Science & Technology Development Journal – Engineering and Technology, 2(4):258-276

Open Access Full Text Article

Review

Data mining in mass spectrometry-based proteomics studies

Le Anh Vu[1,*], Phan Thi Cam Quyen[2], Nguyen Thuy Huong[1]

ABSTRACT

The post-genomic era consists of experimental and computational efforts to meet the challenge of clarifying and understanding the function of genes and their products. Proteomic studies play a key role in this endeavour by complementing other functional genomics approaches, encompasses the large-scale analysis of complex mixtures, including the identification and quantification of proteins expressed under different conditions, the determination of their properties, modifications and functions. Understanding how biological processes are regulated at the protein level is crucial to understanding the molecular basis of diseases and often highlights the prevention, diagnosis and treatment of diseases. High-throughput technologies are widely used in proteomics to perform the analysis of thousands of proteins. Specifically, mass spectrometry (MS) is an analytical technique for characterizing biological samples and is increasingly used in protein studies because of its targeted, nontargeted, and high performance abilities. However, as large data sets are created, computational methods such as data mining techniques are required to analyze and interpret the relevant data. More specifically, the application of data mining techniques in large proteomic data sets can assist in many interpretations of data; it can reveal protein-protein interactions, improve protein identification, evaluate the experimental methods used and facilitate the diagnosis and biomarker discovery. With the rapid advances in mass spectrometry devices and experimental methodologies, MS-based proteomics has become a reliable and necessary tool for elucidating biological processes at the protein level. Over the past decade, we have witnessed a great expansion of our knowledge of human diseases with the adoption of proteomic technologies based on MS, which leads to many interesting discoveries. Here, we review recent advances of data mining in MS-based proteomics in biomedical research. Recent research in many fields shows that proteomics goes beyond the simple classification of proteins in biological systems and finally reaches its initial potential – as an essential tool to aid related disciplines, notably biomedical research. From here, there is great potential for data mining in MS-based proteomics to move beyond basic research, into clinical research and diagnostics.

Key words: bioinformatics, biomedical research, data mining, mass spectrometry, proteomics

[1]*Faculty of Chemistry Engineering, Ho Chi Minh City University of Technology – VNU-HCM, Vietnam*

[2]*Department of Biotechnology, Kien Giang Seed Research Center, Vietnam*

Correspondence

Le Anh Vu, Faculty of Chemistry Engineering, Ho Chi Minh City University of Technology – VNU-HCM, Vietnam

Email: lavu68@gmail.com

History

- Received: 18-3-2019
- Accepted: 18-12-2019
- Published: 31-12-2019

DOI : 10.32508/stdjet.v2i4.483

INTRODUCTION

Proteomics encompass a broad range of technologies that allows the identification and the quantification of proteins in complex biological specimens. Proteomics approaches rely on the ability to detect small changes in protein abundance of an altered state given a control or reference condition. Thus, the identification and quantification of differences between two or more physiological states of a biological system can be defined as changes on the control sample, determining the up- or down-regulation of such protein [1]. These approaches have been extensively applied in biomedical research for the understanding of diseases, including protein-based biomarker discovery for early detection and monitoring of different types of cancer [2], the analysis of abnormal protein phosphorylation patterns associated with diseases [3] and the identification of therapeutic targets [4].

There are many technologies used to extract protein information from biological samples. These techniques cover a range of approaches and quality of extracted data. Commonly used techniques include two-dimensional gel electrophoresis, enzyme-linked immunosorbent assay (ELISA), protein arrays, affinity separation and mass spectrometry (MS) technologies. Many of these methods, such as gel electrophoresis and ELISA are limited in the number of proteins they can analyze because of time-consuming process. They also require specific proteins be selected during the design of the study and proper available antibodies; this can be a challenge for non-model organisms. Meanwhile, MS-based proteomics has become a widely used high throughput method to investigate protein expression and functional regulation. From being able to study only dozens of proteins, state-of-art MS proteomic techniques are now able to identify and quantify ten thousand proteins [5].

Cite this article : Vu L A, Cam Quyen P T, Huong N T. **Data mining in mass spectrometry-based proteomics studies**. *Sci. Tech. Dev. J. – Engineering and Technology;* 2(4):258-276.

MS is used to measure the mass-to-charge (m/z) ratio of molecules. However, the molecules must first be electrically charged and transformed into a gas phase due to electromagnetic fields (Figure 1). Electrospray ionization is a commonly used method for the ionization of molecules. However, other methods are increasingly popular, including matrix-assisted laser desorption/ionization (MALDI) and surface-enhanced laser desorption ionization (SELDI). Once the molecules have been transformed into a gas phase, their m/z ratios are measured by their motion in an electric or magnetic field, this occurs in mass analyzer. There are different types of mass analyzers, including quadrupole systems, time of flight, ion trap and fourier transform. Each of these systems has different strengths and weaknesses, such as the m/z value range that can be detected and the mass spectrometric resolution. Once measured, the m/z values are displayed as mass spectra, describing the molecules present through the peaks at corresponding m/z values[6].

