Journal of Chromatography B, 877 (2009) 1250–1258

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

Review

Technical, bioinformatical and statistical aspects of liquid chromatography–mass spectrometry (LC–MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: A critical assessment☆

Mohammed Dakna[a,*], Zengyou He[b], Wei Chuan Yu[b], Harald Mischak[a], Walter Kolch[c]

[a] Mosaiques Diagnostics & Therapeutics, Hannover, Germany
[b] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
[c] The Beatson Institute for Cancer Research & Sir Henry Wellcome Functional Genomics Facility, University of Glasgow, Glasgow, UK

ARTICLE INFO

Article history:
Received 9 July 2008
Accepted 28 October 2008
Available online 6 November 2008

Keywords:
LC–MS
CE-MS
Biomarkers
Clinical proteomics
Statistical data analysis

ABSTRACT

The search for biomarkers in biological fluids that can be used for disease diagnosis and prognosis using mass spectrometry has emerged to become a state-of-the-art methodology for clinical proteomics. Poor cross platform comparison of the findings, however, makes the need for comparison studies probably as urgent as the need for new ones. It is now increasingly recognized that standardized statistical and bioinformatics approaches during data processing are of utmost importance for such comparisons. This paper reviews two of the currently most promising methods, namely LC–MS and CE-MS techniques, and software tools used to analyze the huge amount of data they generate. We further review the statistical issues of feature selection and sample classification.

Contents

1. Introduction 1250
2. LC–MS 1251
3. CE-MS 1252
4. LC–MS versus CE-MS 1252
5. Bioinformatics of LC–MS and CE-MS data 1253
   5.1. The mzXML file format: storage 1254
   5.2. Software tools for biomarker definition 1254
6. Sample classification using LC–MS and CE-MS data 1254
   6.1. Feature selection strategies 1255
   6.2. Classification 1256
   6.3. Cross-validation 1256
7. Conclusion 1256
   References 1257

1. Introduction

Over the past two decades many proteomics technologies evolved. On one hand, these different methodologies fueled the progress of this field. On the other hand, this progress has led to the lack of comparability of proteomics research findings. These intriguing problems have received great attention by researchers in clinical proteomics. The first argument for explaining this great variability is to attribute it to differences in study designs where patient sampling may include unmatched important confounding factors such as age or gender. A second argument may be the difference among the technologies that are currently in use as well as the difference among software and data analysis platforms. Hence, the community increasingly becomes aware about the need of

☆ This paper is part of the special issue "Quantitative Analysis of Biomarkers by LC–MS/MS", J. Cummings, R.D. Unwin and T. Veenstra (Guest Editors).

* Corresponding author.
E-mail address: dakna@mosaiques-diagnostics.com (M. Dakna).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.10.048

standards that should be reported together with proteomics experiments [1,2].

The analysis of complex biological fluids requires a pipeline of methods for separation, identification and quantification of potential biomarkers [3]. One of the most common techniques is liquid chromatography–mass spectrometry (LC–MS) that generates a two-dimensional data chart. Further, capillary electrophoresis-mass spectrometry (CE-MS) has also been shown to present an attractive platform for clinical proteomics [4–8]. From a bioinformatics point of view, these two methods are similar as the structure and the amount of the data generated is quite comparable.

Here we compare the methodological and technical aspects of both methods. Particular emphasis will be given to the issue of data processing [9,10]. Several of the software solutions that have been developed over the last years are reviewed.

Once the raw data are processed and compiled, feature selection strategies and classification algorithms are used to extract clinically relevant information from the data (e.g. classify the patients in disease and control groups). Guidelines are given to ensure sound reporting of biomarkers with potential discriminatory power [11]. In particular, the multiplicity issue inherent to all high throughput proteomics is outlined. One of the major aims is to define biomarkers differently expressed between two conditions (e.g. disease versus control) while keeping low probability for false positives. Hence, issues of multiple testing and statistical power are of utmost importance for designing proteomics pilot studies to follow the strict statistical guidelines required in clinical applications [12]. We discuss one popular classification procedure: linear discriminant analysis (LDA). The LDA has been applied for almost a century in different context with great success [13]. We opt for LDA because of its simplicity and because many of the pitfalls such as ill-conditioned classification problems as well as collinearity issues of the features used in classification are easily discussed within this context.

The paper is organized as follows. We first describe and compare the technical aspects of LC–MS and CE-MS. We then address the process of raw data processing, storage and generation of biomarkers lists. Aspects of sample classification and related statistical pitfalls are discussed and followed by a conclusion.

2. LC–MS

LC–MS is a hyphenated technique, which combines the separation power of LC with the detection power of mass spectrometry. An LC–MS system consists of the following four main components:

(1) A chromatographic column: The chromatographic column uses a liquid as the mobile phase and a porous solid as the stationary phase. The mobile phase contains the peptides to be separated and moves through the stationary phase. The column separates peptides according to their separation characteristics that result in elution at different time points. In proteomics, two LC methods are commonly used: *reverse phase chromatography* (RV, separating on hydrophobicity) and *strong cation exchange chromatography* (SCX, separating on charge).
(2) An ionization source: It converts eluting peptides into ions. Two typical ionization techniques are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI).
(3) Mass analyzer: The mass analyzer separates ions based on their masses. There are mainly four commonly used types of mass analyzers in proteomics: ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron resonance (FT-ICR).
(4) Detector: It records the relative abundance of ions at different m/z locations.

Fig. 1 depicts a typical LC–MS based proteomics experiment. First, protein mixtures are isolated from biological samples and often enzymatically digested into peptides. The resulting peptides are then separated into different subsets using LC. The eluting peptides are subjected to ionization, resulting in multiple, generally protonated peptides entering the mass spec-

