DISS. ETH NO. 22446

Towards complete proteomics data matrices using targeted analysis of next-generation mass spectrometry data

A thesis submitted to attain the degree of

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

presented by

Hannes L. Röst

MSc, ETH Zurich
born on October 29, 1988
citizen of Uttwil, Switzerland

accepted on the recommendation of

Prof. Dr. Ruedi Aebersold
Dr. Lars Malmström
Prof. Dr. Paola Picotti
Prof. Dr. Niko Beerenwinkel

2014

Summary

Molecular systems biology is an emerging research field which aims to study complex biological processes by a comprehensive, system-wide approach. Technologies that allow quantitative measurements of proteins are central to systems biology since these biomolecules perform a wide range of essential functions in cells. Currently, one of the most powerful technologies to study proteins in high throughput is mass spectrometry-based proteomics, which can be applied in two different flavors. While untargeted shotgun proteomic approaches allow the identification of a large fraction of a proteome, they suffer from poor reproducibility across repeat measurements. On the other hand, targeted proteomic approaches such as selected reaction monitoring (SRM) have been established for the reproducible, sensitive and accurate protein quantification across many experimental conditions, but they have so far lacked the throughput needed for system-wide investigations.

This thesis presents several advances in computational mass spectrometry specifically addressing major challenges of targeted proteomics such as assay specificity and limited throughput. First, we present the development of the SRMCollider simulation software, where we implemented a conceptually novel method to predict the specificity of mass spectrometric assays in targeted proteomics. The software was subsequently used in several large-scale SRM studies to evaluate the specificity of proteome-wide assay collections for *Mycobacterium tuberculosis*, *Saccharomyces cerevisiae* as well as an assay library for over 1000 cancer-associated human proteins.

Next, we present contributions to the development of a novel high-throughput targeted proteomics technology (SWATH-MS) through (i) theoretical simulation using the SRMCollider and (ii) the OpenSWATH analysis software. The OpenSWATH framework provides a completely automated open-source workflow to analyze SWATH-MS datasets, thereby increasing throughput for targeted proteomics by several orders of magnitude compared to standard SRM approaches while retaining similar sensitivity and specificity. This method allowed us for the first time to analyze a whole microbial

proteome in a single injection using targeted proteomics and achieve accurate quantification for over 70 % of the expressed proteome. In conjunction with a novel algorithm using reference-free, non-linear retention time alignment, the pipeline produces very complete and reproducible proteomics data matrices suitable for systems biology studies. The analysis software as well as additional major improvements to fundamental computational infrastructure in the field has been implemented as part of the OpenMS software framework.

We then applied the SWATH-MS technique together with the OpenSWATH analysis pipeline to study the molecular basis of virulence in the human pathogen *Streptococcus pyogenes*. To investigate which proteomic and genetic factors contribute to virulence in *S. pyogenes*, we studied a set of environmental and genetic perturbations. We were able to reproducibly quantify over 900 proteins across 64 clinical isolates, resulting in a data set of unprecedented quality and size. Over 80 protein factors were found to dynamically change when challenged with a virulence-inducing environment and over 150 proteins were identified whose expression was directly influenced by underlying genetic polymorphisms. The detailed study of dynamic and evolutionary mechanisms of *S. pyogenes* virulence provides an improved understanding of the molecular basis of pathogenicity and indicates putative avenues for intervention. Furthermore, the approaches presented in this thesis can serve as a roadmap for similar studies in other pathogens.

In conclusion, this thesis presents computational methods that allow the simulation and analysis of targeted proteomics data generated by SRM and SWATH-MS. The freely available and open-source OpenSWATH software enables for the first time system-wide reproducible protein quantification across a large number of experimental samples by targeted proteomics. This is an essential requirement for many systems biological studies aimed at uncovering molecular mechanisms of health and disease. Therefore, the technology developed in this thesis has the potential to be applied to a wide array of biological questions.

Zusammenfassung

Die molekulare Systembiologie ist ein wissenschaftliches Forschungsgebiet, welches komplexe biologische Systeme mittels eines ganzheitlichen Ansatzes untersucht. Technologien, welche die quantitative Messung von Proteinen (Eiweissstoffen) erlauben, sind dabei von zentraler Bedeutung, weil Proteinen eine Reihe wichtiger Schlüsselfunktionen in einer Zelle zukommt. Zum gegenwärtigen Zeitpunkt zählen die Massenspektrometrie und die darauf basierenden proteinanalytischen Methoden (die Proteomik) zu den mächtigsten Technologien zur Charakterisierung komplexer Proteingemische, welche grob in zwei verschiedene Ansätze unterteilt werden können. Einerseits erlaubt die "shotgun" Methode die Identifikation eines grossen Teils der in einer Probe vorhandenen Proteine (hoher Durchsatz), hat aber den Nachteil einer geringen Reproduzierbarkeit über mehrere Messungen hinweg. Als Alternativen wurden deshalb zielgerichtete Proteomikmethoden ("targeted proteomics") wie z.B. SRM entwickelt, welche sich durch hohe Genauigkeit und Reproduzierbarkeit auszeichnen, aber nur Messungen mit geringem Durchsatz erlauben, was ihr Anwendungspotential für die Systembiologie stark einschränkt.

