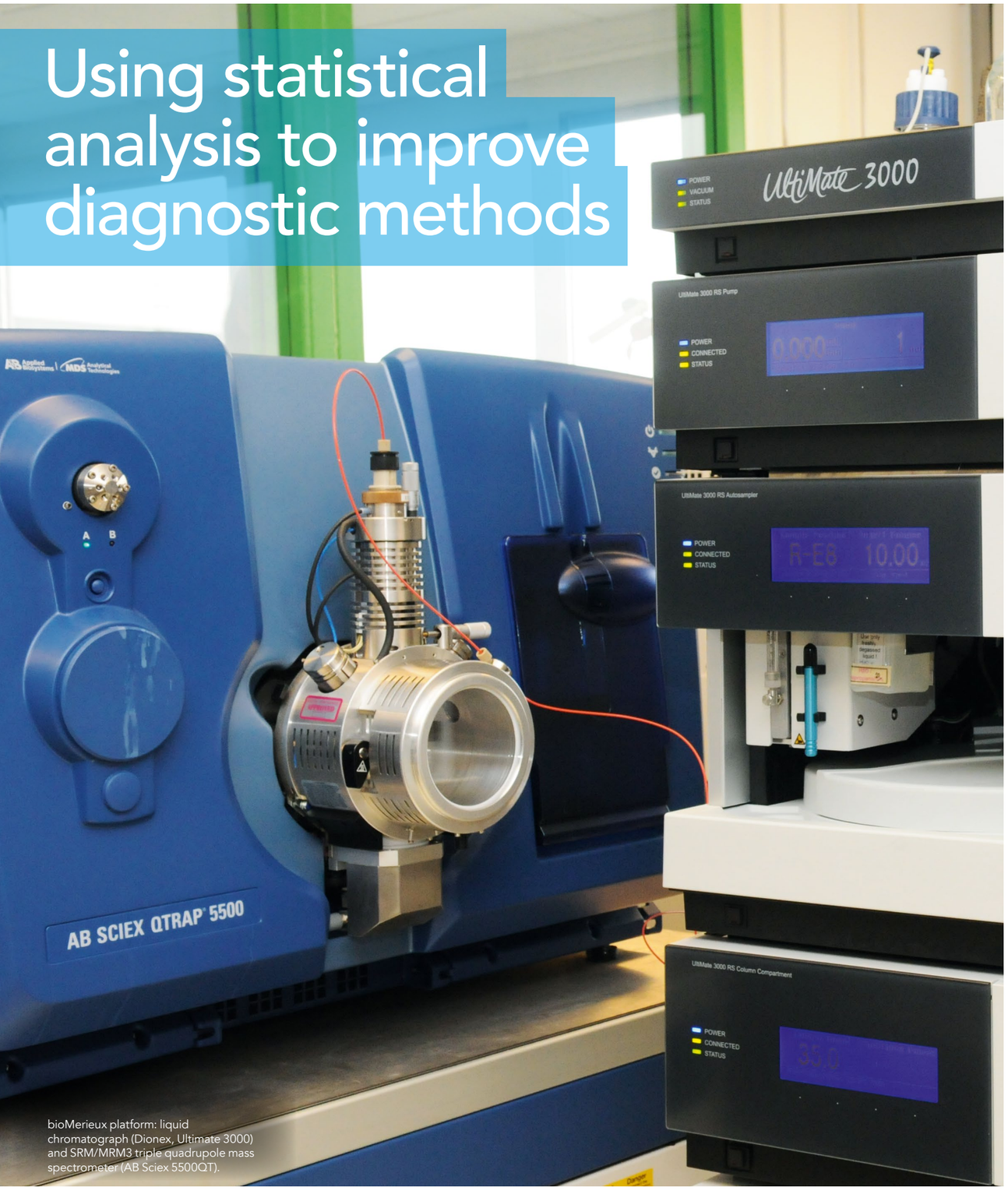
Using statistical analysis to improve diagnostic methods



bioMerieux platform: liquid chromatograph (Dionex, Ultimate 3000) and SRM/MRM3 triple quadrupole mass spectrometer (AB Sciex 5500QT).

Pierre Grangeat is Director of Research at Leti, a technological research institute at CEA Tech, the French Alternative Energies and Atomic Energy Commission. He has coordinated the BHI-PRO project, dedicated to investigating statistical methods to analyse the data obtained from mass spectrometry of biological samples. The results of the project show that this statistical analysis can improve the diagnosis and detection of pathological conditions.

