicklungen bei Instrumentierungs- und Datenanalysestrategien ermöglicht. Eine der entwickelten Techniken, die datenunabhängige Erfassung (Data Independent Acquisition, DIA), war besonders nützlich für Systemansätze, da sie eine erhöhte Kontinuität und Vollständigkeit über große Messreihen hinweg ermöglicht. Da das Proteinkorrelationsprofil (PCP) als Instrument zur Aufklärung der Proteinorganisation über Zellzustände hinweg immer häufiger eingesetzt wird, sind Entwicklungen erforderlich, die sich auf Fraktionierungstechniken, den Durchsatz bei der Probenvorbereitung sowie robuste Datenerfassungsschemata mit hohem Durchsatz konzentrieren. Ein erhöhter Durchsatz führt jedoch zu hochkomplexen Datensätzen, wodurch es auch einer Verbesserung der aktuell angewandten Datenanalysestrategien bedarf.

Diese Arbeit beschreibt biochemische und computergestützte Fortschritte bei der Analyse von Proteinkomplexen aus Zellextrakten und subzellulär angereicherten Fraktionen durch Anwendung nativer Fraktionierungstechniken in Kombination mit hochauflösender "Bottom-up"-MS-basierter Proteomik.

Um Protein-Protein-Wechselwirkungen und Proteinkomplexe auf systemweiter Ebene zu untersuchen, wurden native biochemische Fraktionierungsmethoden etabliert. Diese Methoden zeichnen das quantitative Elutionsprofil jedes Proteins über einen Gradienten auf und ermöglichen durch anschließende korrelationsbasierte Analyse die Identifizierung von Proteinkomplexen. Die etablierten Methoden basieren meist auf Kombinationen der Größenausschlusschromatographie (Size Exclusion Chromatography, SEC) mit quantitativer Proteomik. Im ersten Teil der Arbeit stellen wir die

laufenden Bemühungen zur Steigerung des Durchsatzes, der Probenqualität und der Ausbeute zur Erstellung von Proteinkomplexprofilen vor. Wir fassen die entwickelten biochemischen und MS-Methoden zusammen, welche die Untersuchung von Hunderten von Komplexen über unterschiedliche Zellzustände hinweg ermöglichen und bei einem Durchsatz replizieren, über den auf diesem Gebiet noch nicht berichtet wurde. Diese methodischen Entwicklungen bilden die Grundlage für andere in dieser Arbeit vorgestellten Projekte.

Die subzelluläre Organisation des menschlichen Proteoms ist für zelluläre Prozesse von großer Bedeutung. Trotz vieler Studien, die die subzelluläre Verteilung von Proteinen untersuchten, ist die Fragestellung, wie das Protein innerhalb von Organellen organisiert ist noch nicht schlüssig und allumfassend beantwortet. Wir haben ein Protokoll weiterentwickelt, um die Proteinlokalisierung und Proteinassoziationen in Form von Proteinkomplexen auf Systemebene für 12 Organellen zu untersuchen. Wir verwendeten hochauflösende Separationstechniken in Kombination mit einer robusten Datenerfassung durch SWATH-MS, die zusammen mit einem neuartigen Ansatz des maschinellen Lernens die Abbildung von 4500-Proteinen auf einem bestimmten subzellulären Ort ermöglichten. Mit SEC-SWATH konnten wir den Assemblierungszustand der Proteine innerhalb der Organellen weiter untersuchen. Zusammenfassend ermöglichte dieser Forschungsansatz uns, eine Karte der Organisation des menschlichen Proteoms auf subzellulärer Ebene zu skizzieren.