Die vorliegende Dissertation beschreibt mehrere Verbesserungen im Bereich der rechnergestützten Proteinanalytik mit spezieller Berücksichtigung von zielgerichteten Methoden und deren Einschränkungen bezüglich Spezifität und Durchsatz. In diesem Zusammenhang wurde zuerst die SRMCollider Software entwickelt, welche neuartige algorithmische Ansätze zur Berechnung und Vorhersage der Spezifität von massenspektrometrischen Messkoordinaten ("assays") implementiert. Die Software wurde danach erfolgreich angewandt, um die Spezifität von grösseren Sammlungen von ebensolchen Messkoordinaten in Mensch, Hefe und *Mycobacterium tuberculosis* zu überprüfen.

Weiterhin werden in dieser Arbeit Beiträge zur Entwicklung von einer neuartigen zielgerichteten Proteomikmethode mit hohem Durchsatz (SWATH-MS) beschrieben, wozu (i) theoretische Studien mit der SRMCollider Software durchgeführt wurden und (ii) die OpenSWATH Software entwickelt wurde. Angewandt auf SWATH-MS Daten, erhöht die komplett automatisierte und quelloffene OpenSWATH Software den Durchsatz

der zielgerichteten Proteomik im Vergleich zu SRM um mehrere Grössenordnungen bei vergleichbarer Sensitivität und Spezifität. Diese Methode erlaubte zum ersten Mal die Analyse eines gesamten mikrobiellen Proteoms mittels zielgerichteter Proteomik in einer einzigen Messung, wobei über 70 % aller exprimierten Proteine exakt quantifiziert werden konnten. Zusammen mit einem speziell hierzu entwickelten Algorithmus für die referenzfreie, nichtlineare Retentionszeitkorrektur, erstellt unsere Analysemethode äusserst komplette und reproduzierbare Proteindatenmatrizen, welche sich für systembiologische Ansätze eignen. Unsere Software wurde zusammen mit weiteren Beiträgen zur Grundinfrastruktur der rechnergestützten Proteinanalytik in das OpenMS Projekt integriert.

Im nächsten Schritt wurde die entwickelte Methode zur Untersuchung der molekularen Grundlagen der Pathogenität des menschlichen Krankheitserregers *Streptococcus pyogenes* angewandt. Um die Gene und Proteine zu identifizieren, welche zu dessen Virulenz beitragen, wurden mehrere *S. pyogenes* Stämme mittels Perturbationen in ihrer Umwelt und genetischen Natur untersucht. Wir konnten über 900 Proteine in allen 64 untersuchten Stämmen quantifizieren und eine Proteindatenmatrix in bisher unerreichtem Massstab berechnen. Über 80 verscheidende Proteine zeigten dynamische quantitative Änderungen, wenn *S. pyogenes* mit menschlichem Blutplasma konfrontiert wurde, und über 150 Proteine wurden identifiziert, deren Expression direkt durch genetische Mutationen beeinflusst wurde. Diese detaillierten Untersuchungen der dynamischen und genetischen Virulenzmechanismen des Krankheitserregers eröffnen biologisch wichtige Einblicke in die molekularen Grundlagen der Pathogenität und zeigen mögliche diagnostische Ansätze auf. Weiterhin bietet der hier präsentierte Ansatz einen Grundriss für vergleichbare Studien in anderen Krankheitserregern.

Zusammenfassend werden in der vorliegenden Arbeit mehrere rechnergestützte Methoden präsentiert, welche die Simulation und Analyse von zielgerichteten Proteomikdaten erlauben. Die frei erhältliche und quelloffene OpenSWATH Software erlaubte zum ersten Mal die reproduzierbare Messung einer grossen Anzahl Proteine über viele Messungen hinweg mittels zielgerichteter Proteomik. Dies ist ein essentieller Schritt für viele systembiologische Untersuchungen, welche die molekularen Grundlagen von Krankheit und Gesundheit aufklären wollen. Die in dieser Dissertation entwickelte Methodik hat daher das Potential, sich auch über die beschriebenen Anwendungen hinaus für eine grosse Menge an biologischen Fragestellungen als nützlich zu erweisen.