Protein mass spectra data analysis for clinical biomarker discovery: a global review

Pascal Roy, Caroline Truntzer, Delphine Maucort-Boulch, Thomas Jouve and Nicolas Molinari

Submitted: 12th February 2010; Received (in revised form): 9th May 2010

Abstract
The identification of new diagnostic or prognostic biomarkers is one of the main aims of clinical cancer research. In recent years there has been a growing interest in using high throughput technologies for the detection of such biomarkers. In particular, mass spectrometry appears as an exciting tool with great potential. However, to extract any benefit from the massive potential of clinical proteomic studies, appropriate methods, improvement and validation are required. To better understand the key statistical points involved with such studies, this review presents the main data analysis steps of protein mass spectra data analysis, from the pre-processing of the data to the identification and validation of biomarkers.

Keywords: clinical proteomics; statistics; pre-processing; biomarker identification; validation

INTRODUCTION
The identification of new diagnostic or prognostic biomarkers is one of the main aims of clinical research. A biomarker is defined as a biological constituent whose value differs between groups. Depending on the study design, these groups can be either diagnostic (diseased or healthy subjects) or prognostic (relapse or event free cases) groups. Currently, clinical research is making use of new high throughput technologies, like transcriptomics or proteomics. Proteomics is a rapidly developing field among other omics disciplines, focusing on large biological datasets. While transcriptomics focuses on RNA level data, proteomics is concerned with the next biological level in the universal dogma of genetics, namely proteins. Proteins are the actual effectors of biological functions and the measure of their expression level is in direct connection with their activity. Proteins are less influenced by down-and up-regulation than RNA and might offer a better proxy for biological activity. Furthermore, proteins are exported out of the cell and can be detected in various biological fluids that are easily sampled, like blood. Nevertheless, it should also be pointed out that protein expression levels and activities are not exactly correlated.

Mass spectrometry (MS) is a technology recently used for the separation and large-scale detection of proteins present in a complex biological mixture. This technology is increasingly being used in proteomic clinical research for the identification of new biomarkers and offers an interesting insight for...
biological samples containing large numbers of proteins such as plasma or tumor extracts. Initially, surface enhanced laser desorption/ionization time-of-flight (SELDI-TOF) and thereafter matrix assisted laser desorption and ionization time-of-flight (MALDI-TOF) are the most commonly used MS instruments to achieve these clinical objectives.

MS analysis relies on protein identification by their time-of-flight (TOF), which is related to their mass \( m \) to charge \( z \) ratio, usually written \( m/z \). MS output is a mass spectrum, associating TOF with signal intensities. These intensities are related to the concentrations of proteins in the processed sample. The MS signal resulting from SELDI-TOF or MALDI-TOF measurements is contaminated by different sources of technical variations that can be removed by a prior pre-processing step. Whether this method is quantitative is still questioned by some authors [1]. This aspect deserves further investigations before judging the potential of MS for quantitative analysis [2, 3]. MS nevertheless appears as an exciting tool with great potential [4]. Despite numerous studies stating that MS methods are a powerful approach for detecting diseases in medicine and biology, they still require appropriate methods, improvement and validation [5–7].

This review presents the data analysis steps of MS experiments in the context of high throughput identification studies. Section 2 presents the specificity of high throughput proteomics and the pre-processing steps. After the signal is thus cleaned, Section 3 deals with biomarker identification and validation. A discussion is proposed in Section 4.

**PRE-PROCESSING STEPS**

Data acquisition leads to various sources of experimental noise. As a consequence, the observed signal \( I \) for sample \( i \) can be decomposed into several components described in the following equation: the biological signal of interest \( S \) weighted by a factor of normalization \( k \), a baseline \( B \) corresponding to systematic artifacts usually attributed to clusters of ionized matrix molecules hitting the detector or detector overload, and random noise from the electronic measuring system or of chemical origin \( N \):

\[
I_i(t) = k.S_i(t) + B_i(t) + N_i(t),
\]

where \( t \) refers to TOF values (related with \( m/z \)). The aim of pre-processing steps is to isolate the true signal \( S \). The order in which these steps should be performed still remains an open question and different positions have been taken by different authors. The order of pretreatment steps and interactions between methods were investigated by Arneberg et al. [8] through an original modelling approach. The interested reader is referred to this article for more details. A summary of the entire analytical workflow is presented in Figure 1.

