Design and Validation of Proteome Measurements

Manfred Claassen

Departement Informatik
ETH Zürich
manfred.claassen@inf.ethz.ch

Abstract: Die Proteomik ist ein Teilbereich der Biologie, der die vollständige Charakterisierung eines Proteoms zum Ziel hat. Massenspektrometrie basierte Proteomik hat sich als erfolgreichste Strategie zum Erreichen dieses Ziels herausgebildet. Diese Arbeit stellt statistische Methoden zur optimalen Planung und Validierung von Shotgun-Proteomik-Experimenten vor. Diese Methoden ermöglichen eine effiziente, zuverlässige und zugleich umfassende Proteomcharakterisierung.

Der erste Teil der Arbeit stellt Methoden zur Schätzung von False Discovery Raten für Peptid- und Proteinidentifikationen vor. Diese Methoden ermöglichen die zuverlässige und umfassende Identifikation von ungewöhnlichen chemischen Proteinmodifikationen. Die Anwendung dieser Methoden hat gezeigt, dass diese Varianten zu einem beträchtlichen Anteil der massenspektrometrischen Daten beitragen. Diese Arbeit stellt einen generalisierten Target-Decoy Ansatz zur Schätzung von False Discovery Raten für Proteinidentifikationen vor. Unsere Resultate zeigen, dass die Zuverlässigkeit von Proteinidentifikationen in grossen Studien bis dato bei weitem überschätzt wurde. Angesichts dieser Resultate schlagen wir Richtlinien für die Zusammenstellung von Proteinidentifikationen vor, die eine definierte Konfidenz gewährleisten. Dieser Teil schliesst mit der Formulierung eines generischen Systems zum Vergleich von Proteinidentifikationsmethoden, das die Zuverlässigkeit der Identifikationen berücksichtigt. Ein systematischer Vergleich von tausenden von Proteinidentifikationsvarianten hat gezeigt, dass einfache Methoden bereits optimale Performanz erzielen.

Der zweite Teil der Arbeit entwickelt einen nichtparametrischen Bayesschen Ansatz zur optimalen Planung von Shotgun-Proteomik-Studien. Hierfür wird die Aufgabe der Proteomabdeckungsvorhersage eingeführt. Ein erweitertes infinites Markovmodell wird zur Durchführung der Proteomabdeckungsvorhersage für einfache Shotgun-Proteomik-Experimente vorgestellt. Diese Arbeit stellt das neue Konzept eines fraktalen Dirichlet Prozesses vor, um die Ähnlichkeit der Peptidverteilungen in integrierten Proteomikstudien zu erfassen. Der fraktale Dirichlet Prozess erweitert den hierarchischen Dirichlet Prozess um selbstbezügliche Basismasse. Der fraktale Dirichlet Prozess wird erfolgreich zur Proteomabdeckungsvorhersage für integrierte Proteomikstudien verwendet. Diese Arbeit diskutiert rationale Stopkriterien für derartige Studien und evaluiert diese mit Hilfe der vorgestellten Methoden zur Proteomabdeckungsvorhersage. Schliesslich werden die Methoden zur Proteomabdeckungsvorhersage in einem System zur Planung von Proteomikstudien eingesetzt, dass eine Sequenz von Experimenten bestimmt, die den maximalen erwarteten Zuwachs der Proteomabdeckung erzielt.

Die Mechanismen von komplexen Krankheiten wie zum Beispiel Krebs oder Diabetes sind bis zum heutigen Tage nicht vollständig verstanden. Systembiologie zielt auf die Aufklärung dieser Mechanismen durch ausdrucksfähige Modelle von biologischen Systemen unter Berücksichtigung all Ihrer molekularen Bestandteile. Proteomik steuert neue Technologien zur Datenanalyse bei, die es erlauben, ein Proteom zu charakterisieren, d.h. die Proteinbestandteile eines biologischen Systems zu identifizieren. Jüngste technologische Fortschritte haben Biologen in die Lage versetzt, Proteome systematisch unter verschieden Perturbationen auszumessen und damit zu systemübergreifenden Modellen von biologischen Prozessen auf molekularer Ebene beizutragen [AM03]. Auf Massenspektrometrie basierte Proteomik hat sich als die erfolgreichste Technologie zur Charakterisierung eines Proteoms etabliert [WWY01, BAM$^+$07, BGG$^+$08, dGOC$^+$08, SWR$^+$09]. Die weitreichende Abdeckung eines Proteoms ist mit erheblichem experimentellem Aufwand verbunden, wozu eine sehr grosse Menge an verrauschten massenspektrometrischen Daten [Aeb09] gesammelt und analysiert werden muss. Die notwendigen statistischen Methoden zur rationalen Planung solcher Projekte und Bewertung der Verlässlichkeit der reportierten Proteinidentifikationen sind bisher nur unzureichend verstanden gewesen.