In recent years, with advances in instrumentation and detection techniques, MS has been applied more widely in various areas, including pharmacology and biomedical practice. However, the more the sensitivity, accuracy and performance of MS analysis are improved, the more the quantity, dimensionality, and complexity of the data sets generated by MS have increased significantly. In order to interpret this huge amount of data efficiently, there is growing interest in applying informatics technology based on data mining algorithms to meet current demand.

The aim of this article is to give a brief overview how data mining algorithms could help processing complex MS-based proteomics data, to provide a valuable molecular insight into different biological specimens, and make MS techniques more versatile and translatable in solving biomedical problems. First, we introduce the field of data mining in proteomics studies and highlight the essential concepts. Then specific implementations of data mining algorithms are reviewed, ordered by the steps in a typical workflow. Thereafter, challenges of standardization databases and softwares availability are mentioned. Finally, application of MS-based proteomics in biomedical, as well as limitations and future perspectives of this approach are discussed.

MS-BASED PROTEOMICS AND DATA MINING

In proteomics, mass spectrometry is increasingly used in studies because of its specific and high performance capabilities. The most commonly used method of MS for protein identification is known as the "bottom-up" approach. Using this approach, the molecules measured are peptides generated by the enzymatic digestion of peptides in a sample. The resulting spectra of the fragmented peptides, known as MS tandem spectra (MS/MS), are generated where the peaks describe the amino acids present in the peptides. However, this only provides the identifications of the peptides present in the sample after enzymatic digestion and, therefore, it is still necessary to work from the known peptides to predict which proteins were originally present in the sample. The "bottom-up" approach contrasts with the "top-down"approach, for which MS is used to directly analyze undigested proteins, by ionization and dissociation of intact mass spectrometer proteins. This approach may be more specific than "bottom-up", but it has higher experimental requirements and requires more complex tools to be applicable to a global analysis[7].

Data mining techniques have been widely used to analyze data from many areas of biology; in particular, various machine learning methods have been applied to data generated by analytical techniques of genomics, transcriptomics and metabolomics to classify unknown samples and identify genes relevant to the state of the disease. Currently, similar methods are applied in the field of proteomics, and more specifically, in the analysis of data generated by MS[8]. In many studies, MS generates large data files containing lists of many peaks. Implementing data mining methods therefore is necessary for the identification of proteins related to the interested peaks and to compare the samples. In most cases, the analysis of MS data follows the paths summarized in Figure 2.

Basic Steps in MS-Based Proteomics Data Mining

As mentioned above, the use of MS yields a huge amount of data, where the number of characteristics (peaks) is larger than the number of samples. MS data are typically composed of hundreds to thousands of protein peaks. These data can not be analyzed manually or managed by normal data mining tools. In search of adequate tools to analyze available data and extract useful information, proteomics scientists are increasingly rely on advanced data mining techniques that can address issues such as the wide dimensionality and limited data sets. These advanced techniques include machine learning and artificial intelligence. Current practices of data mining in MS-based proteomics include following steps: Firstly, data was

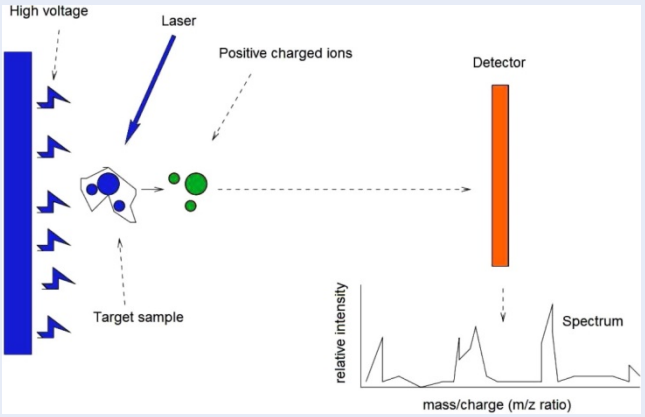

Figure 1: Illustration of mass spectrometry process[6].

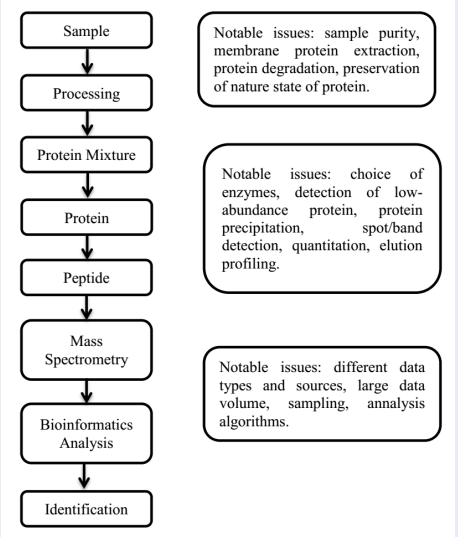

Figure 2: Typical process of MS based proteomics studies.