Currently, there is no standard consensus as to which algorithms to use during the pre-processing steps; this is still an active field of research and a wide variety of algorithms have been proposed. Morris et al. rapidly provided some well-established guidelines for the pre-processing of MS data [9]. More recently, Cruz et al. [10] and Yang et al. [11] have proposed an extensive comparison of some popular and current methods for pre-processing. Most of these methods are available through specific packages, as Bioconductor repository, in the open-source R-Gui software (http://www.R-project.org). A succinct description of the main and most recent methods for each of the pre-processing steps is provided below.

**Figure 1:** Workflow for MS data analysis: from raw spectra to identification of biomarkers and classification.
Noise filtering
The most commonly adopted method is the use of the wavelet transform [12, 13]. For this purpose, spectra are first decomposed into wavelet bases, usually using an undecimated discrete wavelet transform (UDWT), and in this way expressed through detail and approximation coefficients. By shrinking small detail coefficients to zero, noise is then removed from the original spectrum signal. Two possibilities are offered for this thresholding: (i) hard thresholding replaces small coefficients below a given threshold with zeroes, (ii) soft thresholding shrinks coefficients, possibly to zero. Hard thresholding might distort the signal more than soft thresholding, but it remains a simpler approach with good performance.

Du et al. [14] proposed the use of continuous wavelet transforms that simultaneously filter out noise and baseline. Later, Kwon et al. [15] proposed a novel wavelet strategy that accommodates errors that vary across the mass spectrum. Recently, Mostacci et al. [16] proposed an effective multivariate denoising method combining wavelets and principal component analysis (PCA), which takes into account common structures within the signals.

Baseline correction
Two main approaches can be described for removing the baseline: the baseline is considered either (i) as the part of the signal remaining after features of interest have been removed [6], or (ii) as some kind of smooth curve underlying the spectrum [12, 17, 18]. Malyarenko et al. [19] offered a time-series perspective on the problem. In their article, baseline arises from a constant offset and a slowly decaying charge, plus some shift after detector overload events. Another original approach was developed by Dijkstra et al. [20]. These authors set up a mixture model to deconvolve each wavelet and principal component in the mixture model corresponds to the baseline.

Alignment of spectra
Due to physical principles underlying MS, a shift on the m/z axis can appear. Alignment corresponds to an adjustment of the time of flight axes of the observed spectra so that the features of the spectra are aligned with one another. In fact, for a reliable identification of local features, one needs to carefully associate each feature of interest with a specific m/z. The easiest way to perform alignment would be to allow spectra to shift a few time-steps to the left or to the right so that their correlation with other spectra is maximized. Nevertheless, this method may be too simplistic as it does not take into account the differences in shift that may occur along the m/z axis. More elaborate algorithms were thus developed for stronger misalignment. Based on a set of peaks in a reference spectrum, Jeffries et al. [21] proposed stretching each spectrum so that the distance between peaks from the reference and from other spectrum are minimized. The same year, Wong et al. [22] developed a strategy that also aligns spectra using selected local features (usually peaks from the average spectrum) as anchors. In the proposed method, spectra are locally shifted by inserting or deleting some points on the m/z axis. In 2006, Pratapa et al. [23] have proposed an original method for alignment of repetitions of mass spectra, i.e. multiple spectra for the same sample. All spectra are thought of as variations of one and the same latent spectrum, which is inferred by a Hidden Markov Model (HMM). Later, Antoniadis et al. [12] proposed an alignment based on landmarks in the wavelet framework. More recently, Kong et al. [24] proposed a Bayesian approach that uses the expectation–maximization algorithm to find the posterior mode of the set of alignment functions and the mean spectrum for a population of patient, and Feng et al. [25] modelling the m/z by an integrated Markov chain shifting (IMS) method.