Diese Dissertation entwickelt neue statistische Konzepte zum Planen und Validieren von *Shotgun* Proteomik Experimenten und demonstriert wie diese angewendet werden, um effizient zuverlässige und weitreichende Proteomabdeckung zu erzielen.

Die weitreichenste Proteomabdeckung wurde bisher von der Shotgun Proteomik Strategie erreicht. Diese Strategie umfasst Schritte zur Extraktion von Proteinen aus der biologischen Quelle, deren enzymatischem Verdau und optional deren Fraktionierung nach physiko-chemischen Eigenschaften. Die resultierenden Peptidmixturen werden mit Tandem-Massenspektrometrie untersucht. Einzelne Peptidspezies werden dabei im Massenspektrometer isoliert und fragmentiert. Die dabei aufgenommenen Fragmentionenspektren geben Aufschluss über die Massen der entstandenen Fragmentionen. Diese Spektren stellen die Datengrundlage zur Rekonstruktion der Peptide bzw. der respektiven Proteine der initialen biologischen Probe dar. Eine Reihe von Suchmaschinen (*Search Engines*) wurden entwickelt, um in einem ersten Schritt Fragmentionenspektren den Peptidsequenzen einer Proteinsequenzdatenbank zuzuordnen [NVA07]. Die resultierenden Peptid-Spektrum Zuordnungen stellen die Grundlage für den letzten Auswertungsschritt dar, der Inferenz der biologisch relevanten Proteinidentitäten der ursprünglich untersuchten Probe [RM02, NA05].

Der stochastische Zusammenhang zwischen dem Objekt des Interesses, dem Protein, und dessen indirekter verrauschter Beobachtung, dem Fragmentionenspektrum, stellt die erwähnten Inferenzmethoden eine grosse Herausforderung dar und ist Ursache für die inhärente Unsicherheit, die Peptid- und Proteinidentifikationen anhaftet. Die Fähigkeit, diese Unschärfe zu quantifizieren, stellt eine zwingend notwendige Voraussetzung zur Auswertung dieser Identifikationen dar. Der erste Teil dieser Dissertation präsentiert Ansätze zur Quantifizierung dieser Unschärfe in Form von *False Discovery Rates* [BH95]. Diese Ansätze verallgemeinern die *Target-Decoy Strategie* zur Validierung von Peptid-Spektrum Zuordnungen von einfachen Datenbanksuchen [MYL02, EG07]. Allen *Target-Decoy*

Strategien ist die Grundidee gemeinsam, aus den Suchergebnissen gegen eine chimäre Sequenzdatenbank aus tatsächlich erwarteten Proteinsequenzen und nicht vorkommenden Nonsense-Sequenzen den Anteil der falsch positiven Identifikationen abzuschätzen. Die Herausforderung bei der Entwicklung von solchen Strategien besteht in der Gestaltung geeigneter chimärer Datenbanken und einer statistisch kohärenten Ableitung eines Schätzers für *False Discovery Rates* anhand der Identifikationen, die auf unbiologische synthetische bzw. Zufallssequenzen fallen.

In dieser Arbeit habe ich die einfache *Target-Decoy Strategie* für iterierte Datenbanksuchen angepasst. Diese Anpassung ermöglicht die effiziente Berücksichtigung von chemisch modifizierten Peptid- und Proteinvarianten, die nicht explizit in der Proteindatenbank aufgeführt werden (z.B. splice-Varianten). Dieser Ansatz hat die Zusammenstellung von zuverlässigen Identifikationen erlaubt, die aus einer iterierten Datenbanksuche resultierten, bei der hunderte von chemischen Aminosäurenvarianten für die Suche zugelassen waren. Die Anwendung dieser Strategie hat ergeben, dass ein signifikanter Anteil der spektralen Evidenz in Shotgun Proteomik Experimenten auf modifizierte, zum Teil bisher nicht ausreichend berücksichtigte Peptidvarianten zurückzuführen ist [CABc].