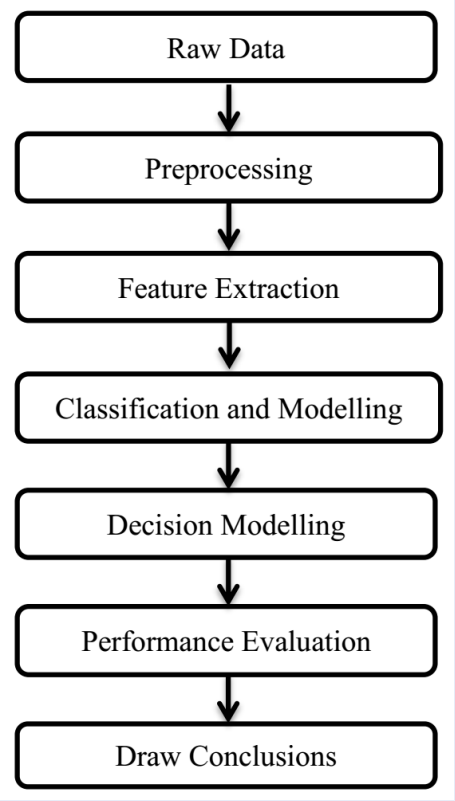

Figure 3: Typical flowchart of the critical steps in data mining.

modeling using identified peaks by pre-processing and feature selection. Then, data sampling was careful applied to process the typical small sample size of MS data. Lastly, the performance of generated model was evaluated.

The critical phases mentioned above must be carefully treated by proteomics researchers to get correct and robust decision models. The steps are repeated iterative and changes are made to explore different aspects of the data. Figure 3 describes the typical flowchart in data mining.

Pre-processing

Raw data obtained from MS is often disturbed. The purpose of pre-processing is to improve data quality. The results of the classification algorithms will be misleading and will be negatively affected when data quality is poor. Therefore, data pre-processing are

crucial in the analysis of untreated proteomics data. Many published studies used the software provided by the manufacturer for pre-processing [9]. The software detects the positions and intensities of the proteins in the samples and performs important pretreatment steps such as first subtraction, intensity normalization, alignment, and peak detection. The criteria specified by the operator are used to filter the peaks. To date, there was no study has been conducted to compare the effectiveness of available reduction techniques and it seems that researchers are optimizing this step in a heuristic way that works best with their own datasets. Although most studies focus on to cut low-frequency noise in spectra, many attempts have also been made to characterize and subtract high-frequency noise components [10].

With the reduction of the baseline complete, standardization is the next step. Since a peak in a spectrum only describes the relative amount of a protein, normalization is performed to make sure meaningful comparisons between the spectra. After preprocessing, obtained peaks are further analyzed by extra size reduction techniques.

Feature Extraction

Feature extraction involved selection of spectra to find peaks (or characteristics) and is usually done by grouping. According to this, each group of m/z points entering a container is described by a value, such as its average or greatest intensity. Subsequently, the characteristics of these containers, such as m/z tray position and estimated intensity, are used as features for data mining. The containers can be independent or overlapping, of equal or adaptive size. By changing the size of containers and grouping method, the researcher can empirically optimize the feature extraction process. Selecting and using only those features important to the modeling process makes the entire data extraction process more accurate and efficient. The data mining algorithm would be faster for a set of data composed of smaller and more significant peaks and give simple and meaningful results. Therefore, it is essential to eliminate irrelevant and redundant features to create better models. However, it must be kept in mind that the function selection process does not always guarantee a correct selection of peaks for the classification problem. Therefore, it is necessary to validate the selected functions when increasing the size of the data sample [11].

Advances in machine learning have led to develop automatic function selection tools. There are two types of feature selection techniques today. A first type analyzes each function independently and removes the functions one by one depending on the relationship between the function and the goal [12]. Selecting independent features is a simple, straightforward and fast process. However, it often happens that a group of entities is more correlated with the desired output. Therefore, the hypothesis of independence of the characteristics can be rather limiting. To overcome the above limitation, some techniques have been proposed to select characteristics in which characteristics are analyzed in groups/subsets [12]. The correlation between the groups of characteristics is considered to the destination output. Although the process requires a lot of computation, it exploits the interrelations of important characteristics when it discovers critical information generally lost during the analysis of independent characteristics.