Normalization
A normalization step is necessary to ensure comparable spectra on the intensity scale. The idea is to consider that the total amount of protein is roughly the same between samples. In other terms, a small proportion of proteins are differentially expressed among all proteins in the sample. Total ion current (TIC) is a useful proxy for the total amount of protein. It is related to the total number of ion collisions with the detector and output by the mass spectrometer. Each of the intensities in the spectrum is divided by the TIC. This method is debatable [26], and more robust but less intuitive approaches are being studied. A good comparison study of these methods is available [27].

Peak detection
Once the true signal is identified, peak detection aims to identify locations in the m/z range that
corresponds to peptides and to quantify corresponding intensities. Several approaches were proposed for this purpose [6], with only the main strategies being described here.

The most intuitive strategy, initially developed by Yasui et al. [28], relies on local maxima, that is to say a point with higher intensity than other points in a neighbourhood. Once identified, these maxima may be filtered so as to keep only peaks that emerge from the noise and (ideally) really correspond to a peptide. For example, when one or all of the following filtering criteria exceed a user defined threshold, this may be used to help define the existence of a local maxima as a meaningful peak: signal-to-noise ratio, intensity and area [29]. Other authors have proposed to move the peak finding problem to the wavelet coefficient space; the goal here is to find high coefficients that cluster together on a similar position for different scales and that thus corresponds to peaks [12, 14, 30]. Another interesting strategy that has been proposed is to use model-based criterion that propose model functions to fit peaks [31].

The above methods subsequently require peak alignment to decide which peaks in different samples correspond to the same biological input [29]. Moreover, only peaks that appear in a large enough proportion of collected samples will be considered in the further analysis, leading to the choice of an arbitrary threshold proportion. To avoid such arbitrary decision making, Morris et al. [9], proposed to identify peaks on the mean spectrum. This has additional advantages: it filters out some instrument noise, plus some noisy peaks that appear in too few spectra.

Finding features of interest

Using peaks as features of interest requires estimating the intensities (as a function of the number of molecules hitting the detector) associated to these peaks. Peaks are actually two-dimensional (2D) features with a height and a width. The simpler approach is to use peak heights as intensities. However, as can easily be verified on a spectrum, peaks are narrow for low m/z but get broader with increasing m/z values. Computing the area under the peak is therefore another proxy for intensity. For a given peak, this is not necessarily a complicated step, but it requires evaluating the peak width which requires some detail. For close or overlapping peaks (on the m/z axis) corresponding to different peptides, estimating the intensity can be a really hard step. If such overlapping peaks are visible in the spectrum, several relatively recent methods [20, 32] using distribution mixtures offer the possibility to deconvolve features, thus enhancing the resolution for each peak and enabling a better intensity reading [33].

At this step, each spectrum is characterized by a finite, common number of peaks. Using adapted statistical methods, biomarker discovery analysis then aims to detect which of these peaks are associated with the factors of interest. Another approach considers spectra as functional data and analyses wavelet coefficients rather than detected peaks [9, 12, 18]. Although promising, this later approach seems little used and merits further investigation.

BIOMARKER IDENTIFICATION AND VALIDATION

Biological constituents associated with either diagnostic or prognostic groups are tested in diagnostic or prognostic studies, respectively. These studies correspond to the first phases of biomarker development formalized by Pepe et al. [34] in the context of early detection of cancer. Guidelines have also been proposed to encourage transparent and complete reporting in the context of prognostic studies to better understand conclusion contribution to scientific knowledge [35].