In einem weiteren Schritt stellt diese Dissertation eine generalisierte Target-Decoy Strategie vor, um *False Discovery Rates* für Proteinidentifikationen zu schätzen. Proteinidentifikationen sind als Gruppen von Peptid-Spektrum Zuordnungen definiert und stellen das eigentlich biologisch relevante Ergebnis eines Shotgun Proteomik Experimentes dar. Eine Reihe verschiedener Methoden zur Inferenz von Proteinidentifikationen wurden in den letzten Jahren vorgestellt [RM02, NA05]. Da die statistische Signifikanz von Proteinidentifikationen jedoch bisher nur zum Teil verstanden wurde, war es schwierig, die Güte der einzelnen Methoden zur Proteininferenz zu bewerten. Obwohl Proteinidentifikationen das biologisch relevante Ergebnis einer Proteomikstudie sind, wurden Konfidenzmasse typischerweise auf dem Level der Peptid-Spektrum Zuordnungen reportiert. Diese Vorgehensweise nimmt implizit an, dass die Konfidenzmasse für Peptid-Spektrum Zuordnungen eine gute Approximation für die entsprechenden Masse für Proteinidentifikationen sind. Wir zeigen, dass die Fehler auf Ebene der Peptid-Spektrum Zuordnungen jedoch in einer nicht-trivialen Weise auf die Ebene der Proteinidentifikationen propagieren. Hierfür stellen wir eine neue Methode vor, die die Statistik dieser Fehlerpropagierung berücksichtigt und auf diese Weise erlaubt, explizit *False Discovery Rates* von Proteinidentifikationen zu schätzen [CRS$^+$09].

Unsere Strategie basiert auf der etablierten Annahme, dass falsch positive Peptid-Spektrum Zuordnungen mit gleicher Wahrscheinlichkeit entweder auf die Target- oder Decoy-Datenbank fallen. Wir verwenden die Anzahl der in der Decoy-Datenbank identifizierten Proteine als einen Schätzer von Proteinen in der Target-Datenbank, die falsch positive Peptid-Spektrum Zuordnungen enthalten. Da das Vorkommen von falsch positiven Peptid-Spektrum Zuordnungen nur eine notwendige, aber nicht hinreichende Bedingung für eine falsch positive Proteinidentifikation ist, stellt dieser Schätzer jedoch noch nicht den gewünschten Schätzer für die falsch positiven Proteinidentifikationen dar. Um den

gewünschten Schätzer abzuleiten, treffen wir die Annahme, dass falsch positive Proteinidentifikationen uniform über die Proteindatenbank verteilen. Wir zeigen, dass unter dieser Annahme die Anzahl der falsch positiven Proteinidentifikationen einer Variante der hypergeometrischen Verteilung folgt. Die Spezifikation dieser Verteilung ermöglicht es den Erwartungswert der Anzahl falsch positiver und damit die *False Discovery Rate* der Proteinidentifikationen anzugeben.

Auf der Basis von *False Discovery Rate* für alle Proteinidentifikationen haben wir eine Target-Decoy Strategie entwickelt, um lokale *False Discovery Rates* für Teilmengen von Proteinidentifikationen zu bestimmen. Damit ist es nun möglich, die Verlässlichkeit von Identifikationsgruppen besonderen Interesses separat zu quantifizieren. In diesem Kontext haben wir unter anderem in der Community umstrittene Identifikationen untersucht, die ausschliesslich durch eine einzelne Peptid-Spektrum Zuordnung nachgewiesen wurden und konnten zeigen, dass diese tatsächlich mit grosser Unsicherheit behaftet sind.

