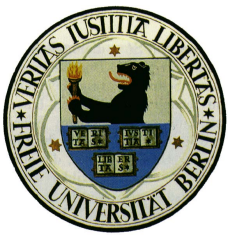
Analysis of sets and collections of Peptide Mass Fingerprint data



Dissertation zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von
Witold Eryk Wolski
aus Warschau

31.05.2007

Gutachter

- Erster Gutachter : Prof. Dr. Udo Heinemann
  (Department of Crystallography, Max-Delbrück-Center for Molecular Medicine)
- Zweiter Gutachter : Prof. Dr. Knut Reinert
  (Institut für Informatik, Fachbereich Mathematik und Informatik, Freie Universität Berlin)

I would like to dedicate this work to Pawel Grec and Martin Mätzig.

Acknowledgements

During the four years of my graduate studies, I have had cause to be grateful for the advice, support and understanding of many people. In particular, my supervisors, Knut Reinert and Hans Lehrach, had been a constant source of good ideas and sound advice.

Much advice and many useful discussions were provided by Maciej Lalowski, Johan Gobom, Ralf Herwig, Thomas Kreitler, Peter Martus, Peter Jungblut, Clemens Gröpl and Andreas Döring.

Last but not least I would like to thank Beate Schröder for her understanding, support and true friendship during my work on this thesis.

Abstract

Recent advances in genomics, which outstanding achievements were exemplified by the complete sequencing of the human genome provided the infrastructure and information enabling the development of several proteomic technologies. Currently no single proteomic analysis strategy can sufficiently address the question of how the proteome is organised in terms of numerical complexity and complexity generated by the protein-protein interactions forming supramolecular complexes within the cell. In order to bring a detailed structural/functional picture of these complexes in whole genomes, cells, organelles or in normal and pathological states several proteomic strategies can be utilised. Combination of technologies will bring a more detailed answer to what are the components of certain cellular pathways (e.g.: targets of kinases/phosphatases, cytoskeletal proteins, signalling molecules), how do they interconnect, how are they modified in the cell and what are the roles of several complex components in normal and disease conditions. These types of studies depend on fast and high throughput methods of protein identification. One of the most common methods of analysis is mass spectrometric technique called peptide mapping. Peptide mapping is the comparison of mass spectrometrically determined peptide masses of a sequence specific digest of a single protein or peptide of interest with peptide masses predicted from genomic databases. In this work several contributions to the computational analysis of mass spectrometric data are presented. During the course of my studies I looked at the distribution of peptide masses in sequence specific protein sequence digests and developed a simple mathematical model dealing with peptide mass cluster centre location. I have introduced and studied the methods of calibration of mass spectrometric peak-list without resorting to internal or external calibration samples. Of importance is also contribution of this work to the calibration of data produced in high throughput experiments. In addition, I studied how filtering of non-peptide peaks influences the identification rates in mass spectrometric instruments. Furthermore, I focused my studies on measures of spectra similarity which can be used to acquire supplementary information, increasing the sensitivity and specificity of database searches.

Original Publications

I **Wolski WE**, Farrow M, Emde AK, Lehrach H, Lalowski M, Reinert K: **Analytical model of peptide mass cluster centres with applications**. *Proteome Sci* 2006, **4**:18.

II **Wolski WE**, Lalowski M, Jungblut P, Reinert K: **Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants**. *BMC Bioinformatics* 2005, **6**:203.

III **Wolski WE**, Lalowski M, Martus P, Herwig R, Giavalisco P, Gobom J, Sickmann A, Lehrach H, Reinert K: **Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process**. *BMC Bioinformatics* 2005, **6**:285.

Contents

List of Figures

List of Tables

Nomenclature

$2D-PAGE$ Two Dimensional Polyacrylamide Gel Electrophoresis

mod modulo operator

$ANOVA$ analysis of variance

CID collision induced dissociation

DFT Discrete Fourier Transformation

ESI Electrospray Ionisation

EST Expressed Sequence Tag

FN false negative

FP false positive

$HPLC$ high performance liquid chromatography

m/z mass over charge

$MALDI$ matrix assisted laser desorption/ionisation

MS Mass Spectrometry

MST minimum spanning tree

OO Object-Oriented

$PAUC$ partial area of interest under ROC curve

$PBMS$ Probability based mascot score

PMF Peptide Mass Fingerprinting

ROC receiver operator characteristic

SDS Sodium dodcylsulphate

TN true negative

TOF time of flight

TP true positive

TPS Thin-Plate Spline

LR/PR peptide rule calibration