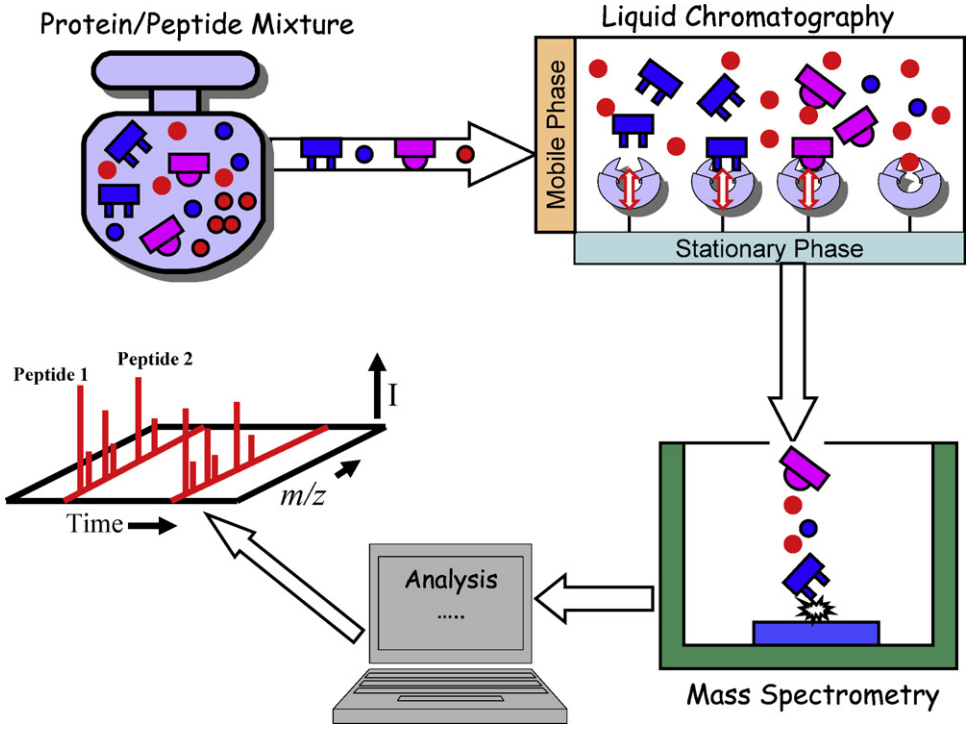

Fig. 1. Outline of LC–MS experiments.

trometer, where a mass spectrum of the peptides is recorded [3].

The output data in LC–MS experiments can be regarded as a two-dimensional image. The horizontal dimension represents the retention (or elution) time and the vertical dimension represents different m/zvalues. Since the LC–MS system generates mass spectra at discrete time points, it may be convenient to use scans instead of retention times. The mass spectrum at each single scan registers the abundance of peptide ions at different m/z locations.

The analysis of LC–MS data consists of multiple tasks, which can be categorized into three groups: low-level processing, mid-level processing and high-level processing [14].

(1) Low-level processing consists of general-purpose tasks such as baseline correction, normalization and filtering.
(2) Mid-level processing consists of peak detection, de-isotoping, charge deconvolution and alignment.
(3) High-level processing consists of feature selection and classification.

All these data analysis tasks have been studied extensively. The low-level and mid-level processing methods are reviewed in detail in [15,14]. In particular, the alignment of LC–MS data are discussed in a recent review paper in detail [16]. High-level processing methods have also been partially discussed in above reviews. A recent review [17] systematically discusses the data analysis tasks at different levels.

The moderate reproducibility of the LC separation process makes the analysis of LC–MS data challenging. Consequently, improvements in LC–MS data analysis methods as well as in software packages are required.

3. CE-MS

The combination of capillary electrophoresis with electrospray ionization mass spectrometry was applied successfully in several clinical proteomics studies [18–22]. Fig. 2 depicts the common setup of a CE-MS experiment. In principle, CE can be interfaced with any mass spectrometer. Mostly capillary zone electrophoresis (CZE) has been utilized in MS coupling, and CE is often (and also here) used synonymously for CZE. Other approaches like capillary isoelectric focusing (C-IEF) appear to be less widely used, mostly due to sophisticated technology that requires exceptional experts in running such analysis, and also due to technical limitations

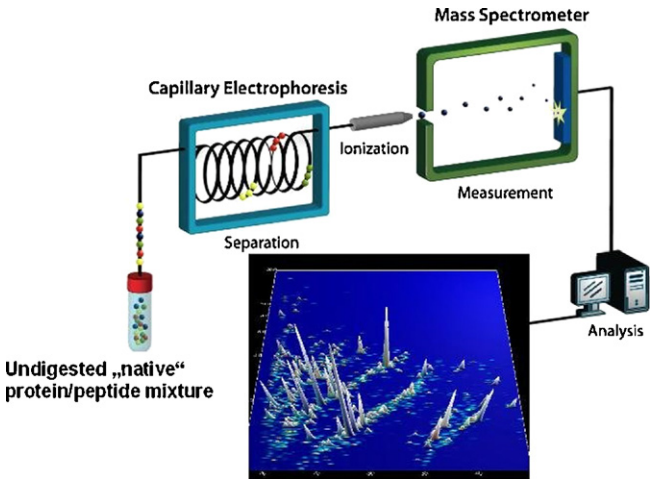

Fig. 2. Outline of CE-MS experiments.

(e.g. the problem of background ampholyte interfering with MS detection). Whereas some manuscripts indicated that proteome analysis may be possible on a large scale using C-IEF-MS [23], this initial optimism unfortunately was not yet substantiated by additional reports. The different types of CE modes that can be applied towards proteome analysis have recently been described in detail in [4,5]. As outlined in [4,5] CE-MS coupling via sheath-flow interfacing is highly stable, and also represents a sensitive detection device [24,25]. Stability and sensitivity have been demonstrated in a number of recent articles [25–29]. CE-MS with sheath-flow interfacing shows a higher detection limit due to higher flow rates. But it also shows higher stability, a major advantage when comparing large number of samples. Consequently, the majority of reports on CE-MS utilize sheath-flow interfacing [25].

A hallmark of CE-separation is the appearance of “streaks” of peptides, when migration time is plotted against mass (Fig. 3). These “streaks” are the result of the simple separation principle used. Separation is accomplished by electrical force applied onto the ion, which in turn depends on the charge and on the flow resistance, depends on the cross-section of the ion. At a low pH, the amino groups are protonated, and protons in general are the sole source of charge under these conditions. The position of each peptide in a CE-separation can, therefore, be calculated with perfect accuracy if its mass and the number of basic amino acids are known, as described in [30].

4. LC–MS versus CE-MS

As already outlined in a previous review [24], CE holds several advantages over LC, which were very recently confirmed in detail in several reviews [31,32]. These advantages are especially beneficial when analyzing a large number of heterogeneous samples that contain interfering compounds, such as lipids, precipitates, etc. Its main advantages are the robustness, ability to fast recondition NaOH, the simple separating principle with high reproducibility, and, with respect to MS interfacing, a buffer that does not change its composition due to gradient. Furthermore, CE-MS enables reproducible and comparable analysis of highly complex samples, as shown in Fig. 4 for rat urine samples [33].