Generally, value distributions of constituents overlap between the groups. In classical clinical studies, one or few biological constituents are tested. These candidate biomarkers correspond to a priori hypotheses issued from a biological pathway. Recent developments in molecular biology have led to high throughput technologies, and consequently the question of how best to adequately design studies. A sequential approach has been adopted to discover new biomarkers. The first step corresponds to identification studies that aim to select a list of candidate biomarkers among a large number of biological constituents, and to estimate the strength of association between those candidate biomarkers and disease status or outcome. The second step corresponds to validation studies designed to retain, among previously selected candidates, the confirmed biomarkers, and to re-estimate the strength of association between those biomarkers and disease status or outcome. Classification refers to the assignment of each spectrum to a group, i.e. for each sample or equivalent individual. Given a serum sample, the aim of classification is to allocate
the patient to a specific diagnostic or prognostic group. Though tightly related to biomarker discoveries, profiling, similar to the signature concept in transcriptomics, has appeared as a new strategy for classification mass-spectra. The idea is to use a protein profile, defined as a list of intensities for different m/z positions, instead of a unique protein concentration. Such a profile can be used in two different ways: (i) as a set of proteins that must (or could later) be individually identified, and (ii) as a set of distinguishing features without searching for biological support for these. This can be an efficient approach in clinical applications.

Identification studies
The aim of identification studies is the selection of candidate biomarkers from a large number of biological constituents. In the proteomic context, a biomarker can be defined as a protein that differentiates samples from different groups. Several methods have been used or developed in order to identify candidate biomarkers. These methods have been discussed in the transcriptomic field. This search for biomarkers is tightly linked with the well-known curse of dimensionality that exists in all high-throughput methods, where the number of variables (peak intensities) exceeds the number of samples. Intuitively, the dimension of the regression space must be reduced, and two approaches have been proposed to do this. In the first approach, the subspace is defined by selecting certain variables within the matrix corresponding to features of interest, whereas in the second case, the subspace is defined by components. While the former of these methods directly focuses on the identification of features that exhibit differences in intensities between the groups, the latter evaluates the importance of features embedded in a classification algorithm.

Feature selection
When two groups are compared, simple approaches consider each feature individually. The aim is to compare intensities between the groups. Each test tells whether the two groups can be distinguished based solely on information from one particular feature. This corresponds to univariate statistical tests, such as the classical Student t-test. The observed test value is compared to the test distribution under the null hypothesis, \( H_0 \), of equality between mean expression levels in the two groups. The significance of the test is calculated from the cumulative distribution function. Since features are never considered simultaneously, correlations cannot be taken into account and information from a set of features might be redundant. Biomarker identification is performed with control of type-1 and -2 error rates adapted to the large number of statistical tests performed (Table 1). The control of false discovery rate (FDR), introduced by Benjamini and Hochberg [36]:

\[
\begin{align*}
\text{FDR} &= E(Q), \\
Q &= \begin{cases} 
V/R, & \text{if } R > 0 \\
0, & \text{if } R = 0
\end{cases},
\end{align*}
\]

\( V \) being the number of false positive tests and \( R \) the total number of rejected null hypotheses, is frequently involved in controlling type-1 error [37]. Storey and Tibshirani [38] proposed to use the \( q \) value, i.e. the expected proportion of false positives among all features as or more extreme than the observed one, as a FDR-based measure of significance. They also provided an estimation of \( \hat{\pi}_0 \), the proportion of features following the null hypothesis.

To control type-2 error, the proportion of detected genes of interest
\[
\text{Power} = \frac{E(S)}{m_1}
\]

has been proposed as a definition of power in the context of multiple testing [39, 40].

Beyond biomarker identification, clinical studies aim to estimate the strength of the association between the candidate biomarker level and disease status or disease outcome. For parametric models, variable selection results from parameter estimation. Because of the selection mechanism involved in identification studies, the strength of association is commonly over-estimated. This optimistic bias is easily understood: only variables from which the corresponding test-values exceed an \textit{a priori} defined

<table>
<thead>
<tr>
<th>Truth</th>
<th>( H_0 \text{ True} )</th>
<th>( H_0 \text{ Wrong} )</th>
<th>Conclusion</th>
<th>Accept ( H_0 )</th>
<th>Reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>V</td>
<td>( m_0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>S</td>
<td>( m_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m-R )</td>
<td>R</td>
<td>( m )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
threshold are selected. Over-estimations of the strength of association and false discoveries are both explained by this well-known regression to the mean phenomena [41]. Optimism increases when the proportion of features of interest decreases and is reduced by increasing the number of samples [42]. Whereas multivariate modelling is useful for identifying information redundancy provided by the selected features with adjusting on the confounders, it does not correct the optimism bias linked to the selection process.