Als nächstes implementierten wir eine Methode, um alle mit aufgereinigten Proteinkomplexe innerhalb eines AP-MS-Experiments experimentell zu identifizieren. Ein AP-MS Experiment erlaubt es, binäre Protein-Protein-Wechselwirkungen zu untersuchen, dabei werden häufig mehrere AP-MS-Experimente kombiniert, um aus diesen multiplen binären Interaktionen, Interaktionsnetzwerke mit hoher Dichte zu erzeugen. Diese Netzwerke werden dann verwendet, um die intrinsische Modularität in der Struktur des Netzwerkes mittels Korrelationsanalyse abzuleiten. Um die Modularität des Netzwerkes in einem einzelnen Experiment zu untersuchen, kombinierten wir Affinitätsreinigung mit anschließender Auftrennung der aufgereinigten Proteinkomplexe mittels nativer Polyacrylamid Gelelektrophorese (blue native PAGE, BNP). Um dies zu ermöglichen, entwickelten wir ein Protokoll das die Probenausbeute und Robustheit erhöhte und gleichzeitig eine Probenvorbereitung mit erhöhtemDurchsatz erlaubt, im Vergleich zu ähnlichen etablierten Protokollen. Zusätzlich kombinierten wir diese Probenaufbereitung mit einer MS-Messmethode welche die Länge der chronographischen Gradienten stark verkürzte. Wir erreichten trotz des nun mehr nur 21-minütigen Gradienten eine hohe quantitative Proteomanalyse Qualität, da wir die DIA Messmethode optimierten. Ausserdem führten wir eine neuartige Datenanalysestrategie für Proteinkorrelationsdaten ein, die uns nicht nur erlaubte die Modularität aufzuschlüsseln, sondern auch die PPIs zu identifizieren ohne dass wir zusätzliche Kontrollen benötigten. Am Beispiel der Proteinkomplexe Prefoldin (PFD) und R2TP/Prefoldin-like (R2TP/PFDL) konnten wir zeigen, dass wir mit unserer Methode die Komplexe und die Modularität aufklären können, nicht beschriebene Komplexe identifizieren können und es die Daten ermöglichen Komplexe die mit PFD and PFDL interagieren zu identifizieren.

Wir verwendeten in Zusammenarbeit mit anderen Arbeitsgruppen AP-MS, um den Effekt von Chimären Fusionstranskriptionsfaktoren auf Proteininteraktionsnetzwerke in Krebszellmodellen zu untersuchen, und so potenzielle Ziele für eine medikamentenbasierende Therapie zu identifizieren. In einer weiteren Kollaboration haben wir einen Multiplexforschungsansatz, bestehend aus einer Kombination von globalen Proteomikmethoden, Interaktionsproteomik und der quantitieren Untersuchung von Phosphorylierungsmuster mittels Proteomik verwendet, um die Auswirkungen von bekannten genetischen Krebsmutationen auf die Assoziationen und den Assemblierungszustand eines Kinasekomplexes, in einem zellulären Model zu untersuchen. Für zwei andere Projekte verwendeten wir eine globale Proteinkomplexanalysemethode um die Auswirkungen auf den Assemblierungszustand, ausgelöst durch eine Perturbation zu untersuchen. Diese Studien zeigten, dass SEC-SWATH die differentielle Untersuchung der Proteomneuorganisation von Zellen in Mitose und Interphase ermöglicht und der Untersuchung ebendieser Umstrukturierungen in embryonalen Stammzellen von Mäusen nach der Abschaltung des geschlechtsspezifischen X-Chromosoms dient.

Zusammenfassend werden in dieser Dissertation biochemische und computergestützte Methoden zur

Analyse von Proteinkomplexen auf subzellulärer Ebene vorgestellt. Dies wurde durch Kombination von Fraktionierungstechniken mit der hochauflösenden massenspektrometrischen Messmethode DIA (Data indenpend acquistion) erreicht. Diese methodischen Ansätze wurden verwendet, um eine räumlich aufgelöste Karte der menschlichen Proteinkomplexlandschaft zu zeichnen und damit neue Einblicke in die funktionelle Organisation des Proteoms zu ermöglichen. In wissenschaftlichen Kollaborationen demonstrieren wir die Anwendbarkeit der Interaktionsproteomik, die Proteinkomplexprofilanalyse für die Untersuchung differenzieller Assemblierungszustände und die Vorteile der Kombination von Affinitätsaufreinigung mit Proteinkoeluationsanalyse.