A disadvantage of CE is its limited loading capacity. Whereas ml quantities can be loaded onto an LC column, a CE can be filled with a maximum of ca. 1 μl and in general only 10–100 nl. Although pH-stacking can be used very effectively, a maximum of 30–50% of the total capillary volume can be filled with sample, which corresponds to 0.5–2 μl when using 50 or 75 μm ID capillaries with 80–100 cm length. This limitation is of minor consideration in CE-MS coupling, since the concentration of analytes in the sample is generally high (with respect to the detection limit in the fmol range). Here the major limitation is due to the dynamic range of the mass spectrometer (4 orders of magnitude at best), and that more abundant peptides will obscure minor signals. Even if the detection limits were lowered significantly or more samples could be loaded, the number of detected peptides will not increase proportionally. In a large fraction of the CE-MS data space, signal is already present. Additional signals at the same or an overlapping position (due to the isotopic distribution of a peptide generally covers 3–6 mass units) cannot be detected with good confidence, as they are likely to be obscured by stronger signals that are already present, and may even result in conflicts in data interpretation.

Although CE can be interfaced with a MS/MS instrument, direct sequencing of the CE does represent a challenging task, because only limited amounts of sample can be loaded onto the capillary (see above), yielding small peaks in the MS which are problematic for subsequent MS/MS analysis. An alternative approach is the

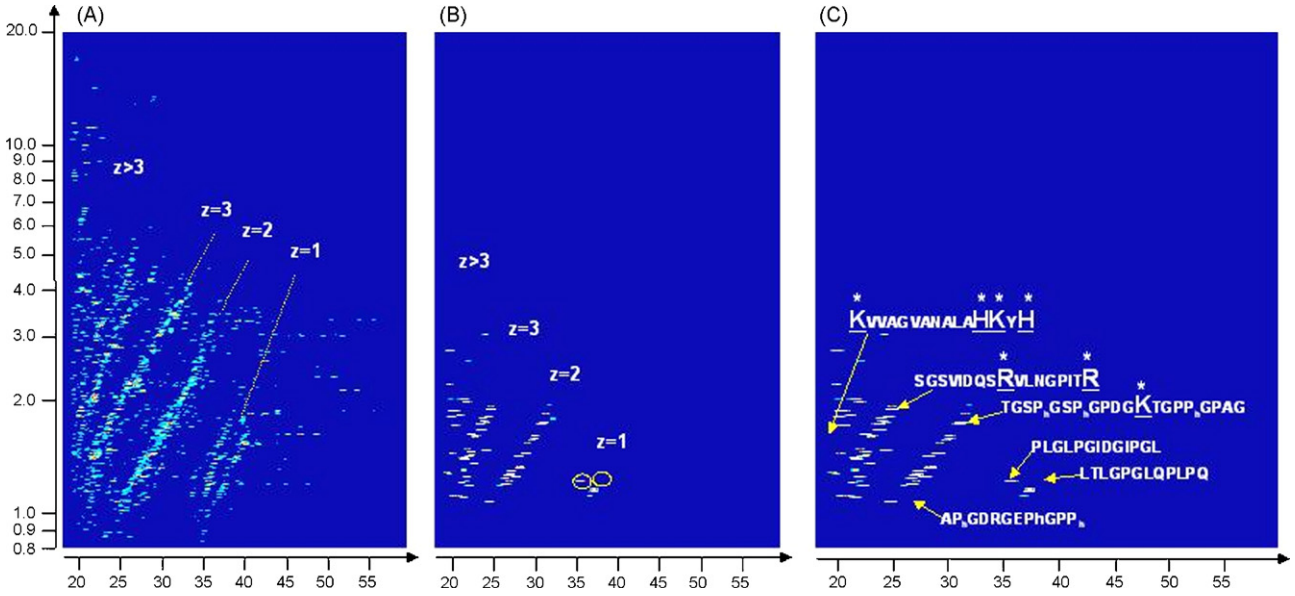

Fig. 3. Compiled CE-MS data from healthy volunteers. (A) Contour plot of the entire urine peptidome. The molecular mass (logarithmic scale) on the *y*-axis is plotted against normalized CE migration time on the *x*-axis. The arrangement of the analyzed peptides in distinct lines is obvious. (B) Contour plot of 107 identified polypeptides. The lines already observed in (A) result from the number of positive charges *x* (at pH 2.2). Peptides marked with circle: Collagen type VI alpha 4 fragment (PLGLPGIDGIPGL); 1217.702 Da; migration time: 34.03 min. Peptide marked with dashed circle: Insulin-like 3 fragment (LTLGPGLQPLPQ); 1232.713 Da; migration time: 36.19 min. (C) Correlation between the effective net charge, molecular mass, and the CE migration time for several examples of determined peptide sequences. Basic amino acids are highlighted with a star. Reprinted from [30] with permission.

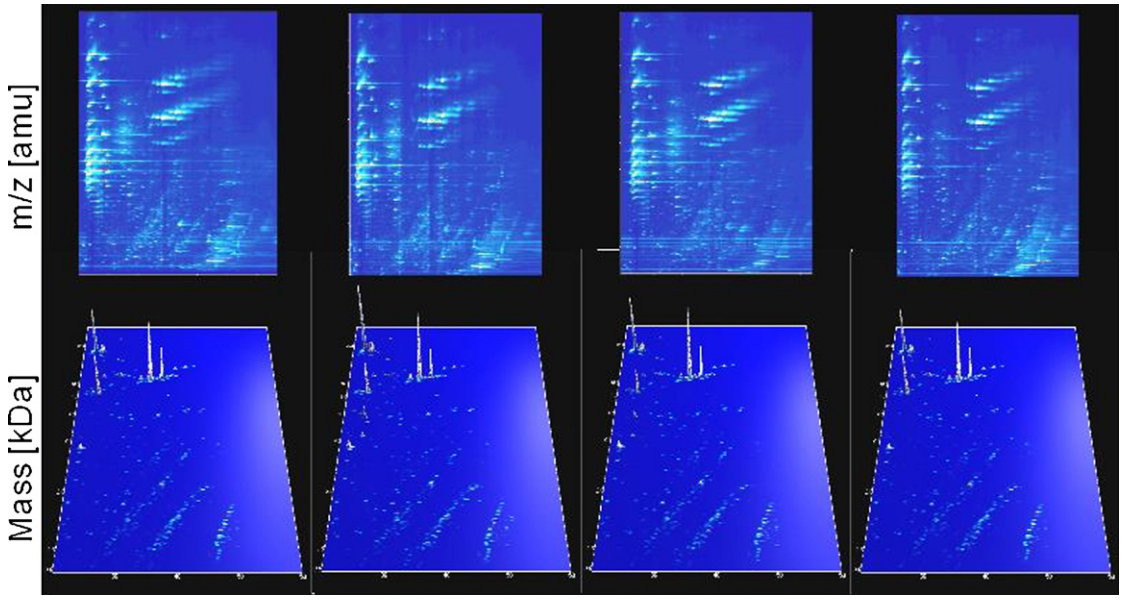

Fig. 4. Reproducibility of urinary rat polypeptides evaluation. Examples of electrophoreograms from four (of 19) consecutive measurements of a single urine sample. The m/z values of the 2-D raw data plots (upper Panel) and the molecular mass (logarithmic scale) of the deconvoluted 3-D plots (lower Panel) on the *y*-axis are plotted against CE migration time on the *x*-axis. The arrangement of the analyzed peptides in distinct lines is obvious and can be comprehended as a result of the number of positive charges at pH 2. An average of 1300 ± 106 polypeptides were detected in each one of the 19 replicates. Reprinted from [33] with permission.

interfacing of MS/MS with LC to allow loading of larger amounts of material. Because theoretical migration time can be used to calculate the exact position of a peptide in CE-MS, sequences can be rather exactly attributed to a position in the CE-MS run. This approach, in combination with highly accurate precursor-ion mass determination with CE-FT-ICR analysis, has proven very successful [34,21,22,35].