Diese Methoden erlauben es zum ersten Mal, die Unschärfe von Proteinidentifikationen in heterogenen Datensätzen beliebiger Grösse akkurat zu quantifizieren. Mit Hilfe dieser Methode haben wir insbesondere herausgefunden, dass der Anteil an falsch positiven Proteinidentifikationen bisher stark unterschätzt wurde. Wir haben weiter festgestellt, dass dieser Unterschied um so stärker ausfällt je grösser das zugrunde liegende Datenvolumen ist. Diese Ergebnisse haben insbesondere Implikationen für die Interpretation von Daten im Kontext von grossen Shotgun Proteomik Studien, die eine grosse Proteomabdeckung zum Ziel haben. Jenseits individueller Studien, trifft diese Erkenntnis auch auf die Bewertung von Proteomics-Datenbanken zu. Unser Ansatz zur Schätzung von *False Discovery Rates* für Proteinidentifikationen kann zur automatischen Kurierung solcher Datenbanken verwendet werden und trägt damit nicht zuletzt auch zu Systembiologieprojekten bei, die auf diesen Ressourcen basieren

Diese Arbeit stellt in einen weiteren Schritt einen formalen Ansatz vor, der es für ein gegebenes Arsenal an Interpretationsmethoden wie z.B. Suchmaschinen und Proteininferenzmaschinen erlaubt, Richtlinien zur optimalen Auswertung von massenspektrometrischen Daten abzuleiten. Bis vor kurzem war ein ausgewogener Vergleich der vielen unterschiedlichen Proteininferenzmethoden nicht möglich, da die Güte von Proteinidentifikationen nicht angemessen quantifizierbar war. Wir schlagen ein Leistungsmass für Proteininferenzmethoden vor, dass einen solchen Vergleich erlaubt. Dieses Leistungsmass misst die erwartete Anzahl korrekter Proteinidentifikationen für eine benutzerdefinierte *False Discovery Rate* auf Ebene der Proteinidentifikationen. Für die Datensätze, die in dieser Arbeit untersucht worden sind, haben wir gefunden, dass die beste Strategie darin besteht, alle spektrale Evidenz von ausreichender Qualität zu berücksichtigen. Wir haben insbesondere herausgefunden, dass für grosse Studien die spektrale Evidenz mit weit grösserer Sorgfalt ausgewählt werden muss, als bisher angenommen. [CRH+10].

Der **zweite Teil** dieser Dissertation entwickelt einen nichtparametrisch Bayesschen Formalismus zur optimalen Planung einer Shotgun Proteomik Studie, die eine effiziente Ab-

deckung eines Proteoms zum Ziel hat. Komplementär zum ersten Teils der Arbeit, deren Gegenstand die bestmöglichen Auswertung vorhandener experimenteller Daten ist, behandelt der zweite Teil die rationale Vorabauswahl von Experimenten mit maximalem erwarteten Informationsgehalt.

Erfahrungsgemäss erzeugen Shotgun Proteomik Studien, die eine grosse Proteomabdeckung zum Ziel haben, einen Grossteil redundanter, uninformativer Daten indem sie die selben Peptide immer wieder aufs Neue messen [SCA09]. Aufgrund von Effizienz- und Sensitivitätserwägungen, ist es von Vorteil diese Redundanz zu vermeiden. Die Vermeidung von redundanten Informationen schont offensichtlich Ressourcen, ohne Einschränkungen in der Proteomabdeckung hinnehmen zu müssen. Darüber hinaus führt die Vermeidung dieser Experimente aber auch zu kleineren Datensätzen, die, wie wir in unserer vorhergehenden Studie zur Verlässlichkeit von Proteinidentifikationen gesehen haben, auch zu einer verminderten Akkumulation von falsch positiven Identifikationen führen. Dieser Effekt könnte den Nachweis von weniger häufig observierten Proteinen erleichtern und sich damit sogar positiv auf die Abdeckung auswirken. Zusammengefasst ist es also von Vorteil, eine Shotgun Proteomik Studie so zu planen, dass sie sich auf die informativen Experimente konzentriert [BCA11].

