BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 2. 144–155
Advance Access publication March 20, 2008

doi:10.1093/bib/bbn007

Algorithms and tools for analysis and management of mass spectrometry data

Pierangelo Veltri

Submitted: 3rd September 2007; Received (in revised form): 25th January 2008

Abstract

Mass spectrometry (MS) is a technique that is used for biological studies. It consists in associating a spectrum to a biological sample. A spectrum consists of couples of values (intensity, *m*/*z*), where intensity measures the abundance of biomolecules (as proteins) with a mass-to-charge ratio (*m*/*z*) present in the originating sample. In proteomics experiments, MS spectra are used to identify pattern expressions in clinical samples that may be responsible of diseases. Recently, to improve the identification of peptides/proteins related to patterns, MS/MS process is used, consisting in performing cascade of mass spectrometric analysis on selected peaks. Latter technique has been demonstrated to improve the identification and quantification of proteins/peptide in samples. Nevertheless, MS analysis deals with a huge amount of data, often affected by noises, thus requiring automatic data management systems. Tools have been developed and most of the time furnished with the instruments allowing: (i) spectra analysis and visualization, (ii) pattern recognition, (iii) protein databases querying, (iv) peptides/proteins quantification and identification. Currently most of the tools supporting such phases need to be optimized to improve the protein (and their functionalities) identification processes. In this article we survey on applications supporting spectrometrists and biologists in obtaining information from biological samples, analyzing available software for different phases. We consider different mass spectrometry techniques, and thus different requirements. We focus on tools for (i) data preprocessing, allowing to prepare results obtained from spectrometers to be analyzed; (ii) spectra analysis, representation and mining, aimed to identify common and/or hidden patterns in spectra sets or in classifying data; (iii) databases querying to identify peptides; and (iv) improving and boosting the identification and quantification of selected peaks. We trace some open problems and report on requirements that represent new challenges for bioinformatics.

Keywords: *mass spectrometry; protein databases; proteomics; protein identification; data management systems*

INTRODUCTION

Mass spectrometry (MS) is recently playing an important role in studying biological samples. It consists in generating a signal (spectrum) of values (*m/z*, intensity) related to the presence of a biomolecule with a certain mass-to-charge ratio (*m/z*), and abundance (intensity) in the original sample. Figure 1 shows an example of MS spectra, where the peak ($m/z = 5736{,}85$) denotes the presence of insuline, whereas the peak ($m/z = 8688{,}14$) denotes the presence of mioglobine in the original sample. MS-based analysis includes data preprocessing phase, database querying and data analysis phase [1–3].

MS instruments generate results as sequences of values stored in flat files that are managed by file systems. There exist a lot of tools for extracting information from such files. Nevertheless, managing and analyzing hundreds of spectra, where each one may occupy Gbyte of memory, is currently a challenging research topic. Moreover, data produced

Corresponding author. Pierangelo Veltri, Dipartimento Medicina Sperimentale e Clinica, Università Magna Graecià - Catanzaro, Italy. Tel: +3909613694149; Fax: +3909613694073; E-mail: veltri@unicz.it

Pierangelo Veltri earned his PhD in computer science from INRIA (France). He is currently an assistant professor at University Magna Graecia of Catanzaro (UMG), Italy, and his main interests are data management for biomedical applications and data integration.

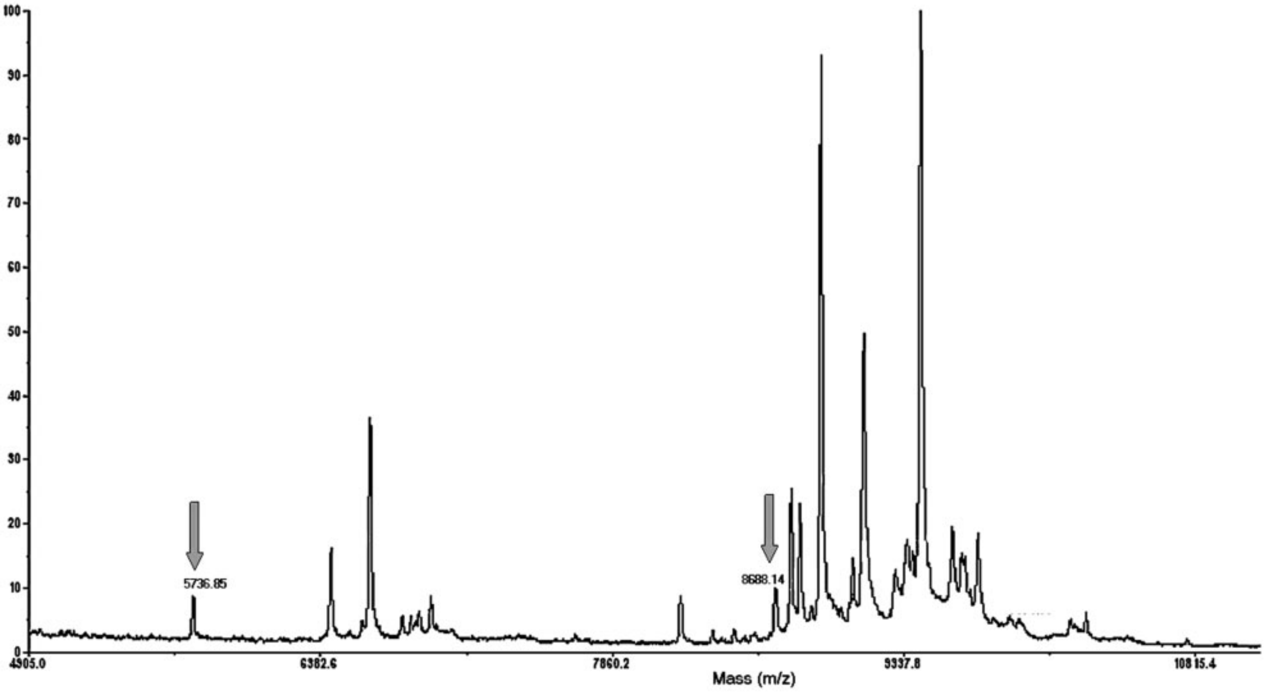


Figure 1: Mass spectrum of a biological sample: note that peaks can be associated to two proteins.

by mass spectrometer are affected by errors introduced during different experiment phases, causing noise, peak broadening, instrument distortion and saturation, miscalibration, contaminants, etc. Binning, alignments, base line subtraction are examples of preprocessing algorithms applied to improve MS data analysis [4–6].