5. Bioinformatics of LC–MS and CE-MS data

The last years have seen a dramatic increase in the number of commercial and open source tools developed to assist researchers during the different stages of proteomics data processing. Here we list some of them and refer the reader to different reviews where detailed discussions of those tools are reported.

Table 1
Tools for storage and quantification of LC or CE MS and MS/MS data.

Name	Website-reference	License
TPP	http://tools.proteomecenter.org/TPP.php [38]	Open source
CPAS	http://proteomics.fhcrc.org/CPAS/Project/Demo/begin.view [39]	Open source
PEPPeR	www.broad.mit.edu/cancer/software/genepattern [40]	Custom
IPP	www.insilicos.com/IPP.html	Commercial
Scaffold	www.proteomesoftware.com	Commercial

5.1. The mzXML file format: storage

At the bottom of the raw data processing, before genuine data analysis can take place, the raw data must be read in the given application software. Since all machines store data in proprietary formats, the use of non-vendor processing tools may pose a problem. Recently, the situation has changed because the emerging XML file formats have been increasingly used for storing the raw data. Several tools are available for converting data from common instruments to mzXML, mzData or mzML format [9,10]. The reader is referred to www.sashimi.sourceforge.net and www.proteomecommons.org for more details. From a statistical point of view, it should be noted that the term "raw data" in this context is not strictly accurate, as manufacturers include some signal processing steps in the collection of such data. The different raw data output formats of mass spectrometers were a major obstacle to uniform proteomics analysis, as the generated data could not be easily managed, organized and shared among researchers at various institutions. With data in the mzXML format there is no need for vendor supplied programming interfaces or dynamically linked libraries. Furthermore, the process of development of downstream analysis software is alleviated from supporting different native formats. Usually the mzXML raw data files may be 2 or 3 times the size of the original native format. After compressing the files, however, the sizes become comparable. A major advantage of mzXML is offset indexing, enabling easy location of each spectrum within the data file. This may speed up the loading time of gigabyte data streams and can advantageously be used for storing additional information as implemented in some software packages (e.g. [36,37]).

The storage and management of LC–MS CE-MS and MS/MS data is a challenging task due to the huge sizes of the raw data files. Some software solutions assist the users in organizing the data as well as for searching proteins and peptides databases. Table 1 lists some of the most widely used storage and management pipelines.

One of the most important forces that drive the computational research in microarray genomics profiling was the availability of public raw data repositories (e.g. the Stanford Microarray Database (SDM)) where experimenters may depose their data in the public domain. The availability of such data attracted the interest of many skilled bioinformaticians and statisticians. The analysis of the available data results in a very fruitful interdisciplinary research. In the field of proteomics, the importance of public domain repositories has only recently been recognized [41]. Examples of such free raw data repositories are the PeptideAtlas www.peptideatlas.org [42] and the Open Proteomics Database http://bioinformatics.icmb.utexas.edu/OPD [41].

5.2. Software tools for biomarker definition

Different software tools are needed at different stages of the proteomic data analysis (see [43] for a review). The aim of such software pipelines is to provide lists of biomarkers that will eventually be used for further analysis (e.g. sample classification). Instrument vendors provide such integrated pipelines. However, from user perspective, most of them represent a black box as neither the source codes nor the details of the algorithms are available. Recently an overwhelming number of commercial and open source packages were developed to analyze LC–MS and CE-MS data. Some features in those packages are quite similar, many others do differ. This may result in different lists of biomarkers derived from the same dataset when using different packages. Review articles have recently summarized some software packages that are currently in use [16,43–46]. Since some of the available tools perform rather parts of the required raw data processing, several have to be combined in different ways for producing a tailored data analysis pipeline that includes at a minimum peak detection, charge deconvolution, alignment and feature detection. In Table 2 we list packages that currently appear to be among the most useful ones. However, this list cannot be considered comprehensive. A more complete list of tools that may be used during different phases of data analysis can be found at www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList.

6. Sample classification using LC–MS and CE-MS data

Data processing results in a LC–MS/CE-MS profile for each sample. The data from all samples in a study are formatted in a $n \times p$ matrix with n representing the number of samples and p denoting the number of variables. The structure of the data matrix is very reminiscent of the output of microarray genomics experiments.

Table 2
Tools for processing raw LC–MS or CE-MS data.

Package name	Website-reference	License	Functionality
MSinpect	http://proteomics.fhcrc.org/CPL/msinspect.html [36,37]	Open source	p,f,cd,i,a
Msight	www.expasy.ch/MSight [47]	Open source	p,f,cd,i,a
Xcms	http://masspec.scripps.edu/xcms/xcms.php [48]	Open source	p,f,cd,i,a
VIPER	http://omics.pnl.gov/software/VIPER.php [49]	Open source	p,f,cd,i,a
OpenMS	http://open-ms.sourceforge.net [50]	Open source	p,f,cd,i,a
Mosavisu	www.mosaiques.com	Commercial	p,f,cd,i,mc
ProTrawler/Regatta	www.bioanalyte.com	Commercial	p,f,cd,i,mc,a
MarkerView	http://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=601522	Commercial	p,f,cd,i
Progenesis	www.nonlinear.com/products/progenesis/lcms/overview.asp	Commercial	p,f,cd,i,a
MassLynx	www.waters.com/waters/nav.htm?locale=en_US&cid=513164	Commercial	p,f,cd,i,a

The functionality is abbreviated as follows: p = peak detection, f = feature detection, cd = charge deconvolution, mc = multiple charge artifacts removal, i = de-isotoping, a = alignment.