Classification and Data Modelling

Humans and animals gain the ability to learn through interaction with the environment. Data learning has been an area of interest for researchers in statistics and computer science [1]. Machine learning algorithms can infer a sample of data through familiarization and repeated interaction with the data. These algorithms vary in their training techniques, their end goal and correspond data. A wide range of algorithms for machine learning has been developed. Some popular algorithms are summarized and compared in Table 1.

A learning process normally includes the task of learning and developing rules or functions from the data set provided by the samples. The development of mathematically precise rules and functions to describe data is called data modeling. The expanded model identifies the properties of the different classes and what separates them for proper classification. In the next phase, called the test, the developed model is validated with new observations to verify that the model produces accurate results. The learning phase and the model estimation are implemented and described using different learning methods or algorithms.

There are two types of machine learning algorithms: supervised and unsupervised. In supervised learning (also called "learning with a teacher"), there is a earlier knowledge of the class to which each case belongs (sample). The training data set includes the input values and the associated output classes (provided by the master). During the learning phase, learning data is used to decide how entities will be selected, weighted, and combined to distinguish classes. The test phase involves the application of weighted characteristics to classify new test data whose class is unknown and which the decision model has never seen

before. Therefore, the goal of the classification methods is to create dataset models by hand and use this model to classify new samples. The learning process would create of a model so that model predictions come close to the desired goal. If the model is able to correctly classify new data, we have reason to believe that it is a good model[5]. The most widely used supervised learning algorithms are Bayesian classifiers, Rule-based learners and Support Vector Machines.

In unsupervised learning (also called "learning without a teacher"), the group to which each sample belongs is unknown or ignored and the data is grouped according to similarity measures. The learning process does not involve a teacher and the algorithm must identify the models in the data. Unsupervised learning can often lead to more than one possible solution. Artificial Neural Networks are typical examples of unsupervised learning used in studies analyzing mass spectroscopy data[13].

In both learning techniques, the goal is to predict (classify) or describe data by developing data models, which are then used to classify or describe new cases. If the data has only two or three characteristics, it would be easy to classify the data. However, developing models can be a daunting task if there are many features to analyze. Large data is not only difficult to visualize, but all possible combinations must be taken into account through comprehensive research techniques during the model training phase. A large number of dimensions with very few samples leads to what is often called over-regulation or over-training. Excessive regulatory models can not generalize and classify new cases with the desired accuracy.

Data Sampling

One of the major challenges in applying machine learning algorithms to biomedical data is the validation of an experimented model with new test data. The decision modeling process requires that the model be developed by training a given set of data (training set), followed by validation of the model in another dataset never seen before during the training (test set). The obvious way to handle this is to split the data into tests and series before constructing the model using stratified random sampling. However, medical data is often very difficult and expensive to acquire. As a result, there are not enough cases available to be divided into subsets of train tests. In addition, the disturbance inherent in most medical data and the complex relationships between characteristics require a sample of sufficient size to efficiently model the data. In addition, the size of the test set controls statistical power and confidence in the developed decision model. As a result, sophisticated sampling strategies are needed to capitalize on the available data.

Cross-validation is one of the most widely used data resampling methods to assess the generalization ability of a predictive model and to prevent overfitting[19]. The data is randomly divided into two sets. The decision model is formed in the first and tested in the second. This random division process is repeated several times to reduce the selection bias. The average of all test estimates provides the average error of the model. If the dataset used for the training is too small, the model may not be able to predict the test cases well. A small series of tests may not result in a validated classifier and may generate a high error rate. As a result, different train test reports are examined (for example, 50-50%, 75-25%, etc.) with cross-validation. A common implementation is cross-validation of k-folds. The data is partitioned into k-disjoint sets. The training of the classifier is performed in sets k-1 and tested in a set of remaining data. This is done for all k-subsets producing k patterns and the estimated error will be the average of the error rates k. For example, a 10-fold cross-validation divides the data into 10 groups. Nine groups are used for training and tests are performed in the left group. This is repeated 10 times until each of the 10 groups has served as a test group. The average test error of the 10 groups is the estimation of the final test error and gives a rough idea of the quality of the model for the classification of the data.

To conclude, effective classification methods by MS data could contribute to early and less-invasive diagnosis and also facilitate developments in the bioinformatics field. As protein MS data growing with data volume becomes complicated and large; improvements in classification methods in terms of classifier selection and combinations of different algorithms and preprocessing algorithms are more emphasized in further work[20].

Performance Assessment

The last phase of the data mining process is the assessment of the models developed by the previously described machine learning algorithms. The accuracy of the classification is calculated by taking the ratio of the number of correctly classified samples to the total number of samples in the test data. However, when the prevalence of a particular class is higher than that of another class, the majority class will distort the result. In such a scenario, measuring accuracy can be misleading[21].