Due to the large volume of data, MS data are attracting database communities as interesting application for new technologies [7, 8]. The current direction consists in enriching raw spectra data with meta-information such as operator, spectrometer type, ionization technique, parameters, biological results. Even if in some scenario it is possible to speed up spectra file management operations (e.g. by using compression techniques), the main focus in using MS is to simplify information extraction from spectra data. Large volume of data may occur while increasing the instrument's resolution or by using a more precise technique. For example, tandem mass spectrometry (MS/MS), where peaks or parts of the MS spectrum are associated to another (tandem) spectrum, produces much more information than MS. Figure 2 shows an example of MS/MS spectrum. After a first MS run, another mass spectrometric analysis is generated from the fragments of a selected peptide peak isolated in a previous MS stage. The fragments, produced via breakdown of the parent peptide through gas collisions, can be correlated to amino acid sequences by dedicated search programs [9]. MS/MS spectra may lead gigabyte of raw data, that need to be compared with information contained in publicly available databases for performing protein/peptide identification (e.g. the *SwissProt* database [10] queried using *Mascot* [11]). MS/MS data analysis uses database search to find *qualitative* information (peptide sequence identification via MS/MS) and *quantitative* information (mass measures) to produce tables of proteins/peptides (quality sample contents) with their relative expression levels (quantity sample contents). Algorithms such as *ProICAT* [12] are used to identify proteins/peptides by querying a protein database. Nevertheless, analysis process still requires improvements. Indeed, it has been shown that in MS/MS analysis many peptides quantified but not qualified may be associated to peptides contained in the analyzed sample [13, 14, 15]. The work of [16] compares *ProICAT*, *Spectrum Mill* and *SEQUEST* data analysis software showing their results are similar in terms of protein quantification, but different, and in some cases complementary, in terms of protein identification. Algorithms to efficiently manipulate

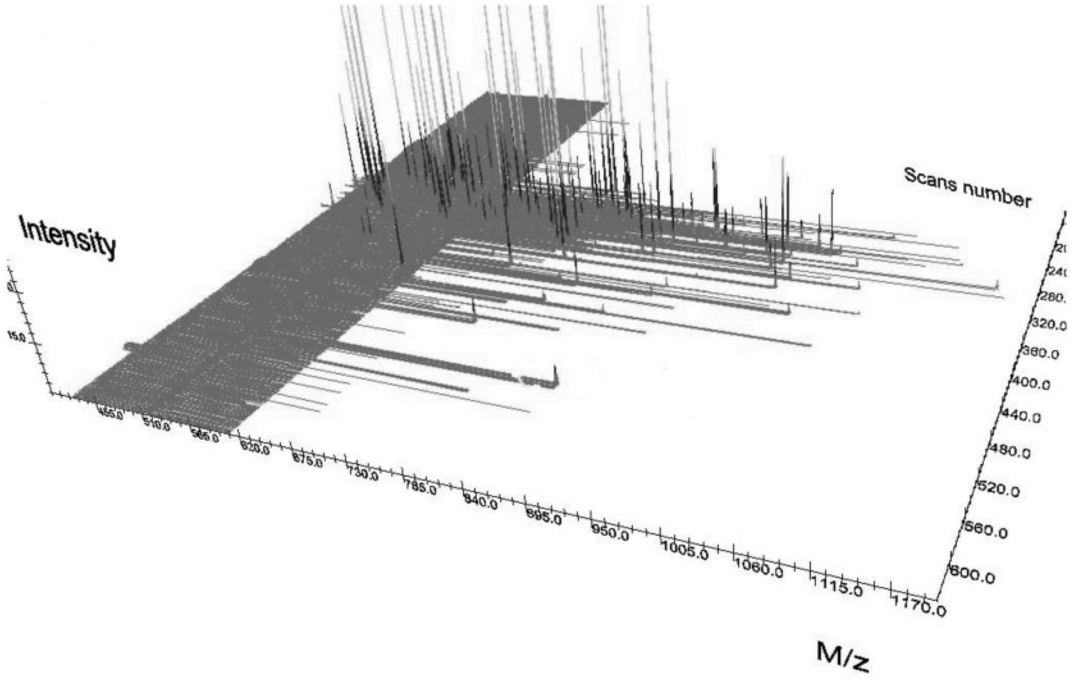


Figure 2: An example of MS/MS Spectrum. The three dimensions are: chromatographic time, mass and intensity.

spectra all studied developed. E.g. *MSight* [17], *Pep3D* [18], *msInspect* [19], *LCMS-2D* [20] and *MZmine* [21, 22] are used to locate peaks in the MS signal, associate them to peptides/proteins and eventually to compare experimental results.

In the rest of the paper we survey on aspects related to spectra data management, preprocessing and mass analysis, and protein identification processes.

MASS SPECTROMETRY TECHNIQUES

MS is a technique allowing to determine with high accuracy the molecular weight of chemical compounds, ranging from small molecules to large, polar biopolymers [23]. An MS platform includes: (i) a system to input the biological sample into the spectrometer, (ii) an ionization system, (iii) an mass analyzer, (iv) an ion detector, (v) a software system to store and analyze spectra. The sample can be inserted directly into the ionization source, or can undergo some type of separation, such as liquid chromatography (LC), gas chromatography (GC) or capillary electrophoresis (CE), where the sample is separated into different components which enter the spectrometer sequentially for individual analysis. Concerning the MS analysis of large, polar biomolecules, commonly used ionization techniques are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), coupled with different kind of mass analyzers such as time of flight (TOF) or quadrupole ion traps. Similarly, Surface-enhanced laser desorption/ionization (SELDI) is a technique used for producing spectra that differs from MALDI in the ionization used technique. Finally, LC-MS produces different spectra related to a single sample and collected at different scan times and LC-MS/MS at each scan time selects one or more precursor ions that are then fragmented and whose spectrum is produced as output.

While in MS (e.g. MALDI-TOF MS) a single sample is analyzed producing a unique spectrum, in MS/MS, the first MS selects some precursor ions (e.g. those having the higher *intensity*), then they collide in a collision cell between the two spectrometers and undergo fragmentation. The resulting daughter ions are analyzed in the second MS. Thus MS/MS produces a spectrum for each precursor ion selected by the first MS. Such spectrum can be used for protein/peptide identification. For very complex and rich samples a preliminary chromatographic separation eases the MS analysis. In LC-MS analysis, a sample is injected onto an LC column and separated into its various components. The components are then passed into the MS through an electrospray interface. The retention time of the solute is defined as the elapsed time between the time of injection of the solute and the time of elution of the peak maximum of that solute. Since different components are passed to the MS at different times, LC-MS produces a set of spectra. Usually the MS is programmed to acquire a spectrum (scan) at fixed time intervals.

From this brief introduction it is clear that MS can produce different kind of data where as a single spectrum can have different meaning according to the type of performed analysis (e.g. MS, MS/MS, LC-MS).

