Using statistical analysis to improve diagnostic methods

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—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 discriminate between the presence and absence of a condition due to a change in their concentrations. The answer to this classification problem—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.

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

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.
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 for 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 target proteins and the background signal 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 interactions 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 proteins in proteomics? Clinical proteomics is one research topic for such proteins. 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 biochips). 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 workflow libraries. For MALDI-ToF users, we have developed a software for simultaneous spectrum deconvolution and baseline removal. The biostatistical tools and methods that we have developed in 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 tools within your organisation? Each partner within the BHI-PRO consortium is considering the integration of the know-how developed in 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 bio-statistics, statistical signal and image processing.

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