Hence, the rich arsenal in software packages for analyzing genomics data may easily be used for proteomics data from LC–MS/CE-MS experiments. However, one must be aware about the high covariance structure in the peak expression which has its origins in common biological and chemical modifications of proteins, e.g. PTMs (Post Translational Modifications). Another technical difference is the sparseness of the data matrix (many entries are zeros). The origin of these zeros may be either biological (i.e. the protein is really absent in the sample) or technical (i.e. the proteins is present but its signal is below the detection limit). These non-random missing values may introduce substantial bias in the downstream statistical analysis so that proper analysis of this zeros is required [51].

The statistical analysis of the $n \times p$ data matrix usually includes some feature selection steps combined with pattern classification algorithms. Several classification approaches from machine learning and computer science research communities have been transferred to proteomics and genomics, which range from simple classification rules to the combination of classifiers [52]. Comparing the results of different procedures is beyond the scope of this manuscript. Here, we review some of the feature selection approaches as well as a simple classification method. The emergence of the free statistical software R (the R-project) [53]www.r-project.org and the related bioconductor-project [54,55]www.bioconductor.org have particularly boosted these topics as both projects offer a huge number of contributed packages that cover almost all relevant algorithms. The adaptation of R to personal needs is straightforward for trained programmers. Furthermore, its interfacing to popular programming languages such as C/C++ or Java is well developed. Given the fact that not all researchers in the omics fields are skilled programmers, several commercial and open source GUIs (graphical user interfaces) have been developed around the R core package. Examples for the open source solutions are the BRB-arrayTools (www.nci.nih.gov/BRB-arrayTools.html) and the rattle R package (http://rattle.togaware.com). Good statistical analysis platforms are also provided by commercial packages like the Rosetta Syllego System www.rosettabio.com/products/syllego, and the GeneSpring GX package www.chem.agilent.com/enUS/Products/software/lifesciencesinformatics/genespringgx/Pages/default. Interested readers may consult, e.g. the web site genome-www5.stanford.edu/resources/restech.shtml for a more complete list. Other knowledge discovery packages may also be used for the clustering, feature selection or classification of LC–MS/CE-MS compiled data. Prominent examples which are freely available are listed in Table 3.

6.1. Feature selection strategies

The curse of dimensionality drives us to reduce the dimension before applying classification algorithms. Most classification methods require the number of the predictors p be smaller than the number of samples n ($p \ll n$). However, in proteomics we often have $p \gg n$. As a consequence, we end up with an ill-conditioned problem. The fundamental assumption for avoiding this problem is that the informative part of the data (at least approximately) lies on a subspace of smaller dimension than the original one.

Table 3
Tools for mining and classifying LC–MS or CE-MS profiles.

Name	Website-reference	Language	License
Rapidminer	www.rapidminer.com [56]	Java	Open source
Weka	www.cs.waikato.ac.nz/ml/weka [57]	Java	Open source
Orange	www.ailab.si/orange [58]	Python	Open source

It is well known that statistical feature selection may improve classification accuracy in the validation set [59,60]. Feature selection procedures may be divided into three categories:

- The embedded feature selection approach, where the ranking of the features is part of the training of the classifier.
- The wrapper feature selection approach, where the ranking of the features is directly related to their contribution to the performance of the classifier.
- The filter feature selection approach, where the ranking of the features is based on some criterion of how well the feature discriminate the classes.

Here we focus on the last approach that may in turn be subdivided into univariate and multivariate filter procedures. The univariate filter (also known as forward filtering) usually relies on some statistical test, which ensures that each individual feature is discriminative between different classes with some predefined statistical confidence. Prominent tests that are widely used in biological data processing are the student t-test for normal data and the non-parametric Wilcoxon–Mann–Whitney test for non-Gaussian data. Since the number of potential biomarkers is high, multiple hypothesis tests are usually performed. This requires appropriate adjustments. To give an example, let us suppose that we perform n independent tests using 0.05 as the critical significance level. The probability for a single test to come to a non-significant (that is a correct conclusion) result is hence $1 - 0.05 = 0.95$ (95%). Since the n tests are independent, the probability that all these n tests to correctly reject the n null hypothesis is simply given by the product of the single results, i.e. $0.95 \cdots 0.95 = 0.95^n$. The probability of at least wrongly reject one of the n null hypothesis is given by $1 - 0.95^n$. Thus, if our experiment performs 300 tests on 300 biomarkers, the error probability is given by $1 - 0.95^{300} = 0.99996$. In other words, we are almost sure that by performing 300 tests on 300 biomarkers, at least one of declared significant findings will be a false positive. Because of the test independence, the probability Pr(k false positives) for having k such false positives among the n biomarkers is simply given by the binomial distribution with the significance level α as the probability of "success" (i.e. having a false positive):

$$\Pr(k \text{ false positives}) = \binom{n}{k} (1-\alpha)^{n-k} \alpha^k. \quad (1)$$

This probability approaches a Poisson distribution for large n and small α with $n\alpha$ being the expected number of false positives. In our example of the 300 biomarkers tested at the significance level of 0.05, this probability is 0.9976679 for $k = 6$ and is 0.822023 even for $k = 12$.

Bonferroni corrections (and their relatives such as the Holm procedure) are the most widely used approach for controlling the experiment-wide false positive rate. Distribution free resampling methods, like the Westfall and Young resampling procedures are also quite popular methods for the control of the family-wise error rate (FWER). A review paper [61] presents a very detailed and practical yet mathematically thorough introduction to this topic. A major drawback of these procedures is that they may lack statistical power. This leads Benjamini and Hochberg to introduce the approach of false discovery rate (FDR) [62]. The FDR is the fraction of false positives among all tests declared significant. Since in biomarker discovery one is usually concerned with finding all those that are differentially expressed between two conditions, we are ready to accept some false positives to accomplish this. However, we also want to assure a low number of false positives. Hence the idea is to control the FDR at some given value δ. Controlling the FDR at $\delta = 0.05$ means that on average 5% of the biomark-

ers declared significant are actually false positives. On the other hand, the other 95% of the biomarkers are indeed true positives. The major drawback of such an univariate approach is that it may select highly correlated features discriminating the classes in similar ways.

The multivariate filter approach transforms a large number of the original variables to a new set of variables which are uncorrelated and ordered so that the first few account for most of the variation in the data. The classical approach is the principle component analysis (PCA), where the new components are chosen such that they maximize the variability of the original predictors across samples [63]. In the partial least squares (PLS), the new components are chosen such that they maximize the covariance between the original predictors and the response variables [64]. Another attractive approach is the sliced reverse regression (SIR), which uses the inverse regression curve for dimension reduction. Recently, [65] compared these three methods for subsequent logistic regression classification and found that the PLS method has the highest performance/computation time ratio. One disadvantage of the multivariate filtering approach is that it may miss biologically important features just because they are highly correlated to some other presumably less meaningful variables for explaining the biology.