Table 1: Comparison of some commonly used machine learning algorithms

Algorithm	Descriptions	Advantages and disadvantages
Bayesian classifiers [14]	Based on Bayes' theorem with an assumption of independence among the predictors, making it particularly useful for large datasets. Despite its simplicity, Bayesian classifiers often works surprisingly well and is widely used because it often goes beyond the most sophisticated classification methods.	Fast and easy to implement. This method is suitable fordatasets with missing values. The main disadvantageis it assumes attributes are independent of each other.
Rule-based learners [5]	Rule-based learners is a computer term used to understand any machine learning method that identifies, learns, or develops "rules" for storing, manipulating, or applying. The defining feature of a rule-based machine student is the identification and use of a set of relational rules that collectively represent the knowledge acquired by the system.	The rules generated are easilyreadable, and is suitable for identification of putative biomarkers, however there isa possibility of overfitting.
Decision trees [15]	The decision tree methodology is a commonly used data extraction method for establishing classification systems based on multiple covariates or for developing prediction algorithms for an objective variable. This method classifies a population into branch-like segments that build an inverted tree with a root node, internal nodes, and terminal nodes.	The output from decision trees can be easily interpreted, but it does depend on the algorithm used and the complexity of the tree generated. It is also well suited to datasets with missing values.
Random forest [16]	Random forests are common learning methods for classification, regression and other activities that work by building a multitude of decision trees at the time of training and leaving the class that is the class mode (classification) or predicting the mean (regression) of individual trees. Random forests correct the habit of supercharging decision trees in their training set.	This method is efficient on large datasets and can handlelarge numbers of attributes, however it is not verysensitive to outliers.
Support Vector Machines (SVMs) [17]	SVMs are machine learning algorithms that analyze the data used for regression and classification analysis. Using a set of learning examples, each of which falls into one of two categories, an SVM learning algorithm constructs a model that assigns new examples to a category or another, making it a non-probabilistic binary linear classifier. An SVM model is a representation of space point examples, assigned so that the examples of distinct categories are divided by the largest possible gap. Thus, the new examples are assigned to this same space and should belong to a category based on the side of the hole in which they are located.	SVMs uses kernels to learn complex functions, however they are very slow and there are multiple parameters to be chosen by the user.
Artificial Neural Networks (ANNs) [18]	ANNs are computer models composed of several simple processing units that communicate by transmitting signals via a large number of weighted connections. Like human brains, neural networks also consist of treatment units (artificial neurons) and connections (weights) between them. The processing units carry the incoming information on their outgoing connections to other units. The "electrical" information is simulated with specific values stored in the weights that allow these networks to learn, memorize and create relationships between the data. A very important feature of these networks is their adaptive nature, in which "learning by example" replaces "programming" to solve problems. After training, ANNs can be used to predict the outcome of new independent input data.	ANNs use a multilayer perceptron to learn complex functions. The output of ANNs are not able to be read and the training of the model can be very slow.

In samples of two classes, there are four possible outcomes when testing the decision model. These are real positive results, true negatives, false positives and false negatives. Sensitivity (actual positive rate) is the ratio of the number of correctly graded positive samples to the total number of positive samples. High sensitivity is highly desirable in medical diagnosis, where the impact of a prediction incorrectly indicates that a sick person is in good health. The false positive rate is the probability that a healthy person is wrongly classified as a sick person (so-called specificity). High specificity is desirable when a false alarm leads to unwanted tests and elaborate treatments. Ideally, for a perfect classification, sensitivity and specificity must be equal to 1 (100%). Clinically acceptable sensitivity and specificity depend on the application. Several studies reported their results using sensitivity and specificity as performance indices[22]. The main limitation of the use of sensitivity and specificity as the only indices of evaluation is their dependence on the prevalence of class and decision threshold. It is therefore difficult to directly compare the results of reported studies using only the sensitivity and specificity measures.

Standards and Databases

Driving by improvements in speed and resolution of MS in the field of proteomics, which involves the large-scale detection and analysis of proteins in cells, tissues and organisms, continues to expand in scale and complexity. There is a resulting growth in datasets of both raw MS files and processed peptide and protein identifications. MS-based proteomics technology is also used increasingly to measure additional protein properties affecting cellular function and disease mechanisms, including post-translational modifications, protein–protein interactions, subcellular and tissue distributions. Consequently, biologists and clinicians need innovative tools to conveniently analyse, visualize and explore such large, complex proteomics data and to integrate it with genomics and other related large-scale datasets. The main challenge for big data mining then would be how we can achieve a transition from association study to causality study. From this point of view, standarization of data provides a new way for system-wide study and could play a key role in such a transition in big-data era.