Diese Dissertation trägt ein formales System zur Planung von Proteomik Studien bei, um deren erwartete Proteomabdeckung zu maximieren. Basierend auf einer kleinen Menge von bereits durchgeführten (LC-MS/MS) Experimenten, eignet sich dieses System zur Vorhersage einer Sequenz von Experimenten mit optimaler erwarteter Proteomabdeckung. Diese Arbeit entwickelt einen nichtparametrischen Bayesschen Formalismus um diese Aufgabe zu implementieren. Dieser Formalismus charakterisiert die Peptidverteilungen, die im Laufe eines LC-MS/MS Experimentes untersucht werden, mit Hilfe von geeignete Varianten von hierarchischen Dirichlet Prozessen [TJBB06]. Die komplexen Ähnlichkeitsmuster zwischen den Peptidverteilungen in multidimensionalen Fraktionierungsexperimenten hat die Formulierung des fraktalen Dirichlet Prozesse, einer neuen Klasse von rekursiven stochastischen Prozessen inspiriert. Wir haben gezeigt, wie diese Prozesse verwendet werden können, um akkurat Proteomabdeckungsvorhersage ausgehend von einer kleine Anzahl von Experimenten durchzuführen.

Zuerst führe ich die Aufgabe der *Proteomabdeckungsvorhersage* ein [CAB09]. Diese Aufgabe umfasst die Schätzung der erwarteten Anzahl von neuen Proteinidentifikationen nach Durchführung einer spezifizierten Sequenz von Shotgun Proteomik Experimenten. *Proteomabdeckungsvorhersage* stellt eine zentrale Aufgabe bei der Planung von Shotgun Proteomik Projekten dar. Neben deren Rolle in der Planung von grossen Projekten, ermöglicht *Proteomabdeckungsvorhersage* die Formulierung von rationalen Stop-Kriterien für bereits fortgeschrittene Projekte, die bereits nahezu maximale Proteomabdeckung erreicht haben. Proteomabdeckungsvorhersage kann darüber hinaus die Entwicklung von neuen experimentellen Methoden mit grossen Abdeckungspotential unterstützen.

Ich habe ein erweitertes infinites Markov Modell entwickelt, das es erlaubt, Proteomab-

deckung für Wiederholungen eines Flüssigkeitschromatografie Tandem-Massenspektrometrie (LC-MS/MS) Experimentes vorherzusagen [CAB09]. LC-MS/MS ist das elementare Experiment einer Shotgun Proteomik Studie. Während eines solchen Experimentes werden Peptide aus einer Vielzahl von unbekannten Peptidverteilungen gezogen. Diese Peptidverteilungen haben üblicherweise überlappenden Support. Das vorgestellte infinite Markov Modell erlaubt es, diese Verteilungen konsistent mit den erhobenen massenspektrometrischen Daten auf Bayessche Art und Weise zu charakterisieren und schlussendlich Proteomabdeckungsvorhersage durchzuführen. Das Modell implementiert einen hierarchischen Dirichlet Prozess. Ein solcher Prozess stellt eine Verteilung über eine Menge von unendlich dimensionalen diskreten Verteilungen dar. Dieser Prozess ist derart gestaltet, dass die einzelnen diskreten Verteilungen mit positiver Wahrscheinlichkeit gemeinsame Atome aufweisen können. Aufgrund der Konjugiertheitseigenschaften dieses Prozesses ist weiterhin für gegebene Observationen eine kompakte Darstellung der prediktive Verteilung einer Prozessinstanz möglich. Aufgrund dieser Eigenschaften ist dieser Prozess geeignet, die prediktiven Peptidverteilungen eines LC-MS/MS-Experimentes für eine Reihe von bereits durchgeführten Experimenten zu charakterisieren. Proteomabdeckungsvorhersage ergibt sich nun natürlich durch Sampling der prediktiven Peptidverteilungen.

Proteomabdeckungsvorhersage für ein *D. melanogaster* Datensatz ergab das folgende Resultat: Die maximal mögliche Abdeckung durch die Anhäufung von falsch positiven Peptididentifikationen ist beschränkt und liegt insbesondere unterhalb der tatsächlich vorhandenden Gesamtzahl der Proteine, der sogenannten Sättigungsabdeckung.

Die meisten grossen Shotgun Proteomik Studien basieren auf multidimensionalen Fraktionierungsexperimenten, die ein Ensemble von überlappenden Peptid- bzw. Proteinverteilungen untersuchen [BCA11]. Ein Modell für eine solche Ansammlung von Experimenten setzt die Berücksichtigung dieses Überlapps voraus.