Similar to the multivariate approach, global statistical tests are also used to assess the significance of a whole set of biomarkers. These may again be beneficial since the covariance structure of data is fully taken into account. Several statistical global tests were introduced [66–69]. These tests eventually shift the analysis from the searching for single biomarkers to identifying pathways and interaction networks. Finally, we note that at least for the microarray data, the use of multivariate feature selection techniques do not necessarily mean a better performance for subsequent classification [70]. It will be interesting to clarify if this is also correct for LC–MS or CE-MS proteomics data.

6.2. Classification

Since it was proposed by Fisher in 1936 [71], the LDA has emerged to be one of the easiest, yet also powerful classification methods. In contrast to sophisticated methods such as neural networks, its stringent simplicity allows for a theoretically solid basis and great practical usefulness. We here follow [13] and summarize some of its properties in order to turn attention to possible pitfalls that may accompany a classification task. Let us suppose that we want to classify samples drawn from multivariate normal densities:

$$g_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{(\mathbf{x}-\mu_\mathbf{k})^T\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{k})}{2}\right\}, \quad (2)$$

where $\mathbf{x} = (x_1 \cdots x_p)$, $\mu_\mathbf{k}$ is the mean vector and Σ is the covariance matrix. Application of the Bayes rule leads to the so-called discriminant function

$$\delta_k(\mathbf{x}) = \mathbf{x}^T\Sigma^{-1}\mu_\mathbf{k} - \tfrac{1}{2}\mu_\mathbf{k}^T\Sigma^{-1}\mu_\mathbf{k} + \log(\pi_k), \quad (3)$$

where π_k is the prior probability for class k. In practice, the parameters π_k, μ_k and Σ are estimated from the training data. For the two class problem (e.g. control and disease group), the LDA rule classifies a new sample to the disease group (group 2 of size N_2) if

$$\mathbf{x}^T\Sigma^{-1}(\mu_2-\mu_1) > \tfrac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\left(\tfrac{N_1}{N_2}\right), \quad (4)$$

otherwise to the control group (group 1 of size N_1).

From Eqs. (3) and (4), we see that the LDA requires an inversion of the covariance matrix Σ. Thus, a full rank matrix is required, which is only the case if the number of variables (biomarkers) p is less than or equal to the number of samples n. The situation in proteomics is rather the opposite ($p \gg n$). Even if $n = p$ the matrix will still not be invertible if some of its vectors are collinear. Hence, the problem of colinearity of biomarkers may lead to instable LDA classification rules. Actually, these numerical instabilities remain even using approximate inversions such as singular value decomposition. To avoid this problem, it is advisable to perform PCA before LDA. Besides leaving one with well posed problem, such step amount to eliminating small eigen-values that may make the LDA rule useless and make direct matrix operations numerically intractable. Another popular solution is to use a regularized version of the covariance matrix $\tilde{\Sigma}$ such as:

$$\tilde{\Sigma} = \alpha\Sigma + (1-\alpha)\lambda I_p \quad (5)$$

where $0 < \alpha < 1$ is a parameter and I_p is the identity matrix of p dimensions. This procedure leads to the so-called regularized LDA.

The performance of the LDA for very high dimensional classification tasks, while often satisfactory, is not optimal [72,73] and the required matrix calculations may become highly involved. Hence other classification methods such as support vector classifier [74] or boosting algorithms [75] may be more suitable for some classification purposes. A direct comparison of the performance of these different methods for classifying LC–MS or CE-MS profiles is beyond the scope of this paper. For MALDI derived profiles a detailed comparison is reported in [76].

6.3. Cross-validation

Different classification algorithms do show different sensitivity to noisy data and outliers as well as different susceptibility to the overfitting problem. This problem is related to the generalization power (i.e. the prediction capabilities for unknown samples) of the chosen method. The lack of detailed error estimation may hazard the diagnostic power of biomarkers derived from proteomics studies and had led to a very controversial discussion about the clinical value of such findings [77–86]. One way to avoid falsely low error estimates is to use cross-validation. The data in a study may be divided into training and validation sets. The training data are used for feature selection and training the classifier to derive a decision rule that is then applied on the validation set. Several papers have shown that performing feature selection on the entire dataset often grossly underestimates the true generalization error [87,88]. Methods for estimating the test error include leave-one-out cross-validation, k-fold validation, and random subsampling validation. The choice of validation methods largely depends on the sample size available. The suitability of those methods for the omics data has been discussed in detail in [89–91].

7. Conclusion

The clinical utility of proteomic analysis is an area of still unrealized potential. It offers the promise of diagnosis, prognosis, and therapeutic follow-up of human diseases. Given the current status of measurement reproducibility and lack of standardization, comparative studies are of great importance.

The biological complexity of biofluids made the hyphenation techniques, where different combinations of separation techniques coupled with various detection schemes the methods of choice. Most methods generate two-dimensional data charts of huge size. The analysis of the generated data warrants an interdisciplinary joint effort to address its complexity. Of particular importance are the bioinformatical and statistical topics. The need of standards in the field was recognized as very stringent and the international Human Proteome Organization (HUPO) www.hupo.org launched the MIAPE [1] standard for sharing experimental data between research groups. Conformity with the Clinical Data Interchange

Standards Consortium (CDISC) www.cdisc.org and the Health Level 7 (HL7) www.hl7.org guidelines still have to be addressed. The vendor independent file formats like mzXML and the availability of public data repositories are milestones on this road. Public access to data is the requirement for clinical proteomics development that was grossly underestimated in the initial studies. The adoption of standards will boost meta-analysis using raw data from different centers. For example, by making their data publicly available, the claim of diagnosing ovarian cancer at the earliest stages of the disease [92] could not be confirmed [80,81].

For study design, all sources of variability must be carefully considered [12,93,94]. For instance, biological inter- and intra-patient variability due to confounding factors such as age and gender may influence the protein profile obtained from a given sample. Another variability is due to differences in sample collection, handling, storage, instrumentation and raw data processing. To obtain sound results, a sufficiently large number of samples must be employed to best represent the target population of interest as non-tailored sample sizes may also bias the findings. Further, any finding must be validated in an independent test set before they can be considered valuable for reporting [2]

The cross center comparisons and reproducibility certainly boost the ultimate goal of clinical proteomics to determine which proteins or groups of proteins are responsible for a specific function or disease. Accurate identification of protein is crucial if the biological role of biomarkers is to be discovered and exploited as potential therapeutic targets. A number of the novel putative biomarkers discussed represent fragments of proteins that have undergone disease-specific PTMs. LC or CE MS/MS based techniques must still undergo strict scrutiny for testing if they are capable of detecting these subtle differences.

While it is evident that a lot of additional work is required to bring clinical proteomics to its full potential, we also want to point out that it has already proven a valuable and successful approach in several recent studies [18–22].