In the past half century, a remarkable but silent revolution has taken place in the field of diagnostics. A physician's art in diagnosing a condition is no longer limited to visual or aural inspections, with powerful tools now available that allow inspection up to the molecular level. One such tool, the mass spectrometer, promises to be particularly disruptive, as it can provide a molecular profile of a biological sample – the analysis of which lies at the heart of modern diagnostic methods.

PROTEOMICS: AN INTRODUCTION

Complex molecular chains, called proteins, are produced by cells in our bodies during normal functioning. The composition of proteins and the concentrations in which they occur tend to vary depending on our physiological state. By studying the profiles of these proteins in blood or urine samples, it could be possible to infer the presence or absence of a pathological condition. This is the main focus of clinical proteomics.

Mass spectrometry plays a key role in clinical proteomics, and the efficacy of both methods can be measured through a simple, yet common, experiment: the blood test.

PROTEINS AND PATHOLOGY

Depending on their role in the body, cells produce many different proteins in varying concentrations. So, the first question to ask is: which proteins indicate a pathological condition? Each condition typically alters the concentrations of several proteins. These changes constitute a biomarker, which forms a signature profile for the condition. Once this has been identified, a procedure is then required to classify the sample based on the biomarker it contains. However, this is not as straightforward as you might think, with many challenges needing to be overcome in doing so.

A simple idea might be to detect changes in concentrations of individual proteins, and to combine the outcomes. However, changes in one protein may correlate to changes in another protein, causing a duplication of information. Ideally, techniques used to identify the biomarker should take such correlations into account.

Protein concentrations exhibit random variations between individuals with the same pathological condition – a process referred to as biological variability. This leads us to the second question: which protein concentrations indicate the condition? The answer to this classification problem – to decide if a sample corresponds to a healthy or pathological status – must be a statistical procedure. Automated procedures are preferable here to tests that require human intervention.

Mass spectrometry provides peptide fragment concentration measurements for each protein, along with small variations from the true values. These variations arise due to the functioning of the mass spectrometer which has to fragment the peptides, and from the biochemical preparation process required to isolate proteins and to cut the protein into peptides. This causes 'noise' – a technical variability in the measured values. Thus the proteins are decomposed into smaller molecular chains, the peptide fragments, whose concentrations are measured. As such, the correspondence between mass spectrometry measurements and protein concentrations are not one-to-one, but given by a hierarchical graph. The quantification algorithm the team has developed on the BHI-PRO project takes into account this graph structure and also includes an estimation of the unknown parameters that describe the technical variability on each branch of the graph.

STATISTICAL ANALYSIS FOR IDENTIFICATION

On the BHI-PRO project, Bayesian methods were developed to address the biomarker identification, quantification and classification problems.

In the identification problem, the measured set of proteins can be divided into two groups: discriminant and non-discriminant. Proteins from the former group help us discriminate between the presence and absence of a condition due to a change in their concentrations. First, a complete set of candidate biomarkers is identified and drawn up, by considering all possible ways to group the proteins. Then, using the Bayesian method, it becomes possible to identify the grouping that best explains the measured data. On the BHI-PRO project an analytical

By looking at measured data, can we compute the protein concentrations in the biomarkers and identify the pathological condition that gave rise to this?

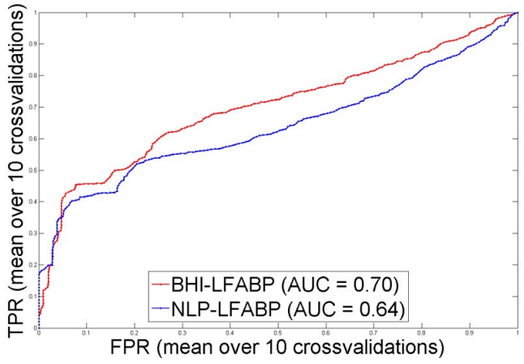


Figure 1 presents the receiver operating characterisation curves linking the True Positive Rate (TPR) to the false positive Rate (FPR) of the sample classification algorithm for this LFABP protein. The curve for the BHI-PRO quantification and a classification algorithm based on Quadratic Discriminant Analysis (QDA) is shown in red. The curve for the reference algorithm is blue. The BHI-PRO algorithm gives equivalent or slightly better sensibility (TPR) performances for a given specificity or FPR than if applied after the reference NLP algorithm.

expression has been proposed to score each protein set and to select the optimum one.