Diese Anforderung inspirierte das neue generelle Konzept des *fraktalen Dirichlet Prozesses* [CABa]. Dieser stochastische Prozess generalisiert den hierarchischen Dirichlet Prozess [TJBB06] durch die Einführung von selbstbezüglichen Basismaen. Der hierarchische Dirichlet Prozess beschreibt eine Verteilung über eine Menge von (unendlich dimensionale) diskrete Verteilungen. Jede einzelne diskrete Verteilung folgt einem Dirichlet Prozess, deren Basismass eine gemeinsame diskrete Verteilung ist. Diese Konstruktion gewährleistet, dass die Verteilungen eines hierarchischen Dirichlet Prozesses ähnlich zu einander sind und insbesondere Atome teilen können. Der *fraktale Dirichlet Prozess* verwendet ein Ensemble von Basismassen, die Linearkombination von allen zu ziehenden Verteilungen sind. Dieses Konzept erlaubt es explizit, den Überlapp, bzw. die Ähnlichkeit zwischen beliebigen Teilmengen einer Gruppe von diskreten Verteilungen zu erfassen. Die Charakterisierung der a posteriori Verteilung des *fraktalen Dirichlet Prozesses* führt auf ein System von Rekursionsgleichungen. Die Lösung dieser Gleichungen ermöglichte die Formulierung eines Gibbs Samplers für Bayessche Inferenz der Prozessparameter.

Ich stelle eine Variante des *fraktalen Dirichlet Prozesses* vor, um Proteomabdeckungsvor-

hersage für multidimensionale Fraktionierungsexperimente durchzuführen [CAB11]. Die Anwendung dieser Methode auf einen Datensatz für das Bakterium *L. interrogans* ergab, dass Proteomabdeckung bereits mit einigen wenigen Experimenten sehr genau vorhergesagt werden kann. Extrapolation des gegebenen Datensatzes hat weiterhin gezeigt, dass die Sättigungsabdeckung bereits erreicht wurde und damit weitere Experimente keinen nennenswerten Zuwachs an neuen Proteinidentifikationen erwarten lassen [SBM$^+$]. Eine weitere Anwendung dieser Methode bestätigte das Erreichen der Sättigungsabdeckung im Kontext Studien des menschlichen Proteoms [BMS$^+$] und des Nematoden *C. elegans* [SJR$^+$].

Dieser Teil der Dissertation schliesst mit der Beschreibung eines Ansatzes zur optimalen Planung einer Shotgun Proteomik Studie [CABb]. Optimale Planung wird in diesem Zusammenhang als kombinatorisches Optimierungsproblem definiert. Ziel dieser Optimierung ist es eine Sequenz von Experimenten zu finden, deren erwartete Proteomabdeckung maximal ist. Die erwartetete Proteomabdeckung einer Sequenz von Experimenten kann mit Hilfe der oben beschriebenen Methoden angegeben werden. Wir zeigen, dass das Optimierungsproblem auf ein Maximum K-Cover Problem reduziert werden kann. Wir explorieren verschiedene Möglichkeiten, dieses Problem zu lösen und für die Aufgabe der optimalen Experimentplanung einzusetzen.

Die statistischen Konzepte, die in dieser Dissertation entwickelt wurden, sind nicht auf Anwendungsszenarien in der Massenspektrometrie basierten Proteomik beschränkt. Der erste Teil der Arbeit generalisiert die Target-Decoy Strategie zur Schätzung von *False Discovery Rates* von Proteinidentifikationen, d.h. von Aggregaten von Peptid-Spektrum Zuordnungen. Die Target-Decoy Strategie ist generell anwendbar zur Bewertung der Zuverlässigkeit von Inferenzreslutaten, die auf der Zuordnung von Observationen (z.B. Peptid-Spektrum Zuordnungen) zu Hypothesen (z.B. Proteinen) basieren, die aus einer statischen Kollektion von Hypothesen ausgewählt werden (z.B. Proteindatenbank). Nach Zusammenstellung einer geeigneten Kollektion von Decoy-Hypothesen, kann die Target-Decoy Strategie einfach angewendet werden. Ein weiteres Einsatzgebiet wäre zum Beispiel eine Retrieval-Aufgabe, die die Zuordnung von Melodie-Abschnitten zu Liedern aus einer Musikdatenbank umfasst.