SPECTRA DATA HANDLING

Spectra are usually stored in text file managed by file systems, and meta information are generally treated in separate files. Nevertheless, as discussed above there is a great interest in treating spectra data together with their own meta information describing them. The need for a uniform and widely accepted access and manipulation strategy for such large datasets arises especially when information coming from single experiment on a specified MS platform has to be shared and validated among different laboratories [2]. Main interests in data spectra handling are: (i) data modelling and storing, (ii) data and information integration [24–26], (iii) data preprocessing, (iv) data querying using (experiments) metadata. Companies that produce software tools for MS analysis, provide modules designed to support all experimental phases. For instance, the Applied Biosystems company, one of the leaders in producing mass spectrometers [12], produces laboratory information management system (LIMS), an integrated tool allowing to design and execute workflow-based applications. In particular, it supports main phases of proteomics experiments: sample preparation, protein separation (using, for instance, 2D gel), data analysis and visualization, modules for protein identifications from spectra and databases querying for protein identification. Moreover, tools such as Sample-Manager [27] offers the possibility of integrating data obtained by bio-clinical instruments to Electronic Patient Records. Nevertheless, spectra data management is still realized by using the file system, storing only the physical locations of raw spectra files in the database. The suite [28] stores the physical address of the file containing raw data. Data are then loaded whenever they are necessary and managed in main memory. Such an approach does not allow any data management and indexing, limiting experimental tests that need comparing several data, as for instance required by some preprocessing or data mining algorithms. Moreover, no integration of experimental results has been still offered, to allow biological experts in generating its own local database with results related to previous experiments.

Recently data format has being proposed by proteomics community [29] to promote the use of XML-based [8] formalism to standardize spectra representation. Spectra are represented in a compressed form, enriched with metadata information simplifying spectra sharing and distribution. Each spectrum is treated as an encoded string and it is simply shareable among scientists, while (*m/z*, intensity) couples cannot directly be accessed. For instance, in the sequel a snapshot of mzData formalism of data representation is reported.

```
<?xml version="1.0" encoding="UTF-8"?>
<mzData version="">
  <description>
    <admin>
      <sampleName>sample</sampleName>
      <sourceFile>
        <nameOfFile>sample.jdx</nameOfFile>
        <pathToFile>/home/spectra/sampleFiles.jdx</pathToFile>
        <fileType>gdx</fileType>
      </sourceFile>
      <contact><name>Producer</name></contact>
    </admin>
    <instrument>
      <instrumentName>QSTAR XL LC-MS/MS</instrumentName>
    </instrument>
    <dataProcessing>
      <software><name>Analyst QS</name></software>
    </dataProcessing>
  </description>
  <spectrumList count="">
    <spectrum id=1>
      <mzArrayBinary>
        <data precision="32" endian="little" length="5">
          Q0lvdkNOIBBDTiEJQ04iA0NPfyQ=
        </data>
      </mzArrayBinary>
      <intenArrayBinary>
        <data precision="32" endian="little" length="5">
          QEAAAEEAAABAYAAAQGAAAEEAAAA=
        </data>
      </intenArrayBinary>
    </spectrum>
  </spectrumList>
</msData>
```

The metadata allows to represent information such as: spectrometer type (e.g. Applied); spectra type (MALDI, SELDI, MS/MS, etc.); experimental date and operator; biological sample description (human tissue, serum, healthy patient, etc.); notes about spectra contents (interesting proteins, biological patterns, etc.); general notes about the experimental results (e.g. patient disease); information about data manipulation (e.g. data being preprocessed or not); laboratory identification (full address, department and eventual laboratory number). They are defined using XML schema used to describe a spectrum. In the sequel we report a short example of schema definition syntax:

```
<?xml version="1.0" encoding="UTF-8" ?>
  ...
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"...>
  <xs:element name="mzData"> ... </xs:element>
  <xs:complexType name="sourceFileType"> ... </xs:complexType>
  <xs:complexType name="personType">  ... </xs:complexType>
  <xs:complexType name="softwareType">  ... </xs:complexType>
  <xs:complexType name="dataProcessingType">  ... </xs:complexType>
  <xs:complexType name="spectrumType">  ... </xs:complexType>
  <xs:group name="binaryDataGroup">  ... </xs:complexType>
  <xs:complexType name="spectrumSettingsType">  ... </xs:complexType>
  ...
  <xs:complexType name="descriptionType">  ... </xs:complexType>
  ...
</xs:schema>
```

The elements name are used to describe the information and to structurally guide the generation of the document containing the MS experiment information. In particular, `sourceFileType` element contains description and location of the source raw file generated by the instrument; the `personType` element contains operator identification information (i.e. name, institution, contact information); the `softwareType` element contains Software information (the software that produced the peak list); the dataProcessingType contains information on data preprocessing and preparation; the spectrumType element contains information about eventual relationship among spectra data (e.g. this occurs when a spectrum is obtained from processing another one); the binaryDataGroup element contains the spectrum values in a base64 encoded format; the spectrumSettingsType element contains the spectrometer parameters set for acquisition; the descriptionType contains free text notes. Instances of such a schema may be stored in XML databases or simply used as exchange formalism among MS laboratories, and the schema can be used to formulate queries.

Spectra data can be also modeled in a relational database. In this case, each spectrum can be stored in a relation, containing mass-to-charge and intensity values. Metadata can be stored in other relations. The advantage is the use of a valid and very known technology, in terms of data management, query optimization and indexing techniques [30]. For instance, the SpecDB spectra database [31] is an example of use of relational database for managing spectra data. It implements basic spectra management functions and stores spectra in their different stages (raw, preprocessed and prepared) keeping trace of the different phases of proteomics experiments. To deal with the huge volumes of mass spectra data, and to allow easy and efficient access to portions of spectra each spectrum is stored in a set of relations, each one containing portions according to a predefined window of *m/z* values. Thus, a spectrum with values of *m/z* contained in the range 0–20 000 Da is divided into a set of relations such that intensities of peaks that belong in defined ranges (for instance in portions of 500 Da) are assigned to each relation.

In the general case, a database storing and managing spectra data must contain procedures to fulfill the following requirements: (i) efficiently storing and retrieving data (single spectrum, set of spectra and portions of spectra), (ii) import/export functions (e.g. loading of raw spectra available in different text files, exporting of spectra in XML-based formalism), (iii) query/update functions able to enhance performance of data preprocessing and analysis (e.g. avoiding full main memory processing).