STATISTICAL ANALYSIS FOR QUANTIFICATION AND CLASSIFICATION

A Bayesian Hierarchical Inversion method has been proposed by Dr Grangeat, and his team, in collaboration with Prof Giovannelli from Bordeaux University, to address the quantification and classification problems. This method seeks to invert the problem: by looking at measured data, can we compute the protein concentrations and identify the pathological condition that gave rise to this?

The method comprises two steps, and requires two separate sets of measurements. During the first step, the training stage, protein concentrations in the biomarkers are estimated from labelled samples in the first set of measurements. These samples are obtained from individuals who are either known to have the condition or be free of it, and exhibit both biological and technical variability. The Bayesian method provides a combined estimation of the numerous protein concentrations in the biomarkers, while averaging out the sources of variability in the measured data.

In the second stage, the classification stage, Bayesian statistics use the estimated protein concentrations in the second set of measurements to classify unlabelled samples. These samples are obtained from individuals who are yet to be diagnosed, with the Bayesian Hierarchical Inversion process able to automatically quantify the protein profile and to estimate the class each sample belongs to in order to perform the diagnosis.

However, a key challenge with using Bayesian methods lies in computing the protein concentration using statistical averages from a set of peptide fragment signals delivered by the mass spectrometer. The team tackles this problem by using empirical averages of repeated draws from a simulation of the probability distribution of the protein concentrations. This is similar to estimating the average temperature of London in summer from a sequence of measurements with a thermometer influenced by sensor parameters variation to get a robust approximation of the average temperature.

CLIPP platform: MALDI TOF mass spectrometer (Brucker, Xtreme).

STATISTICAL ANALYSIS OF TECHNICAL VARIABILITY

A method has been proposed by Professor Roy from Lyon University and his team, in collaboration with Dr Grangeat, to address the following technical variability question: what are the performances on the estimation of unknown quantities of proteins when high technical variability is influencing the measurement? A biostatistical protocol including an experimental design and a model-based variance decomposition were developed to quantify the technical variability of the measurements and to separate it from the biological variability. The performances on the estimation of the protein concentration delivered by the BHI-PRO algorithm have been compared to the one using the non-linear algorithm already applied in the current practice (NLP) (Table 1). This illustrates that the quantification algorithm strongly influences the technical variability of the estimated quantities. Reducing the technical variability allows a better capture of biological variability and to discriminate groups more powerfully.

BAYESIAN INFLUENCE

The BHI-PRO algorithms were applied to a set of 206 samples drawn from the same cohort of individuals tested for colorectal cancer. These samples were almost equally drawn from non-malignant and malignant individuals with varying stages of the tumour. For each sample, 21 proteins were quantified from the SRM mass spectrometry measurement using the Bayesian Hierarchical Inversion method. It was demonstrated that it outperforms the current non-linear algorithm (NLP) which selects the best multiple reaction monitoring (MRM) transition for quantification, without requiring any form of human intervention to analyse the mass spectrometry data (Figure 1).

Quantification algorithm	Biological variability %	Technical variability %
NLP	27.2	70.6
BHI-PRO	54.8	36.2

Table 1: Percentages of biological and technical variability on the estimation of the concentration of LFABP protein from the MRM mass spectrometry measurements. Their sum is not 100 because of the shared interaction and the model error. The Bayesian Hierarchical Inversion quantification algorithm (BHI-PRO) gives better quantification performances than the Non-Linear reference algorithm currently used (NLP) reducing the part of technical variability by nearly two.

Then, by applying BHI-PRO identification method on these protein profiles, the discriminant protein was correctly identified within one hour. This highlights the influence of BHI-PRO work: armed with only a small set of labelled samples, biomarkers can now be identified much faster than before.

Statistical methods, as featured within the Bayesian techniques of the BHI-PRO research project, are a fantastic addition to a diagnostician's toolkit. These techniques improve the accuracy of biomarker identification and quantification methods, and make mass spectrometry on the whole, a much more reliable process. This might contribute to the use of mass spectrometry for clinical diagnosis.

Armed with only a small set of labelled samples, biomarkers can now be identified and quantified much faster and more reliably than before

Q&A

How did you get interested in proteomics while working in the energy commission?