Der zweite Teil der Arbeit führt den fraktalen Dirichlet Prozess ein. Dieser Prozess stellt ein neues Mass über einer Menge von diskreten Massen dar und generalisiert den hierarchische Dirichlet Prozess. Wir konnten zeigen, dass ein solcher Prozess besonders geeignet ist, um die Ähnlichkeitsmuster zwischen Peptidverteilungen in integrierten Shotgun Proteomik Experimenten einzufangen. Es wird interessant sein, ob es weitere Anwendungsszenarien gibt, die ebenfalls Strukturen aufweisen, die besonders gut durch den fraktalen Dirichlet Prozess repräsentiert werden.

Diese Dissertation schlägt neue statistische Methoden vor, die Experimentatoren rationale Entscheidungskriterien liefert, welche Daten zu messen sind und wie diese am Besten ausgewertet werden können, um letzten Endes effizient die grösstmögliche und zugleich

verlässliche Proteomabdeckung zu erzielen. Die auf diese Weise erzeugten Ergebnisse stellen eine wichtige Ressource für weitere quantitative orientierte Proteomik Ansätze dar und stärken auf diese Weise die Rolle von proteomischen Daten im Kontext von interdisziplinären Systembiologieprojekten, die auf heterogenen Datenressourcen basieren.

Diese Arbeit adressiert Fragen und entwickelt Methoden, die an der Grenze zwischen Biologie und maschinellem Lernen positioniert sind. Diese Arbeit macht die Verwendbarkeit von Konzepten des maschinellen Lernens zur Beantwortung von biologisch relevanten Fragen deutlich und veranschaulicht zugleich, wie biologische Fragestellungen neue Aufgaben und Konzepte für das Feld des maschinellen Lernens inspirieren kann. Ich bin überzeugt davon, dass beide Disziplinen in Zukunft weiterhin von dieser Synergie profitieren werden.

Literatur

[Aeb09] R. Aebersold. A stress test for mass spectrometry-based proteomics. *Nat Methods*, 6(6):411–2, 2009.

[AM03] R. Aebersold und M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[BAM+07] E. Brunner, C. H. Ahrens, S. Mohanty, H. Baetschmann, S. Loevenich, F. Potthast, E. W. Deutsch, C. Panse, U. de Lichtenberg, O. Rinner, H. Lee, P. G. Pedrioli, J. Malmstrom, K. Koehler, S. Schrimpf, J. Krijgsveld, F. Kregenow, A. J. Heck, E. Hafen, R. Schlapbach und R. Aebersold. A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol*, 25(5):576–83, 2007.

[BCA11] M. Beck, M. Claassen und R. Aebersold. Comprehensive proteomics. *Current Opinion in Biotechnology*, 22:3–8, 2011.

[BGG+08] Katja Baerenfaller, Jonas Grossmann, Monica A. Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem und Sacha Baginsky. Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science*, 320(5878):938–941, 2008.

[BH95] Yoav Benjamini und Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, Seiten 289–300, 1995.

[BMS+] M. Beck, J. Malmstroem, A. Schmidt, M. Claassen, O. Rinner, F. Herzog und R. Aebersold. Comprehensive Proteome Map of a Human Cell Line. *in prep.*

[CABa] M. Claassen, R. Aebersold und J. M. Buhmann. The Fractal Dirichlet Process. *in prep.*

[CABb] M. Claassen, R. Aebersold und J. M. Buhmann. Optimal Design of Integrated Proteomics Experiments. *in prep.*

[CABc] M. Claassen, R. Aebersold und J. M. Buhmann. Reliable, Efficient and Comprehensive Identification of Modified Peptides with an Iterated Target-Decoy Database Search Strategy. *in prep.*

[CAB09] M. Claassen, R. Aebersold und J. M. Buhmann. Proteome coverage prediction with infinite Markov models. *Bioinformatics*, 25(12):i154–60, 2009.

[CAB11] M. Claassen, R. Aebersold und J. M. Buhmann. Proteome Coverage Prediction for Integrated Proteomics Datasets. *Journal of Computational Biology*, 18(3):283–293, 2011.