The parameter estimation bias results in an over-estimation of the predictive ability of statistical models when applied to new samples. More sophisticated methods for the selection of predictors exist, jointly known as penalized regression. The idea behind these techniques is to constrain the regression coefficients. Several constraints have been proposed to shrink the parameter estimators identified as L1 or L2 penalizations [43–45]. Those penalizations result in filtering out weak predictors, and often improve prediction accuracy. The extension of the threshold gradient descent (TGD) method [46] to the survival model [47] is similar.

**Classification algorithms**

Given Y, the vector of disease status or outcome status, dimension reduction techniques aim to explain Y with information from X, the matrix of features of interest. X-based reduction methods, i.e. non supervised methods, only use information from X, while X- and Y-based methods, i.e. supervised methods, also use information from Y. A typical example of X-based methods is PCA, where components define the subspace of X that maximize the projected X-variability. The supervised X- and Y-based methods define the subspace that maximizes the projected X- and Y-covariability. Several methods can be used, and an *a priori* knowledge of that structure may guide the choice of the analysis method [48]. Partial least square regression (PLS) is an example of a dimension reduction technique that directly maximizes the components associated with Y [49]. Linear discriminant analysis (LDA) first requires dimension reduction, such as PCA (PCA – LDA) or PLS (PLS – LDA). A subset of the first components is usually sufficient to catch most of the data covariance, and the optimal number of components can be chosen by cross-validation, as proposed by Boulesteix in the case of PLS + DA.

Wu *et al.* [50] proposed a comparison of certain classification algorithms for MS data. They compared LDA and quadratic discriminant analysis (QDA) to other supervised learning algorithms: k-nearest neighbors (KNN), bagging, boosting, random forest (RF) and support vector machines (SVM). KNN uses a vote of nearest samples (in a space defined by intensities for different features) to choose a class for a new sample. Bagging, boosting and RF are methods to aggregate classifiers as classification and regression trees (CART). Hastie *et al.* [45] present in detail learning algorithms and their properties.

**Combined methods**

These methods combine feature selection and classification algorithms. Classification algorithms involve a large set of potential biomarkers in classifier building. In the combine methods, biomarkers that most help in classification are retained and used in the classifier, while the others are discarded. Classification and regression trees are an example, as well as their combination with bagging, boosting or random-forest. Sparse PLS combines PLS regression and variable selection into a one-step procedure [51]. Reynes *et al.* [52] used a genetic algorithm (GA) to select feature that contributes to a voting rule, whereas Koomen *et al.* [53] combined a GA to select features that contribute to large Mahanalobis distance between groups with FDR controlled Student’s *t*-test.

**Non-peak-based classification**

Some classification methods were developed that do not rely on the peak concept. Instead, these methods use every single point of the spectra to look for regions that can distinguish different groups of samples. Indeed, the search for part of the spectrum containing useful information can be performed under the guidance of variables related to clinical events. The pioneering study by Petricoin *et al.* [54] is a typical example of this strategy using every point of the spectrum as a potential predictor. They developed an algorithm that combined a genetic algorithm with self-organizing map to select a few points on the *m/z* axis as the best set of group predictors. Li *et al.* [55] combined GA for feature selection and KNN to perform classification in a restricted subspace. Tong *et al.* [56] used the same initial idea. They consider each point of a spectrum as a potential feature of interest, in the same way as Petricoin *et al.*, but select *m/z* positions using CART associated in a
decision forest (a specific tree aggregator with constraints on tree heterogeneity and quality).