References

[1] C.F. Taylor, N.W. Paton, K.S. Lilley, P.A. Binz, R.K. Julian, A.R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T.A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates, H. Hermjakob, Nat. Biotechnol. 25 (2007) 887.

[2] H. Mischak, R. Apweiler, R.E. Banks, M. Conaway, J.J. Coon, A. Dominizak, J.H. Ehrich, D. Fliser, M. Girolami, H. Hermjakob, D.F. Hochstrasser, V. Jankowski, B.A. Julian, W. Kolch, Z. Massy, C. Neusüss, J. Novak, K. Peter, K. Rossing, J.P. Schanstra, O.J. Semmes, D. Theodorescu, V. Thongboonkerd, E.M. Weissinger, J.E. Van Eyk, T. Yamamoto, Proteomics Clin. Appl. 1 (2007) 148.

[3] R. Aebersold, M. Mann, Nature 422 (2003) 198.

[4] V. Kasicka, Electrophoresis 29 (2008) 179.

[5] V. Dolnik, Electrophoresis 29 (2008) 143.

[6] E. Schiffer, H. Mischak, J. Novak, Proteomics 6 (2006) 5615.

[7] D. Fliser, J. Novak, V. Thongboonkerd, A. Argiles, V. Jankowski, M. Girolami, J. Jankowski, H. Mischak, J. Am. Soc. Nephrol. 18 (2007) 1057.

[8] H. Mischak, B.A. Julian, J. Novak, Proteomics Clin. Appl. 1 (2007) 792.

[9] P.G. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu, R. Aebersold, Nat. Biotechnol. 22 (2004) 1459.

[10] S.M. Lin, L. Zhu, A.Q. Winter, M. Sasinowski, W.A. Kibbe, Exp. Rev. Proteomics 2 (2005) 839.

[11] J. Listgarten, A. Emili, Drug Discov. Today 10 (2005) 1697.

[12] J.A. Lewis, Stat. Med. 18 (1999) 1903.

[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, 2001.

[14] J. Listgarten, A. Emili, Mol. Cell. Proteomics 4 (2005) 419.

[15] D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, A. Emili, Mol. Cell. Proteomics 10 (2004) 984.

[16] A.H. America, J.H. Cordewener, Proteomics 8 (2008) 731.

[17] M. Vandenbogaert, S. Li-Thiao-Té, H.M. Kaltenbach, R. Zhang, T. Aittokallio, B. Schwikowski, Proteomics 8 (2008) 650.

[18] S. Decramer, S. Wittke, H. Mischak, P. Zürbig, M. Walden, F. Bouissou, J.L. Bascands, J.P. Schanstra, Nat. Med. 12 (2006) 398.

[19] D. Theodorescu, S. Wittke, M.M. Ross, M. Walden, M. Conaway, I. Just, H. Mischak, H.F. Frierson, Lancet Oncol. 7 (2006) 230.

[20] E.M. Weissinger, E. Schiffer, B. Hertenstein, J.L. Ferrara, E. Holler, M. Stadler, H.J. Kolb, A. Zander, P. Zürbig, M. Kellmann, A. Ganser, Blood 109 (2007) 5511.

[21] L.U. Zimmerli, E. Schiffer, P. Zürbig, D.M. Good, M. Kellmann, L. Mouls, A.R. Pitt, J.J. Coon, R.E. Schmieder, K.H. Peter, H. Mischak, W. Kolch, C. Delles, A.F. Dominiczak, Mol. Cell. Proteomics 7 (2008) 290.

[22] K. Rossing, H. Mischak, M. Dakna, P. Zürbig, J. Novak, B.A. Julian, D.M. Good, J.J. Coon, L. Tarnow, P. Rossing, J. Am. Soc. Nephrol. 19 (2008) 1283.

[23] P.K. Jensen, L. Pasa-Tolic, G.A. Anderson, J.A. Horner, M.S. Lipton, J.E. Bruce, R.D. Smith, Anal. Chem. 71 (1999) 2076.

[24] W. Kolch, C. Neusüss, M. Pelzing, H. Mischak, Mass Spectrom. Rev. 24 (2005) 959.

[25] A. Gaspar, M. Englmann, A. Fekete, M. Harir, P. Schmitt-Kopplin, Electrophoresis 29 (2008) 66.

[26] A.D. Zamfir, J. Chromatogr. A 1159 (2007) 2.

[27] F.W. Tempels, W.J. Underberg, G.W. Somsen, G.J. de Jong, Electrophoresis 28 (2007) 1319.

[28] R. Haselberg, G.J. de Jong, G.W. Somsen, J. Chromatogr. A 1159 (2007) 81.

[29] J. Hernandez-Borges, T.M. Borges-Miquel, M.A. Rodriguez-Delgado, A. Cifuentes, J. Chromatogr. A 1153 (2007) 214.

[30] P. Zürbig, M.B. Renfrow, E. Schiffer, J. Novak, M. Walden, S. Wittke, I. Just, M. Pelzing, C. Neusüss, D. Theodorescu, C. Root, M. Ross, H. Mischak, Electrophoresis 27 (2006) 2111.

[31] E.J. Song, S.M. Babar, E. Oh, M.N. Hasan, H.M. Hong, Y.S. Yoo, Electrophoresis 29 (2008) 129.

[32] R. Bakry, C.W. Huck, M. Najam-ul-Haq, M. Rainer, G.K. Bonn, J. Sep. Sci. 30 (2007) 192.

[33] M. Frommberger, P. Zürbig, J. Jantos, T. Krahn, M. Mischak, A. Pich, I. Jus, P. Schmitt-Kopplin, E. Schiffer, Proteomics Clin. Appl. 1 (2007) 650.

[34] D. Theodorescu, E. Schiffer, H.W. Bauer, F. Douwes, F. Eichhorn, R. Polley, T. Schmidt, W. Schofer, P. Zürbig, D.M. Good, J.J. Coon, H. Mischak, Proteomics Clin. Appl. 2 (2008) 556.

[35] J.J. Coon, P. Zürbig, M. Dakna, A.F. Dominiczak, S. Decramer, D. Fliser, M. Frommberger, I. Golovko, D.M. Good, S. Herget-Rosenthal, J. Jankowski, B.A. Julian, M. Kellmann, W. Kolch, Z. Massy, J. Novak, K. Rossing, J.P. Schanstra, E. Schiffer, D. Theodorescu, R. Vanholder, E.M. Weissinger, H. Mischak, P. Schmitt-Kopplin, Proteomics Clin. Appl. 2 (2008) 964.

[36] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C.W. Lin, J. Chen, D. Goodlet, J. Whiteaker, A. Paulovich, M. McIntosh, Bioinformatics 22 (2006) 1902.