[CRH+10] M. Claassen*, L. Reiter*, M. O. Hengartner, J. M. Buhmann und R. Aebersold. Generic Comparison of Protein Inference Engine Families. *Molecular & Cellular Proteomics*, submitted, 2010.

[CRS+09] M. Claassen*, L. Reiter*, S. P. Schrimpf, M. Jovanovic, A. Schmidt, J. M. Buhmann, M. O. Hengartner und R. Aebersold. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol Cell Proteomics*, 8(11):2405–2417, 2009.

[dGOC+08] L.M.F. de Godoy, J.V. Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, F. Fröhlich, T.C. Walther und M. Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, 2008.

[EG07] J. E. Elias und S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–14, 2007.

[MYL02] R.E. Moore, M.K. Young und T.D. Lee. Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[NA05] A.I. Nesvizhskii und R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics*, 4(10):1419, 2005.

[NVA07] A. I. Nesvizhskii, O. Vitek und R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, 4(10):787–97, 2007.

[RM02] J. Rappsilber und M. Mann. What does it mean to identify a protein in proteomics? *Trends Biochem Sci*, 27(2):74–8, 2002.

[SBM+] A. Schmidt, M. Beck, J. Malmstroem, H.N. Lam, M. Claassen, D. Campell und R. Aebersold. Proteome-wide high-throughput screening using directed mass spectrometry: Application to the human pathogen L. interrogans. *in prep.*

[SCA09] A. Schmidt, M. Claassen und R. Aebersold. Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr Opin Chem Biol*, 13(5-6):510–7, 2009.

[SJR+] S.P. Schrimpf, M. Jovanovic, L. Reiter, M. Claassen, J. Malmstrm, A. Sendoel, E. Brunner, B. Roschitzki, C. Panse, R. Schlapbach, P.E. Hunziker, R. Aebersold und M.O. Hengartner. Complementary Separation Techniques to Identify Complex Proteomes. *in prep.*

[SWR+09] S.P. Schrimpf, M. Weiss, L. Reiter, C.H. Ahrens, M. Jovanovic, J. Malmström, E. Brunner, S. Mohanty, M.J. Lercher, P.E. Hunziker et al. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol*, 7(3):e48, 2009.

[TJBB06] Yee W. Teh, Michael I. Jordan, Matthew J. Beal und David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2006.

[WWY01] M. P. Washburn, D. Wolters und J. R. Yates, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242–7, 2001.

Manfred Claassen wurde geboren am 10. Mai 1977, in Maracay, Venezuela. Er studierte Biochemie und Informatik an der Universität Regensburg, Eberhard Karls Universitt Tübingen und der Universit Claude Bernard Lyon. An der Universität Tübingen schloss er 2004 den Diplomstudiengang Biochemie und 2006 den Diplomstudiengang Informatik ab. Während des Studiums arbeitete er als wissenschaftlicher Mitarbeiter in der Brock Gruppe (Tübingen, Immunfluoreszenzmikroskopie), der Pietro Gruppe (Lyon, Multi Drug Resistance) und der Kohlbacher Gruppe (Tübingen, Proteinstrukturmodellierung). Von September 2006 bis Mai 2010 arbeitete Manfred Claassen als wissenschaftlicher Mitarbeiter in den Gruppen von Ruedi Aebersold und Joachim Buhmann an der ETH Zürich. Im Rahmen seiner interdisziplinären Doktorarbeit entwickelte er neuartige Konzepte maschinellen Lernens zur Auswertung von massenspektrometrie basierten Proteomik Daten. Nach Abschluss seines Doktorats im Mai 2010 blieb er der ETH füer einen kurzen Postdoc bis Dezember 2010 erhalten. Eine seiner publizierten Arbeiten wurde auf der ISMB 2009 mit dem Ian Lawson Van Toch Memorial Award for Outstanding Student Paper ausgezeichnet. Desweiteren erhielt Manfred Claassen 2010 das Stipendium fr Angehende Forschende des Schweizerischen Nationalfonds SNSF. Seit Januar 2011 ist er als Postdoc in der Gruppe von Daphne Koller an der Stanford University tätig und forscht an probabilistischen Modellen zur Beschreibung von molekularen Systemen in der Immunologie.