Finally, most of the existing solutions are related to commercial release of the MS hardware support. For instance, the Applied Biosystems [12] QSTAR XL Hybrid LC-MS/MS system coupled with the Dionex Corporation UltiMate Nano and Capillary LC system is able to generate both a proprietary non-standard formalism and a JCAMP-DX file, a chemical spectroscopic data exchange format stored as text file and containing a header with experiment metadata, and a body containing data as a table of (*m/z, intensity*) couples.

Data preprocessing

Data management has to consider preprocessing [4] procedures, in charge of reducing spectrum noise and dimension. Preprocessing aims to correct intensity and *m/z* values in order to: (i) reduce noise, (ii) reduce amount of data and (iii) make spectra comparable. Noise reduction and normalization are conducted in part by the spectrometer and in part by external preprocessing tools. This noise varies across the *m/z* axis and it generally varies across different fractions, so that a one-value-fits-all strategy cannot be applied. We now sketch some of the used techniques for spectra preprocessing.

Base line subtraction uses an iterative algorithm to attempt to remove the baseline slope and offset from a spectrum by iteratively calculating the best fit straight line through a set of estimated baseline points (Figure 3). The baseline points are determined by fitting the line through the spectrum and then discarding all data points with intensity above a threshold from the fitted line. The number of points above and below the line is then counted. If there are fewer points above the line than below, they are considered peaks and discarded. Then a new line is fit through the remaining data points. This process is repeated until the number of points above the line is less than or equal to those below the line. This final line is subtracted from the spectrum to get the baseline corrected spectrum.

Normalization enables the comparison of different samples since the absolutes peak values of different fraction of spectrum could be incomparable. The purpose of spectrum normalization is to identify and remove sources of systematic variation between spectra due, for instance, to varying amounts of sample or degradation over time in the sample or even variation in the instrument detector sensitivity. There exist different techniques such as the *Canonical Normalization*, the *Inverse Normalization* [5, 32], the *Direct Normalization* and the *Logarithmic Normalization* [6, 33].

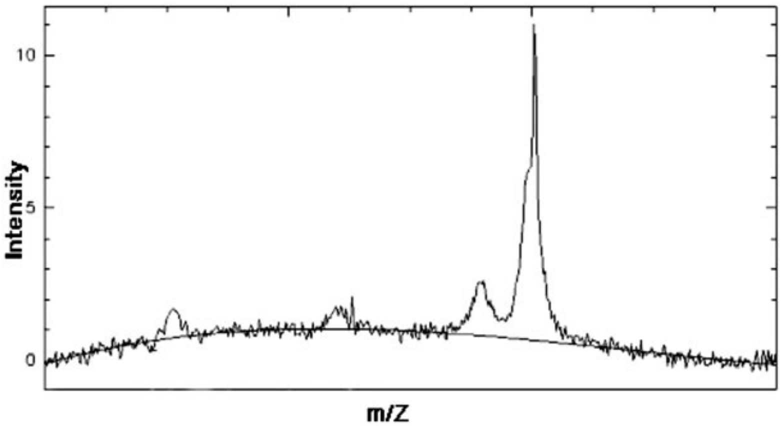


Figure 3: An example of preprocessing algorithm. Figure reports an example of baseline curve to be used with baseline substraction.

Binning performs data dimensionality reduction by grouping measured data into bins. This process involves grouping adjacent values and electing for each group a representative member. The binning algorithm takes a subset of N peaks from a spectra, represented by the couples $[(I_1, m/z_1), (I_2, m/z_1), \ldots, (I_N, m/z_N)]$, and substitutes all of them with a unique peak $(I, m/z)$, whose intensity I is an aggregate function of the N original intensities (e.g. their sum), and the mass m/z is usually chosen among the original mass values (e.g. the median value or the value corresponding to the maximum intensity). Such basic operation is conducted by scanning all the spectrum by using a sort of sliding window [5, 32].

Peaks alignment consists in *alignment* of correspondent peaks across samples. Without alignment, the same peak (e.g. the same peptide) can have different values of *m/z* across samples. To allow an easy and effective comparison of different spectra, peaks alignment methods find a common set of peak locations in a set of spectra, in such a way that all spectra have common *m/z* values for the same biological entities.

Finally, *quantization* of spectra reduces the range of possible values and obtains a further noise reduction. For MS data a non-uniform quantization model is used, in which quantization step size is larger for intensity values close to ground noise mean value.

Figures 4 and 5 show an example of raw and preprocessed spectra of a MALDI data set.

Data visualization

Even if advanced data analysis are used to identify interesting information in spectra, visual tools for spectra analysis are very useful for biologist. Data visualization tools aim to support analyst in finding interesting peaks in spectra, selecting them for further analysis and to visualize interesting peaks obtained by automatic analysis tools. The main functions required to a visualization tool are:

(1) 3D visualization of an LC-MS spectrum or of many MALDI (SELDI)-TOF spectra allowing to manipulate images;
(2) 2D visualization of MALDI-TOF and LC-MS spectra. The former visualized in the (*m/z, intensity*) plane, while the latter visualized in the (*m/z, intensity*), (*m/z, retention time*) and (*retention time, intensity*) planes;
(3) interactive visualization of ion properties on the (*m/z, retention time*) plane;
(4) peak selection and zooming for *ad hoc* inspection allowing to use peak information for database querying;
(5) different data format manipulation and data conversion.

There exist many spectra visualization. JDXview [34] displays various kinds of spectra in JCAMP-DX format. It allows zooming and measuring of distances on a spectrum and supports graphics output in vector graphics format. Pep3D [18, 35] supports LC-MS and LC-MS/MS spectra but does not provide 3D visualization or mzData conversion. Data are represented as a 2D density plot. For MS/MS experiments using collision-induced dissociation, links are embedded in the image to the daughter spectra and the corresponding peptide sequences. Nevertheless, Pep3D does not provide 3D visualization or mzData conversion. mzViewer [36] is a simple lightweight viewer that allows only 2D visualization of mzData. It is a stand-alone application, but Java

