Design and Validation of Proteome Measurements

DISS. ETH Nr. 19037

Design and Validation of Proteome Measurements

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of

Doctor of Sciences

presented by

Manfred Claassen
Dipl. Biochem. & Dipl. Inf.
University of Tübingen
born May 10 1977
citizen of Germany

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann examiner
Prof. Dr. Ruedi Aebersold co-examiner
Prof. Dr. Oliver Kohlbacher co-examiner

2010

Abstract

Proteomics is a branch in biology that aims to comprehensively characterize a proteome. Mass spectrometry based proteomics has proven to be the most powerful approach to achieve this goal. This thesis introduces statistical concepts to optimally design and validate shotgun proteomics experiments and thereby enables to efficiently achieve reliable and extensive proteome coverage.

The first part reports methods to estimate false discovery rates for peptide and protein identifications. These approaches enabled to reliably and comprehensively identify unusually modified protein variants. It turned out that these variants contribute to a significant fraction of the spectral evidence. This work presents a generalized target-decoy approach to estimate false discovery rates for protein identifications. This work shows evidence that the reliability of protein identifications in large studies has so far been largely overestimated and provides guidelines to compile identifications at well defined confidence. This part concludes with formulating a generic framework to compare protein inference engines based on protein identification false discovery rates. A systematic comparison of thousands of protein inference variants revealed that simple approaches yield optimal inference performance.

The second part develops a nonparametric Bayesian approach to optimally design shotgun proteomics studies. Therefore the proteome coverage prediction task is introduced. An extended infinite Markov model is presented to perform proteome coverage prediction for simple shotgun proteomics experiments is presented. To capture the intricate similarities among peptide distributions arising in integrated shotgun proteomics studies, this work developed the general concept of the fractal Dirichlet process that augments the hierarchical Dirichlet process by introducing self-referential base measures. The fractal process is successfully applied to predict proteome coverage for integrated shotgun proteomics datasets. Rational stop criteria for these studies are discussed and evaluated by means of the proteome coverage prediction approaches. Finally the proteome coverage

approaches are integrated into a study design framework that enables to determine an experimental sequence that achieves maximal expected increase in proteome coverage.

Zusammenfassung

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 Proteomikstu-

dien 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.