For MS-based proteomics, the Standards Initiative Proteomics of the Human Proteome Organization (HUPO-PSI) is an organization that plays a pioneering role in development of standard terminologies, file format and minimum requirements for MS based proteomics data[23]. The most common formats are: (i) mzML, which stores raw MS data, as well as the peak list of the processed spectrum[24]; (ii) mzIdentML, which has information on peptides and proteins obtained from MS data[25]; and (iii) mzQuantML, which has detailed quantitative information[26]. While mzML and mzIdentML have been applied for a long time, standards for quantitative data are still rarely applicable. This is mainly due to the lack of quantitative standard support from popular analysis tools.

The need for large and easily accessible data repositories is essential for the benefit of proven data exchange in other research areas such as genomics and transcriptomics. Public databases will integrate data obtained from many laboratories and as a result, data can be analyzed by applying a new tools or new algorithms. Therefore, some public databases have been developed, for example, the PRIDE Archive[27], GPMDB[28], PeptideAtlas[29], Massive[30] and the Human Proteome Map[31] (Table 2). These databases are designed to provide a user-friendly interface, featuring graphical navigation with interactive visualizations that facilitate powerful data exploration in an intuitive manner. Moreover, they also offer a flexible and scalable ecosystem to integrate proteomics data with genomics information, RNA expression and other related, large-scale datasets.

Because of the nature of biological data, conducting research in life science to some extent has to change its style in the era of big data, e.g., from academic exploration individually to more cooperative study in systematic, standardized and pipelining ways. The main challenges here could be to establish interoperable databases, make sustainable tools available to the research community, create tool development centers, construct resources and infrastructure, such as cloud computing to serve the huge amount of researches, generate standards, vocabularies and ontologies of big biological data, develop new systems of infrastructure and tools, and obtainbuy-in from the scientific community, such as cloud service. Clearly, aforementioned challenges can be solved in a more engineering manner, and a well-designed experiment system matching some systematic, standardizing data processing pipe-line will be an important factor for a successful study.

Softwares and Tools

Many computer programs have been developed for the analysis of MS-based proteomics data[32,33]. Besides the available software, many useful tools have

Table 2: MS-based proteomics database (available on internet, accessed by November 15, 2019).

Databases	Descriptions	URL Address
Chorus	A Sustainable Cloud Solution for Mass Spectrometry Data	https://chorusproject.org/
GPMDB	Open source system for analyzing, validating, and storing protein identification data	http://gpmdb.thegpm.org/
Human Proteome Map	An interactive resource to the scientific community by integrating the massive peptide sequencing result from the draft map of the human proteome project.	http://www.humanproteomemap.org/
jPOST	Japan ProteOme STandard Repository/Database	https://jpostdb.org/
MassIVE	A community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data	https://massive.ucsd.edu/
PeptideAtlas	A resource for target selection for emerging targeted proteomics workflows	http://www.peptideatlas.org/
PRIDE Archive	The proteomics identifications database	https://www.ebi.ac.uk/pride/archive/
ProteomicsDB	A protein-centric in-memory database for the exploration of large collections of quantitative mass spectrometry-based proteomics data.	https://www.proteomicsdb.org/

also been reported in the programming languages of BioPython[34], BioJava[35] and BioPerl[36]. Furthermore, various bioinformatic tools used to data conversion, quantification, visualization and identification of peptides/proteins have also been noted (http://tools.proteomecenter.org/; http://wiki.nbic.nl/index.php/ProteomicsTools; http://www.msutils.org/wiki/pmwiki.php/Main/SoftwareList). Some tools has a role as components of larger platforms to form master data processing processes (Table 3).

Alternatively, there is a large number of software packages for the analysis of quantitative proteomics data, available in both commercial and free distributions (Table 4). This list is intended to serve as a useful reference and guide to the selection and use of different pipelines to perform quantitative proteomics data analysis depending on the type of instrument, method or platform used.

Some excellent reviews on existing software are available, such as[37–39]. For instance, in[37], three different software platforms, Progenesis, MaxQuant and Proteios were compared for peptide-level quantification in shotgun proteomics using a spike-in peptide data set with two different spike-in peptide dilution series. The performance of the software workflows was evaluated with different metrics, including harmonic mean of precision and sensitivity, mean accuracy, coverage and the number of unique peptides found[37]. The comparison suggested that Progenesis performed best, but a noncommercial combination of Proteios with imported features from MaxQuant also performed well[37]. Algorithms, such as peak picking and retention time alignment, usually included within a quantitative shotgun proteomics label-free workflow, have also been evaluated and compared separately[40–42]. While such separate comparisons are interesting and the evaluation by Chawade *et al.*[37] is informative, a thorough comparison of multiple workflows on protein level is still missing, especially in terms of differential expression analysis. Here, some of the most popular free software applications applied to proteomics profiling biomarker discovery and cluster analysis will be mentioned.