**Validation studies**

Data splitting of the study sample into a learning set and a test set has been proposed as a process of internal validation. The classifier is trained on the learning set and later validated on the test set. Such an internal validation is a required evaluation procedure to avoid over-fitting, i.e. avoid the use of characteristics of the learning set that are not of interest but rather specific to this particular set. The use of one particular learning set on the $2^n$ possible splits, and the power limitation of the procedure result in a preference for the leave-K-out or bootstrap methods [57–60]. When the statistical analysis of the identification study first requires a selection of differential biological constituents and then a classification algorithm, the test set must be excluded from all pre-processing and biomarker identification steps. Hilario et al. [61] drew attention to the potential misemployment of cross-validation techniques. They suggest not performing the pre-processing steps on the full dataset before a split between training and test set is decided. This can result in artificially over-estimated biomarker performances. Even if an internal validation has been performed, external validation is a necessary step of biomarker validation. Pepe et al. [62] underlined the difference between the strength of association between a biomarker level and disease status or outcome estimated by the odds-ratio, and its ability to classify subjects. They proposed to use the receiver operating characteristic (ROC) curves to estimate the classification performance of a biomarker or its incremental value in addition to the usual diagnostic or prognostic clinical and biological factors. This approach has been extended to the analysis of positive and negative predictive values using predictive receiver operating characteristic (PROC) curves [63].

**Statistical power**

The power of transcriptomic studies has been shown with increasing the number of non-differentially expressed genes considered in the study [39]. This theoretical result applies to proteomics although the fewer tests performed in proteomics lead to smaller power decrease. In addition to power calculations developed in the context of transcriptomic analysis [39, 40], Jouve et al. [64] have underlined that the high instrumental variability encountered in MS, together with FDR control, builds a detrimental synergy leading to a low statistical power for usual MS study sample sizes. Larger identification studies and improvement of measurement variability are both needed.

**DISCUSSION**

Proteomic technologies have been recently adapted to the new field of clinical proteomics. Our article deals with the different steps involved in proteomic data analysis. These methods assume that the biological sample is of high quality. It is therefore quite useful to add a preanalytical step concerning sample preparation. The pivotal article by Petricoin et al. [54] emphasized the impact of preanalytical factors in proteomic studies and the predominant role of good study design towards better reproducibility and hence, biologically meaningful results [5]. The various origins of errors and biases were well identified in the preanalytical and analytical steps, leading to a good discussion concerning the conditions necessary to avoid them (temperature handling, time delay or freeze/thaw cycle). The quality of the pre-analytical steps, as well as the amount of samples available, is crucial, regardless of the technology involved later on.

An inadequate quality of samples will impact the fractionation steps, such as major protein depletion, that allows the investigation of low-concentration constituents and MS analysis. Albrethsen [65] made an interesting review about reproducibility in several protein profiling studies using MALDI-TOF instruments and highlighted the several sources of technical variation. Villanueva et al. [66] and Callesen et al. [67] showed that using an optimized serum sample preparation method allows overcoming the challenge of reproducibility. It was also demonstrated that high throughput MS may be a promising biomarker discovery tool as long as analytical sources of variation are identified and controlled [68].

In addition to these sample quality considerations that could benefit in the future from good quality control, experiments must be designed so as to avoid potential sources of bias and confounding factors; inadequate study designs may lead to a confounding effect between technical factors and the biological factor of interest. Two strategies have to be employed to this end: blocking and randomizing. The first one allows avoiding systematic bias due to the block effect to balance the samples within each block
(plate) with respect to the study factors of interest. The second one allows avoiding unknown or unexpected bias like position on the plate or tip differences. Addition of technical replicates to biological replicates in study designs have recently been discussed [69]. Technical replicates allow controlling technical variability, and knowledge about biological and technical variability can then be used for sample size calculations [70].

Today, cancer clinicians would like to combine proteomic profiles and classical clinical variables in the same models to improve assessment of cancer prognosis. The first intuitive way to do this is to involve classical clinical and ‘omic’ variables in the same models [71–73]. However, this approach is not satisfying since it does not take into account that the effects of ‘omic’ variables are overestimated with regard to the classical clinical variables [42, 74].