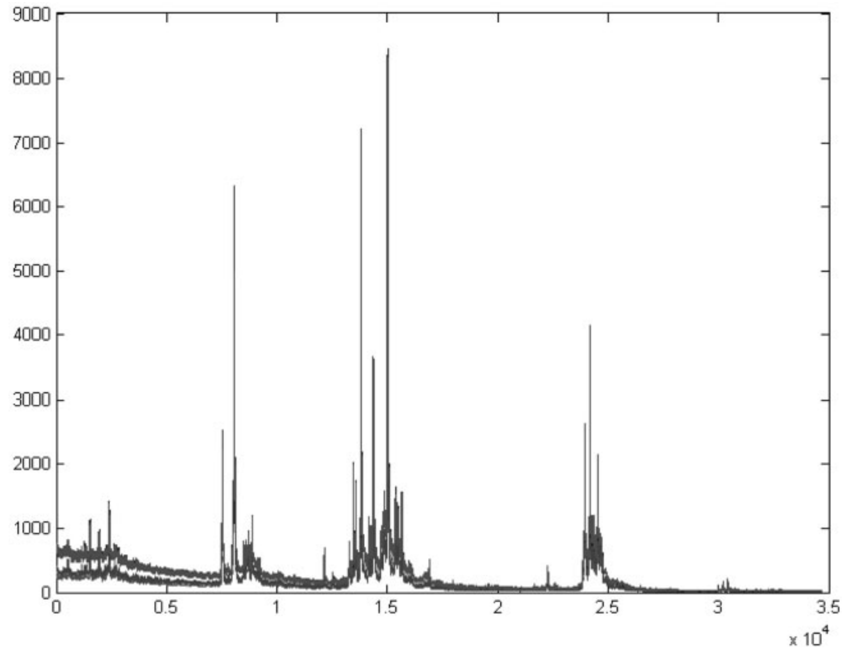


Figure 4: Example of preprocessing results: a set of spectra instances before preprocessing.

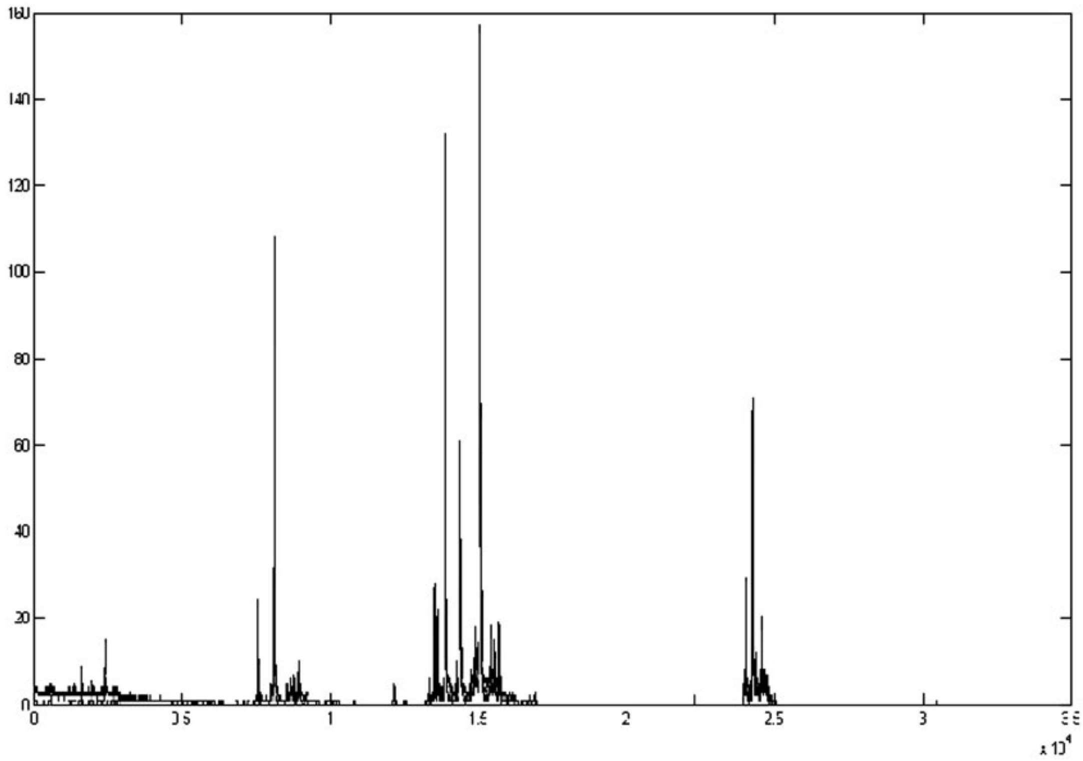


Figure 5: Example of preprocessing results: the spectra are now preprocessed.

classes are provided for incorporating the mzViewer into other applications. SpectraViewer [37] is a simple Java-based tool for the 3D visualization of spectra data and their conversion into the mzData format. It uses the Java3D API for graphic manipulation and is deployed to the user through the Java Web Start technology [37, 38]. Many other visualization tools are available in the open source community.

Advanced data manipulation

The problem of dealing with large volume of spectra data arises from the need for identifying differently expressed proteins or peptides in different samples. The analysis of several spectra coming from biological samples belonging to different subjects (e.g. healthy and diseased) focuses on identifying discriminant values in spectra (*m/z*, intensity couples corresponding to biomarkers) that are responsible of diseases. Novel approaches for extracting information from spectra have been recently developed. In [32], data-mining techniques have been used to identify discriminants in a female population, distinguishing ovarian cancer diseased from healthy ones. Similarly, data-mining techniques for spectra data have been applied in [39] for SELDI MS data, to identify discriminants in rectal cancer disease. In [40] machine-learning algorithms have been used to identify biomarkers in SELDI MS data generated on tens of patients to figure out cerebral accident discriminants. Data-mining techniques can be applied to classify MALDI spectra data according to discriminant biomarkers which can be associated to different peaks in MALDI data.