MaxQuant

MaxQuant is a quantitative proteomics software package designed for analyzing large-scale mass-spectrometric data sets, developed by the Max Planck Institute of Biochemistry[43]. It supports all main labeling techniques like SILAC, Di-methyl, TMT and iTRAQ as well as label-free quantification. MaxQuant is a comprehensive software that performs several analysis steps: a) Peak detection and scoring of peptides: MaxQuant corrects systematic inaccuracies of measured peptide masses and corresponding retention times;b) Mass calibration: It detects mass and intensity of peptide peaks in MS spectra and assemble them into 3D peak hills over m/z retention time

plane, followed by filtration to identify isotope patterns; c) Database searches for protein identification: Peptide and fragment masses (in case of an MS/MS spectra) are searched in an organism specific sequence database, and are then scored by a probability-based approach termed peptide score d) Protein quantification: High mass accuracy is achieved by weighted averaging and through mass recalibration. The software is written in C# and freely available on http://www.coxdocs.org.

PEAKS

PEAKS Studio performs LC-MS/MS data analysis and statistics according to the experimental design. Following the identification of peptides with MS/MS spectra, the resulting peptide sequences are used to determine the original protein components of the samples. PEAKS studio main features include: a) Peptide/Protein identification: de novo sequencing, database search, post-translational modification (PTM) search with 500+ modification, sequence variant and mutation search; b) Protein quantification in complex biological samples: Label-free, label-based: TMT (MS2, MS3)/iTRAQ, SILAC, ^{18}O labeling, ICAT and c) Supporting fragmentation types: CID, HCD, ETD/ECD, EThcD, IRMPD, and UVPD. PEAKS Studio is licensed commercially by Bioinformatics Solutions Inc.[44] and a free trial available on http://www.bioinfor.com/download-peaks-studio/.

OpenMS

OpenMS is an open-source software C++ library for LC/MS data management and analyses, developed at the Free University of Berlin, the University of Tübingen, and the ETH Zurich[45]. It provides a large number of tools (more than 200) to analyze proteomics datasets, in the form of command lines. These tools can perform the following tasks: a) Import, export and conversion of vendor formats and several open community-driven XML formats; b) Preprocessing of spectra: Filtering based on various properties, Peak picking, Baseline and noise filtering; c) MS2 spectrum identification: Support for third-party peptide search engines, own customisable and extensible basic search engine, indexing of peptides in custom protein databases with SeqAn, statistical validation via posterior error probability and FDR/q-value calculation, combining results of different peptide search engines with ConsensusID; d) Visualisation of spectra (on all MS levels), features and peptide identifications in our TOPPView; e) Finding RNA and protein-protein crosslinks; f) Identification of phosphorylation sites with Luciphor. OpenMS is free software and runs under Windows, macOS and Linux.

Applications in the Post-Genomic Era

Sample classification from protein mass spectra

Application of data mining suggests a novel algorithm for pattern classification from protein mass spectra, which is a slight variation of the "nearest centroid" classiffication, the proposed "Peak Probability Contrast" (PPC). It is first described in the study of Tibshirani *et al.*[46]. Briefly, PPC works by extracting peaks from each spectrum, and then determining the optimal peak height split point for discriminating between the classes at each site. Then it computes the proportion of spectra in each class with peak heights above the split point and uses these proportions to build a nearest centroid classifier. In particular, when applied to spectra from both diseased and healthy patients, the PPC technique provides a list of all common peaks among the spectra, their statistical significance, and their relative importance in discriminating between the two groups. Compared to other statistical approaches for class prediction, this method performs as well or better than several methods that require the full spectra, rather than just labeled peaks. The algorithm consists of six sequential steps,as shown in Figure 4. The development of this method, so as to find a relative small number of peak clusters for class prediction, is expected to facilitate the identification of biologically significant and relevant proteins for specific biological states, such as tumor development and progression[47–50].

Clustering mass spectra peak-lists

Data mining algorithms are also applied to proteomics data, in an attempt to group proteins based on their spectral similarities[51]. Notably, clustering validation methods are used to find the clustering method which most faithfully captures the underlying distribution of the samples. These work also show that the application of clustering algorithms in proteomics can assist in (a) identifying peak features responsible for categorizing samples, (b) formulate hypotheses on the possible function and role of unidentified proteins and (c) reveal proteins which act jointly as biomarkers in a concrete biological state[52–54].

The proteomics data on which clustering is performed are the mass spectra peak-lists (not the raw mass spectra) which derive from a mass spectrometer. In order to apply cluster analysis, these peak-lists are represented as vectors in a multidimensional space, where each vector element is a feature of a specific mass (e.g., its intensity) or a group of masses. To deal with the

Table 3: Programs and tools for MS-based proteomics data mining (available on internet, accessed by November 15, 2019).