In the very recent literature, certain authors have proposed models that take this optimism into account. These models simultaneously construct a classifier combining both types of variables and determine whether ‘omic’ data represent additional predictive value. In 2008, Binder and Schumacher [75] proposed an offset-based boosting approach in the context of survival data. Shortly later, Boulesteix et al. [76] and LeCao et al. [77] proposed a more general approach, in the sense that it was not limited to survival data. This approach is based on PLS, RF and the pre-validation strategy suggested by Tibshirani and Efron [74]. In a more recent article, Boulesteix et al. [78] proposed a permutation-based procedure to test the additional predictive value of high-dimensional molecular data. In addition, Le Cao et al. [79] recently proposed a joint analysis of multiple datasets from different technology platforms. More precisely, they proposed a regularized version of canonical correlation analysis to analyse correlations between two data sets, and a sparse version of PLS regression that includes simultaneous variable selection in the two types of data sets.

In this work, we decided to focus on the analysis of proteomic data from MS, which provides an efficient tool for discovering new biomarkers in a large amount of samples. One potential drawback of MS is that the data generated are not directly quantitative. In fact, samples analysed with MS are highly complex, thus limiting the use of MS as a quantification tool. Once selected and statistically validated, candidate markers have to be identified and quantitatively validated. Tandem MS (MS/MS) completes this type of work. This technique involves multiple steps of MS selection in which each of the peptides of interest (candidate markers) are isolated and fragmented. A sequence search engine—like MASCOT for example—then tries to match these observed spectra with known peptides whose sequences are stored in dedicated databases. This label-free technique also provides a direct mass spectrometric signal intensity for the identified peptides. To reduce the sample complexity and thus allow a better quantification, proteins are often first isolated through separation techniques like liquid chromatography (LC). Such label-free techniques allow a direct comparative quantification of proteins [80, 81]. These methods need heavy analytical processes and for this reason are not extensively used on large-sample datasets at the moment. In addition to this biological aspect, label-free proteomics generate an amount of data higher than those generated by MS, thus leading to additional statistical questions for which statistical analytical tools are still immature [82–84]. However, label-free proteomics sounds promising and current work is evaluating their use in large clinical proteomics studies.

Some alternative methods to standard MS for differential proteomics have appeared in the last years, like isobaric tags for relative and absolute quantification (iTRAQ) or difference gel electrophoresis (2D-DIGE). These methods aim at both selecting and quantifying candidate biomarkers. Briefly, the iTRAQ technique utilizes four isobaric amine specific tags to determine relative protein levels in up to four samples simultaneously. 2D-DIGE is a 2D gel separation technology for proteins in which the different samples to be compared are labelled with different dyes (Cy3 and Cy5), which enables signal detection at different wave length emissions. Interested readers may read the works of Wu et al. [85] who provided a comparative study of these methods. Although interesting, these two latter techniques cannot be used in large-scale studies at the moment. In fact, both techniques are limited by the low number of samples that can be analysed simultaneously. Many hundreds of samples can be simultaneously analysed with MS, while only <10 can be analysed using 2D-DIGE or iTRAQ. This issue results in a lack of reproducibility for the former, and costs (both in material and time) that are too high for well powered discovery studies for the latter. This is a drawback for biomarker discovery in large-scale studies.
Proteomics represents, in our opinion, a key field for the future of clinical research. However, improvements can be made in the methodologies involved. This improvement can only result from a tight collaboration between biostatisticians, computer scientists, biologists and clinicians.

**Key Points**
- Efficient pre-processing is an essential pre-requisite for retrieving meaningful proteomic biological information from raw spectra and reaching meaningful clinical conclusions.
- The identification and validation of pertinent biomarkers requires large, well-designed studies.
- Methodology improvement would benefit from a tight collaboration between biostatisticians, computer scientists, biologists and clinicians.

**References**


41. Truntzer C, Maucort-Boulch D, Roy P. Consequences of regression to the mean in the high dimensional setting 2010. submitted manuscript.