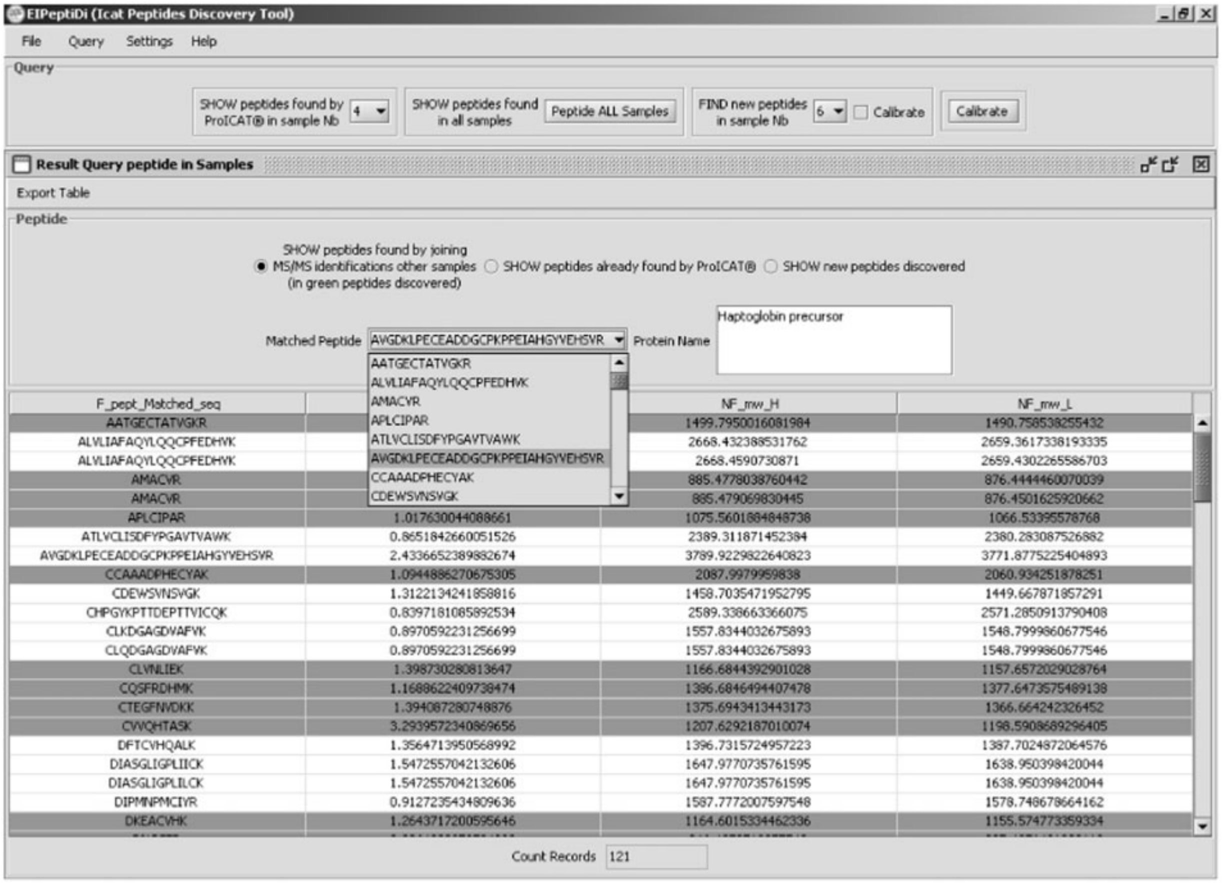
Algorithms that do not require human intervention are needed for rapid and repeatable quantitative processing of spectra that often contain hundreds of discrete peaks. *Peaks extraction* consists of separating real peaks (e.g. corresponding to peptides) from peaks representing noise. Although sometimes such task can be performed by using the data-processing embedded in mass spectrometer, custom identification methods fitting both informatics and biological considerations are more effective. Data-mining analysis can be performed to extract and identify interesting peaks in mass spectra. Tasks such as classification, clustering, pattern analysis, and the corresponding data mining algorithms and tools (Q5, C5, K-means) [41] may be used to manipulate spectra. Data dictionaries and ontologies can be used to guide experiment workflows [42].

Recently, MALDI data analysis has been performed by using time series [43]. In this case, a preprocessed mass spectrum is mapped as a sequence $S = [((m/z)_1, I_1), \ldots, ((m/z)_n, I_n)]$, where for each pair the first value refers to the-mass-to charge ratio and the second one is the associated intensity value. A mass spectrum so defined is modelled as a time series $T = [(x_1, t_1), \ldots, (x_n, t_n)]$ whose values x_i correspond to the spectrum intensity values I_i, and time steps t_i correspond to the *m/z* values. Indeed, the notion of time implicitly lies in the sequence of mass-to-charge values. T can be rewritten as $T = x_1, \ldots, x_n$ when as usual sampling period (periods) is well specified. Advanced mapping rules can be used to model such time series into a compact representation which possibly synthesizes the significant variations in the time series profile. This method gave interesting results.

PROTEIN IDENTIFICATION

In proteins identification research has focused on obtaining a list of significant peaks, each of which representing a protein/peptide contained in the original biological sample. Protein identification is performed by using software modules querying publicly available databases such as MASCOT [11] containing peaks expressions of known chemical species (e.g. proteins/peptides). Ideal spectra are compared with the experimental ones to identify composition of the biological samples. Database search usually may return several results with different matching probabilities requiring spectrometer expertise. Also, for MS/MS spectra protein/peptide identification is performed comparing experimental spectra to theoretical ones by querying publicly available databases (such as *Swiss Prot,* database [10] or the recent initiative PRiDE [44].

Recently, increasing attention has been devoted to fully exploiting the *quantitative* information obtained by LC-MS/MS experiments [21, 45, 46], where quantitative stands for detecting changes in protein abundances. Concerning quantitative aspects, the simple registration of the ion intensity of peptide peaks in MS is usually not an accurate way of acquiring information on the amount of such species. MS quantitation can be improved by using isotopic labeling methods [47], which allow for precise relative quantification between two or more protein mixtures (see for instance ICAT technique [48]). *Qualitative* information is then correlated to *quantitative* information in order to produce tables of proteins/peptides (quality sample contents) with their relative expression levels (quantity sample contents). Nevertheless, experimental observations showed that, at least in the case of plasma/serum samples, there are many missing values, i.e. peptides present in sample even if not identified by the software routine, proving that identification process may be improved. Moulder *et al.* [16] have compared some ICAT data analysis software and have shown that ProICAT, Spectrum Mill and SEQUEST give



ElPeptiDi (Icat Peptides Discovery Tool)

Query

SHOW peptides found by ProICAT® in sample Nb: 4

SHOW peptides found in all samples: Peptide ALL Samples

FIND new peptides in sample Nb: 6 □ Calibrate

Calibrate

Result Query peptide in Samples

Export Table

Peptide

● SHOW peptides found by joining MS/MS identifications other samples (in green peptides discovered) ○ SHOW peptides already found by ProICAT® ○ SHOW new peptides discovered

Matched Peptide: AVGDKLPECEADDGCPKPPEIAHGYVEHSVR

AATGECTATVGKR
ALVLIAFAQYLQQCPFEDHVK
AMACVR
APLCIPAR
ATLVCLISDFYPGAVTVAWK
AVGDKLPECEADDGCPKPPEIAHGYVEHSVR
CCAAADPHECYAK
CDEWSVNSVGK

