New statistical algorithms for clinical proteomics

T. Conrade

February 17, 2006

Abstract

Background: Mass spectrometry based screening methods have been recently introduced into clinical proteomics. This boosts the development of a new approach for early disease detection: proteomic pattern analysis.
Aim: Find, analyze and compare proteomic patterns in groups of patients having different properties such as disease status or epidemio-logical parameters (e.g. sex, age) with a new pipeline to enhance sensitivity and specificity.
Problems: Mass data acquired from high-throughput platforms frequently are blurred and noisy. This extremely complicates the reliable identification of peaks in general and very small peaks below noise-level in particular.
Approach: Apply sophisticated signal preprocessing steps followed by statistical analyzes to purge the raw data and enable the detection of real signals while maintaining information for tracebacks.
Results: A new analysis pipeline has been developed capable of finding and analyzing peak patterns discriminating different groups of patients (e.g. male/female, cancer/healthy). First steps towards distributed computing approaches have been incorporated in the design.

Introduction

Todays mass spectrometry protein fingerprinting techniques rely on the analysis of highly complex spectra obtained by high-troughput platforms in clinical settings. Algorithms have to deal with a substantial amount of noise, systematic errors introduced by the machines, biochemical degradation processes in the samples, and difficult and error-prone ionization techniques. Within this "crossfire of influences" they seek to isolate and analyze signals leading to biomarkers possibly being useful for further clinical studies.

Preliminary Achievements

In the first 8 (of 12 estimated) months of this project a pipeline has been designed and implemented that preprocesses the raw data, provided by the group of Prof. Thiery[1], distills relevant signals and analyzes these features for several tasks.

[1]More than 10.000 spectra from about 750 apparently healthy individuals, taken and evaluated at the University Hospital Leipzig.

Dagstuhl Seminar Proceedings 05471
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2006/542

First stage low-level preprocessing techniques such as smoothing and reduction of the baseline applied to the raw data prepare the signal. During the second stage peaks in the adjusted signal are sought for and properties, such as width, height, or shape of these peaks are stored in a database. Unlike other methods we do not apply noise-filtering because our studies have shown this to result in loss of many pivotal marker. These sets of peaks for each spectrum will be used as the basis for detection of markers for discrimination of patient groups and further analyzes, such as correlation detection between peaks and particular blood-parameters.

Each of these individual steps has a sanity check and every intermediate result is assigned a quality value together with a link to the preceding result to provide reliable, verifiable results.

The third level of algorithms are the actual analyzes working on the reliable distilled features from the preceding steps. These analyzes cover two main directions at the moment:

- *Finding Differential Peak Patterns* between groups of patients (e.g. male/female, cancer/healthy)
- *Analyzing correlations* of peaks to blood parameters or patient meta-data

For the automated, quick and easy usage of the pipeline and to hide the wiring of the algorithms from the user a **.NET 2.0** web front-end has been designed that directly interacts with the **MS-SQL 2005** database. This front-end executes and controls the algorithms, manages the data and visualizes the results.

With this suite of algorithms and the vast amount of clinical data from our project partners[2] we were able to outperform state-of-the-art commercial software products such as **ClinProtTools**[3] by more than 10% in specificity and sensitivity. **These results will be submitted for publication until the end of Februray 2006.**

Future Goals

Up until now the project has generated very important experience for handling mass spectrometry data and crucial algorithms for preprocessing and analysis of them. This toolbox has been tied together in an execution pipeline with defined input and output interfaces, wrapped up by a web front-end and provides its data to other services in a defined and documented high-performance database environment. Thanks to its modular implementation, parts are easily exchangeable and new algorithms can be plugged-in without any problems.

This project was initially planned as a proof-of-concept to evaluate whether statistically relevant results can be generated for analyzes of highly complex and noisy mass data. This has been fully achieved and has even exceeded former expectations. The next two paragraphs describe necessary and optional extensions to this successfull initial research to devise new ways of analyzing mass spectrometry data and exploit their wealth of lurking information content.

[2] Group of Prof. Thiery, University Hospital Leipzig

[3] By Bruker Daltronics, worldmarket leader in MALDI-TOF mass spectrometry devices. See `www.bdal.com`