[37] D. May, M. Fitzgibbon, Y. Liu, T. Holzman, J. Eng, J. Whiteaker, A. Paulovich, M. McIntosh, J. Proteome Res. 6 (2007) 2685.

[38] A. Keller, A. Nesvizhskii, E. Kolke, R. Aebersold, Anal. Chem. 74 (2002) 5383.

[39] A. Rauch, M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman, P. Hussey, M. Igra, B. MacLean, C.W. Lin, A. Detter, R. Fang, V. Faca, P. Gafken, H. Zhang, J. Whiteaker, D. States, S. Hanash, A. Paulovich, M. McIntosh, J. Proteome Res. 5 (2006) 112.

[40] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M.A. Gillette, S.A. Carr, Mol. Cell. Proteomics 5 (2006) 1927.

[41] J.T. Prince, M.W. Carlson, R. Wang, P. Lu, E.M. Marcotte, Nat. Biotechnol. 22 (2004) 471.

[42] F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S.N. Loevenich, R. Aebersold, Nucleic Acids Res. 34 (2006) 655.

[43] R. Matthiesen, Proteomics 7 (2007) 2815.

[44] P.M. Palagi, P. Hernandez, D. Walther, R.D. Appel, Proteomics 6 (2006) 5435.

[45] F. Lisacek, C. Hoogland, P. Lescuyer, D.F. Hochstrasser, R.D. Appel, Proteomics Clin. Appl. 1 (2007) 900.

[46] M.C. Codrea, C.R. Jimenez, J. Heringa, E. Marchiori, Comp. Methods Prog. Biomed. 86 (2007) 281.

[47] P.M. Palagi, D. Walther, M. Quadroni, S. Catherinet, J. Burgess, C.G. Zimmermann-Ivol, J.-C. Sanchez, P.-A. Binz, D.F. Hochstrasser, R.D. Appel, Proteomics 5 (2005) 2381.

[48] C.A. Smith, E.J. Want, G.C. Tong, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.

[49] J.S. Zimmer, M.E. Monroe, W.J. Qian, R.D. Smith, Mass Spectrom. Rev. 25 (2006) 450.

[50] N. Pfeifer, A. Leinenbach, C.G. Huber, O. Kohlbacher, BMC Bioinform. 8 (2007) 468.

[51] P. Wang, H. Tang, H. Zhang, J. Whiteaker, A.G. Paulovich, M. McIntosh, Proc. Pacific Symp. Biocomput. 11 (2006) 315.

[52] R. Polikar, IEEE Circuits Syst. Mag. 6 (2006) 21.

[53] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.

[54] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, Genome Biol. 5 (2004) R80.

[55] R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (Eds.), Springer Series in Statistics for Biology and Health, 2005.

[56] I. Mierswa, M. Ingo, R. Wurst, M. Klinkenberg, T. Scholz, Euler, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.
[57] I.H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[58] J. Demsar, B. Zupan, G. Leban, Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana, 2004.
[59] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature Extraction, Foundations and Applications. Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.
[60] I. Guyon, A. Elisseeff, J. Mach. Learn. Res. 3 (2003) 1157.
[61] S. Dudoit, M.J. van der Laan, Multiple Testing Procedures with Applications to Genomics, Springer Series in Statistics, 2008.
[62] Y. Benjamini, Y. Hochberg, J. R. Stat. Soc. B 57 (1995) 289.
[63] T. Speed, Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC, 2003.
[64] A. Boulesteix, Stat. Appl. Genet. Mol. Biol. 3 (2004), Article 33, www.bepress.com/sagmb/vol3/iss1/art33.
[65] J. Dai, L. Lieu, D. Rock, Stat. Appl. Genet. Mol. Biol. 5 (2006), Article 6, www.bepress.com/sagmb/vol5/iss1/art6.
[66] H. Hotelling, J. Educ. Psychol. 24 (1933) 417, 498.
[67] Y. Lu, P.Y. Liu, P. Xiao, H.W. Deng, Bioinformatics 21 (2005) 3105.
[68] A.C. Culhane, G. Perriére, E.C. Considine, T.G. Cotter, D.G. Higgins, Bioinformatics 18 (2002) 1600.
[69] J.J. Goeman, S.A. van de Geer, F. de Kort, H.C. van Houwelingen, Bioinformatics 20 (2004) 93.
[70] C. Lai, M.J.T. Reinders, L.J. van't Veer, L.F.A. Wessels, BMC Bioinform. 7 (2006) 235.
[71] R.A. Fisher, Ann. Eugenics 7 (1936) 179.
[72] P. Dipillo, Communication in Statistics-Theory and Methodology A5 (1976) 843.
[73] P. Dipillo, Commun. Stat.-Theory Methodol. A6 (1977) 933.
[74] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.
[75] Y. Freund, Mach. Learn. 43 (2001) 293.
[76] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao, Bioinformatics 19 (2003) 1636.
[77] E. Check, Nature 429 (2004) 496.
[78] E.P. Diamandis, Clin. Chem. 49 (2003) 1272.
[79] G.L. Hortin, Clin. Chem. 51 (2005) 3.
[80] J.M. Sorace, M. Zhan, BMC Bioinform. 4 (2003) 24.
[81] K.A. Baggerly, K.R. Morris, Coombes, Bioinformatics 20 (2004) 777.
[82] K.R. Coombes, J.S. Morris, J. Hu, S.R. Edmondson, K.A. Baggerly, Nat. Biotechnol. 23 (2005) 291.
[83] E.P. Diamandis, J. Natl. Cancer Inst. 96 (2004) 353.
[84] D.F. Ransohoff, J. Natl. Cancer Inst. 97 (2005) 315.
[85] D.F. Ransohoff, Nat. Rev. Cancer 4 (2004) 309.
[86] D.F. Ransohoff, Nat. Rev. Cancer 5 (2005) 142.
[87] S. Varma, R. Simon, BMC Bioinform. 7 (2006) 91.
[88] P.V. Purohit, D.M. Rocke, Proteomics 3 (2003) 1699.
[89] U.M. Braga-Neto, E.R. Dougherty, Bioinformatics 20 (2004) 374.
[90] B.J.A. Mertens, M.E. de Noo, R.A.E.M. Tollenaar, A.M. Deelder, J. Comput. Biol. 13 (2006) 1591.
[91] D.I. Broadhurst, D.B. Kell, Metabolomics 2 (2006) 171.
[92] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Lancet 359 (2002) 572.
[93] D.A. Colantonio, D.W. Chan, Clin. Chim. Acta 357 (2005) 151.
[94] Lyons-Weiler, Cancer Inform. 1 (2005) 1.