Protein Name: Haptoglobin precursor

F_pept_Matched_seq		NF_mw_H	NF_mw_L
AATGECTATVGKR		1499.7950016081984	1490.758538255432
ALVLIAFAQYLQQCPFEDHVK		2668.432388531762	2659.3617338193335
ALVLIAFAQYLQQCPFEDHVK		2668.4590730871	2659.4302265586703
AMACVR		885.4778038760442	876.4444460070039
AMACVR		885.479069830445	876.4501625920662
APLCIPAR	1.017630044088661	1075.5601884848738	1066.53395578768
ATLVCLISDFYPGAVTVAWK	0.8651842660051526	2389.311871452384	2380.283087526882
AVGDKLPECEADDGCPKPPEIAHGYVEHSVR	2.4336652389882674	3789.9229822640823	3771.8775225404893
CCAAADPHECYAK	1.0944886270675305	2087.9979959838	2060.934251878251
CDEWSVNSVGK	1.3122134241858816	1458.7035471952795	1449.667871857291
CHPGYKPTTDEPTTVICQK	0.8397181085892534	2589.338663366075	2571.2850913790408
CLKDGAGDVAFVK	0.8970592231256699	1557.8344032675893	1548.7999860677546
CLQDGAGDVAFVK	0.8970592231256699	1557.8344032675893	1548.7999860677546
CLVNLIEK	1.398730280813647	1166.6844392901028	1157.6572029028764
CQSFRDHMK	1.1688622409738474	1386.6846494407478	1377.6473575489138
CTEGFNVDKK	1.394087280748876	1375.6943413443173	1366.664242326452
CVVQHTASK	3.2939572340869656	1207.6292187010074	1198.5908689296405
DFTCVHQALK	1.3564713950568992	1396.7315724957223	1387.7024872064576
DIASGLIGPLIICK	1.5472557042132606	1647.9770735761595	1638.950398420044
DIASGLIGPLILCK	1.5472557042132606	1647.9770735761595	1638.950398420044
DIPMNPMCIYR	0.9127235434809636	1587.7772007597548	1578.748678664162
DKEACVHK	1.2643717200595646	1164.6015334462336	1155.574773359334

Count Records: 121

Figure 6: An example of tool for improving the protein identification process in MS/MS process. The discovered peptides are highlighted. The tool is ElPeptiDi, freely available on line.

comparable results in terms of protein quantitation and identification. Moreover, in order to identify a larger number of proteins, the data analysis process has to be performed several times on the same samples. In [13] it has been proved that cross-validating results obtained by search analysis on many experiments improves the protein identification process. A proposed tool boosts the performance of protein/peptide identification process, increasing the number of identified proteins. For instance, in Figure 6 proteins discovered by using the tool proposed in [13] are reported.

The protein identification process is very important in many medical applications, e.g. for early detection of diseases. Identifying the absence of proteins (and thus of their activities) in some organism can be co-related to diseases. Spectrometry relied to protein identification and data-mining analysis is used for analysis of spectra data for clinical studies. There exist many software tools for the protein identification, such as Mascot [11], Sequest [49], X!Tandem [50], Omssa [51] and also platform suite for protein identification [52]. Their functionalities vary depending on the used techniques. Most of the used systems are based on querying publicly available databases; nevertheless new techniques have been developing. For instance, the de novo-based techniques use *m/z* peaks ratio to identify, peptide sequences avoiding the use of databases, using scoring functions. Systems using de nove are, for instance, the Sherenga algorithm [53], the Peaks system [54], PepNovo [55] (see also lists of applications on the web site [56]. Currently, the most used approaches to identify proteins/peptides are those based on accessing to a database comparing experimental spectra with sequences present in the queried database [11, 49–51]. Other software are available, mostly differing by the used techniques to compare experimental and ideal spectra.

CONCLUSIONS

In this paper we surveyed on the use of mass spectrometry and the automatic support required to software tools, mostly for extracting information from spectra. There are still many efforts to do to allow the use of spectrometry in clinical laboratories. Mainly there are open topics related to improving the definition and the population of spectra databases for different laboratories and spectrometer types. An effort has been doing for offering common user interface where formulate queries in a simple and common language, for instance, supported by ontologies, on data coming from different laboratory sources. At the same time, algorithms of data cross validation comparing quantitative informations (peaks) to known information (peaks associated to known proteins) need to be defined thus to improve available software. Nevertheless, the interest is high; it is indeed commonly accepted that mass spectrometry may allow to identify biomarkers that characterize diseases, thus that low expression of proteins in a biological samples may be used to make diagnosis and consequently to design therapies.

Key Points

- Research for curing human diseases is currently interested in the study of proteome, thus that studying protein functionalities may help in early detection of important diseases.
- Mass Spectrometry (MS), a technique widely used from physicists for accurate mass measures, has being adopted from biologists as instrument to identify the presence of proteins (and thus their functions) in biological samples.
- Spectra produced during MS experiments contain a large number of data, mostly representing biological information but only a small portion contains potentially relevant information for biological studies. Extracting relevant information and distinguishing the non-relevant from relevant ones, is an important task for spectrometrists. Preprocessing techniques have to be performed together with and automatic information extraction algorithms, to filter out noise generated during experiments and non-relevant information.
- Experimental results containing mass spectra and their associated information (such as quantified and identified protein/peptides associated to peaks in the spectra) are precious results that need to be structured and stored in database management system, to allow re-use for future experiments performed both by the same laboratory or by scientist community. There is currently a great interest in storing and publishing data results, enriching the community that already use protein databases.
- There is a commonly accepted necessity of defining standards for spectra representation, containing raw data and experimental information (spectrometer type, laboratory, protein/peptides results). Recently, the HUPO international organization designed the guide lines for hardware and software designer and vendors.
- Information on spectra containments (proteins associated to peaks in spectra) are contained in several and different databases. Algorithms for efficiently querying and gathering information from databases are required to guarantee the best results in information extraction from single spectra.

References

1. Jain AK, Dubes RC. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.
2. Beer I, Barnea E, Ziv T, *et al*. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004;**4**(4):950–60.
3. Petricoin EF, Ardekani AM, Hitt BA, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;**359**:572–7.
4. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. *Proteomics* 2003; **3**(9):1692–8.
5. Gopalakrishnan V, William E, Ranganathan S, *et al*. Proteomic data mining challenges in identification of disease-specific biomarkers from variable resolution mass spectra. In: *Proceedings of SIAM Bioinformatics Workshop 2004*, pp. 1–10. Lake Buena Vista, FL, April 2004.
6. Yasui Y, McLerran D, Adam BL, *et al*. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J Biomed Biotechnol* 2003;**1**(4):242–48.
7. Chen L, Huang Z, Ramakrishnan R. Cost-based labeling of groups of mass spectra. In *SIGMOD 2004,* 2004.
8. World Wide Web Consortium. *XML Language, 2005*. http://www.w3.org/XML/ (20 February 2008, date last accessed).
9. Steen H, Mann M. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol* 2004;**5**(9):699–711.
10. ExPASy Proteomics Server. *Swiss Prot Database, 2006*. http://www.expasy.org/sprot/ (20 February 2008, date last accessed).
11. Matrix Science Ltd. *Mascot: A Powerful Search Engine Which Uses Mass Spectrometry Data to Identify Proteins from Primary Sequence Databases, 2004*. http://www.matrixscience.com/ (12 March 2008, date last accessed).
12. Applied Biosystems. http://www.appliedbiosystems.com/ (20 February 2008, date last accessed).
13. Cannataro M, Cuda G, Gaspari M, *et al*. The EIPeptiDi tool: enhancing peptide discovery in ICAT-based LC MS/MS experiments. *BMC Bioinformatics* 2007;**8**(255) doi:10.1186/1471-2105-8-255.
14. Kratz A, Ferraro M, Sluss PM, *et al*. Case records of the massachusetts general hospital. weekly clinicopathological exercises. laboratory reference values. *New England J Medicine* 2004;**15**(351):1548–63.
15. Anderson NL, Polanski M, Pieper R, *et al*. The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* 2004;**3**(4): 311–26.
16. Moulder R, Filen JJ, Salmi J, *et al*. A comparative evaluation of software for the analysis of liquid chromatography-tandem mass spectrometry data from isotope coded affinity tag experiments. *Proteomics* 2005;**11**(5):2748–60.
17. Palagi PM, Walther D, Quadroni M, *et al*. Msight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005;**5**(9):2381–84.
18. Li X, Pedrioli PGA, Eng J, *et al*. A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Anal Chem* 2004; **76**(13):3856–3860.