The division on microTechnologies for Biology and Healthcare of Leti is developing technologies for lab-on-chip. One important research topic was the development of lab-on-chip devices on silicon for Liquid Chromatography associated with mass spectrometry, and for sample preparation to extract targeted molecules such as proteins from raw samples. Also, at the life research institute of the Atomic Energy Commission in Grenoble now called BIG (Biosciences and Biotechnology Institute of Grenoble), proteomics is a main research topic. Signal processing is mandatory to analyse mass spectrometry measurement. Thus, I started with my background in image reconstruction applied to tomographic devices to investigate new methods to reconstruct protein profiles.

How did you identify statistical tools as a potential solution for this problem?

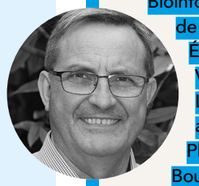
Proteins of interest are present in very low concentrations. Typically, the order of magnitude of the ratio between the targeted proteins and the total content is in the order of 1 per 100 million or more. Thus, there is a large variability in the analytical process. So, statistical tools are very relevant to describe the uncertainty and the variability both on the concentration of the proteins within the sample and the interaction of those proteins with the analytical chain. Also, mass spectrometry analytical chains are complex processes, starting from protein level, going to peptide level, and then to fragment level. Hierarchical statistical models are appropriate to describe such multilevel interactions.

Are there any other potential applications for these methods in proteomics?

Clinical proteomics is one research topic for proteomics. But fields such as life science or pharmaceutical research are also of major interest for proteomics application. The statistical methods we have investigated could be generalised to all the proteomics analytical devices such as the immunological recognition (ELISA test) or the protein bioarrays. The main difference between genes and proteins from the point of view of the analytical process is that there is no efficient way to duplicate proteins whereas PCR can duplicate genes efficiently. Thus, proteomics analytical process will always be linked to small signal levels, requiring statistical tools for data analysis.

What are some immediate next steps you have in mind for this work?

The next step might be the integration of the protein quantification software we have developed within an automated mass spectrometry analytical chain, or the integration of the protein selection software within protein analytical software libraries. For MALDI-ToF users, we have developed a software for simultaneous spectrum deconvolution and baseline removal. The biostatistical tools and methodologies for comparing the performances of analytical software and analytical chains are also of general interest for the scientific community.



Can you see any other applications for these techniques within your organisation?

Each partner within the BHI-PRO consortium is considering the integration of the know-how developed on the BHI-PRO project within its current research or developments. Typically, this will include the application of those statistical tools to other application fields such as microorganisms recognition using mass spectrometry, the study of pollutants in the environment (air, water, …), breath gas analysis, statistical analysis for genomics, and more generally biostatistics, statistical signal and image processing.

Fields such as life science or pharmaceutical research are also of major interest for proteomics application

Detail

RESEARCH OBJECTIVES

The BHI-PRO project focuses on applying statistical techniques to mass spectrometry in order to make data analysis more straightforward. More specifically, this research is dedicated to the discovery and validation of new protein biomarkers.

FUNDING

Agence Nationale de la Recherche (ANR 2010 BLAN 031301); bioMérieux; CEA

COLLABORATORS

Université Grenoble Alpes, Grenoble, France; Electronic and Systems for Health Laboratory in CEA-Leti, Grenoble, France; Data Analysis and Systems Intelligence Laboratory in CEA-List, Saclay, France; Université Paris-Saclay, Saclay, France; bioMérieux, Marcy l'Etoile and Grenoble, France; Université de Bordeaux, Laboratoire de l'Intégration du Matériau au Système (IMS), Talence, France; Service de Biostatistique-Bioinformatique des Hospices Civils de Lyon, CNRS UMR 5558, LBBE, Équipe Biostatistique Santé, Villeurbanne, Université de Lyon, Lyon, France; CLinical and Innovation Proteomic Platform (CLIPP), Université de Bourgogne, Dijon, France

BIO

Pierre Grangeat has been Research Director at CEA LETI since 2002. His fields of interests are in signal processing, inverse problems, source separation for biomedical devices including biomolecular fluids analysis (blood, air, water), and digital health. From 2011 to 2013, he was the coordinator of the ANR BHI-PRO project.

CONTACT

Dr Pierre Grangeat
CEA, Leti, MINATEC Campus
17, avenue des Martyrs 38054
GRENOBLE, cedex 9
France

E: pierre.grangeat@cea.fr
T: +33 4 3878 4373
W: http://www.leti-cea.com/cea-tech/leti/english