19. Bellew M, Coram M, Fitzgibbon M, *et al.* A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms. *Bioinformatics* 2006;**22**(15):1902–09.
20. Du P, Sudha R, Prystowsky MB, Angeletti RH. Data reduction of isotope-resolved lc-ms spectra. *Bioinformatics* 2007, doi:10.1093/bioinformatics/btm083.
21. Katajamaa M, Miettinen J, Oresic M. Mzmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006;**22**(5):634–36.
22. Katajamaa M, Oresic M. Processing methods for differential analysis of lc/ms profile data. *BMC Bioinformatics* 2005;**6**:179 doi:10.1186/1471-2105-6-179.
23. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
24. Branden C-I and Tooze J. *Introduction to Protein Structure,* 2nd edn. New York: Garland Publishing, 1998.
25. Bujnicki JM. *Practical Bioinformatics,* 1st edn. Berlin: Springer, 2004, pp. 933–1891.
26. Lacroix Z, Critchlow T, (eds). *Bioinformatics: Managing Scientific Data*. San Francisco: Learn How, 2001.
27. Thermo Scientific. *Sample Manager*. http://www.thermo.-com/com/cda/home (20 February 2008, date last accessed).
28. Nonlinear Bioinformatics Solutions. *Non linear phoretyx*. http://www.nonlinear.com/ (12 March 2008, date last accessed).
29. Human Proteome Organization. http://www.hupo.org (20 February 2008, date last accessed).
30. Molina HG, Ullman J, Widom J. *Database System: The Complete Book*. Upper Saddle River, NJ: Prentice-Hall, 2002.
31. Cannataro M, Tradigo G, Veltri P. Sharing mass spectrometry data in a grid-based distributed proteomics laboratory. *Information Processing & Management* 2007;**43**(3):577–91.
32. Petricoin EF, Ardekani AM, Hitt BA, *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;**9306**(359):572–7.
33. Wu B, Abbott T, Fishman D, *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003;**1**(19):1636–43.
34. JDXview tool. http://merian.pch.univie.ac.at/∼nhaider/cheminf/jdxview.html (20 February 2008, date last accessed).
35. The Pep3D tool. http://tools.proteomecenter.org/Pep3D.php (12 March 2008, date last accessed).
36. mzViewer. http://www.bioinformatics.bbsrc.ac.uk/projects/mzviewer/ (12 March 2008, date last accessed).
37. SpectraViewer. http://dns2.icar.cnr.it/cannataro/projects/SpectraViewer/ (12 March 2008, date last accessed).
38. Sun Microsystems. *Java Web Start Technology, 2007.* http://java.sun.com/products/javawebstart/ (20 February 2008, date last accessed).
39. Smith FM, Gallagher WM, Fox E, *et al.* Combination of SELDI-TOF-MS and data mining provides early-stage response prediction for rectal tumors undergoing multimodal neoadjuvant therapy. *Ann Surgery* 2007;**2**(245):572–77.
40. Prados J, Kalousis A, Sanchez JC, *et al.* Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics* 2004;**8**(4):2320–32.
41. Han J, Camber M. *Data Mining: Concepts and Tecnhiques*. San Francisco: Morgan Kauffman, 2001.
42. Cannataro M, Veltri P. MS-Analyzer: preprocessing and data mining services for proteomics applications on the Grid. *Concurrency and Computation: Practice & Experience* 2007; **19**(15):2047–66.
43. Gullo F, Ponti G, Tagarelli A, *et al.* A time series based approach for classifying mass spectrometry data. In: IEEE, editor, *Proceeding of International Conference on Computer Mased Medical Systems (CBMS)*, 2007.
44. EMBL EBI. *Proteomics Identifications Database*, 2008.
45. Gaspari M, Verhoeckx KC, Verheij ER, *et al.* Integration of two-dimensional lc-ms with multivariate statistics for comparative analysis of proteomic samples. *Anal Chem* 2006; **78**(7):2286–96.
46. America AH, Cordewener JH, van Geffen MH, *et al.* Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional lc-ms. *Proteomics* 2006;**6**(2):641–53.
47. Swanson SK, Washburn MP. The continuing evolution of shotgun proteomics. *Drug Discov Today* 2005;**10**(10):719–25.
48. Gygi SP, Rist B, Gerber SA, *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;**17**(10):994–9.
49. Eng J, Yates J. *SEQUEST, 2008.* http://fields.scripps.edu/sequest/ (20 February 2008, date last accessed).
50. The Global Proteome Machine Organization. *X! Tandem*, 2008. http://www.thegpm.org/TANDEM/ (12 March 2008, date last accessed).
51. NCBI. *OMSSA*, 2008. http://pubchem.ncbi.nlm.nih.gov/omssa/ (20 February 2008, date last accessed).
52. Seattle Proteome Senter. *Trans-proteomic pipeline*, 2008.
53. Dancik V, Addona TA, Clauser KR, *et al.* De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999;**6**:327–42.
54. Bioinformatics Solutions. *PEAKS, 2008.* http://www.bioinformaticssolutions.com/products/peaks/denovo.php (20 February 2008, date last accessed).
55. Frank A, Pevzner P. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005;**77**(4): 964–73.
56. Proteome Software. *Software for Protein Identification, 2008.* http://www.proteomesoftware.com/Proteome_software_link_software.html (20 February 2008, date last accessed).