Programs/Tools	Descriptions
Input Processing/Data Handling	
InsilicosViewer	viewer for displaying mzXML data
massWolf	Waters MassLynx raw-to-mzXML converter
mzBruker	Bruker raw-to-mzXML converter
mzWiff	ABI/MDS Sciex Analyst raw-to-mzXML converter
MzXML2Search	mzXML to SEQUEST dta, MASCOT generic and Micromass pkl converter
mzXMLViewer	viewer for displaying mzXML data
RAMP	mzXML data parsers
readmzXML	mzXML parser based on RAMP
ReAdW	Thermo Xcalibur-to-mzXML converter
validateXML	mzXML validation script
Database and Spectral Library Search	
Comet	an open source tandem mass spectrometry (MS/MS) sequence database search tool
SpectraST	a spectral library building and searching tool designed primarily for shotgun Proteomics applications
X!Tandem	open source Proteomics software that attempt to find the best sequence model for a given MS/MS spectrum of a peptide
Probability Assignment and Validation	
iProphet	validation of distinct peptide sequences; can also combine search results of multiple search engines
PeptideProphet	validation of PSMs made by tandem mass spectrometry (MS/MS) and database searching; probabilities are assigned to the peptide identifications made by programs like SpectraST, Comet or X!Tandem
ProteinProphet	statistical model for validation of peptide identifications at the protein level
Protein Quantification	
ASAPRatio	Automated Statistical Analysis on Protein Ratio
Libra	Four channel quantification software
XPRESS	software to calculate the relative abundance of proteins in the sample
Miscellaneous	
Cytoscape	platform for visualizing and integrating molecular interaction networks
Pep3D	viewer for LC-MS and LC-MS/MS results
PeptideSieve	A tool for predicting proteotypic peptides
PepXMLViewer	analyze unassigned but high quality spectra
Prequips	Prequips is an interactive software environment for integration, visualization and analysis of LC-MS/MS data
QualScore	peptide dataset interaction
SpecArray	tool for analyzing and comparing LC-MS runs
SubsetDB	FASTA manipulation

Table 4: Software packagesfor analysis of quantitative proteomics data (available on internet, accessed by November 15, 2019).

Software	Technique	Type of data	Instruments	Input files	Distribution
MaxQuant	SILAC, ICPL, Label free, iTRAQ, TMT	MS^1/MS^2	Orbitrap, FT-ICR (Thermo)	.raw (Thermo)	Free
OpenMS	iTRAQ, SILAC, Label free	MS^1/MS^2	Any via mzXML or mzML	.dta, mzData, mzXML, mzML	Free - Open Source
Proteios	TRAQ, TMT	MS^1/MS^2	Any via mzML	mzML	Free - Open Source
Census	^{15}N, SILAC, iTRAQ	MS^1/MS^2	Any via mzXML	MS1/MS2, DTASelect, mzXML, pepXML	Free
VIPER	^{18}O, ICAT	MS^1/MS^2	Any via mzXML	.pek, .CSV, .mzXML, .mzData, .raw (Thermo)	Free - Open Source
BioWorksTM	^{15}N, SILAC, iTRAQ	MS^1/MS^2	Any via mzXML	MS1/MS2, DTASelect, mzXML, pepXML	Commercial
PEAKS® Q	iTRAQ, SILAC, Label free	LC-MS	Any via mzXML	mzXML, pepXML	Commercial
Progenesis LC-MSTM	ICAT, SILAC, iTRAQ	MS^1/MS^2	Any via mzXML or mzML	mzML, mzIdentML	Commercial
ProQuant	Label free image recognition	LC-MS	Thermo, Waters	OpenRaw	Commercial

SILAC: Stable isotope labeling with amino acids in cell culture; ICPL: Inductively coupled plasma; iTRAQ: Isobaric tags for relative and absolute quantitation; TMT: Tandem mass tag; ICAT: Isotope-coded affinity tag.

high dimensionality of the generated peak-list vectors mass "containers" (i.e., contiguous nonoverlapping regions in the m/z axis) can be defined before analyzing the samples of an experiment. The process of binning performs dimensionality reduction by grouping consecutive masses and selecting a representative feature of those masses for each group(e.g., mean, log, maximum intensity value). Moreover, one can preprocess the peak-lists vectors by performing scaling or normalization.

The suggested clustering algorithms for these data are the hierarchical as well as the k-means clustering. For a better comprehension of the clustering results several visualization methods are also exploited (i.e., dendrograms, heatmaps and cluster sets). In the clustering results that derive from this method, not only well separated protein clusters can be easily discerned, but also the spectral containers that are most influential in partitioning the proteins into clusters. Furthermore, the presented method offers the option of integrating the identification results for the proteins – members of each cluster, as well as their Gene Ontology annotation [55]. By exploiting both the identification and the Gene Ontology classification information for most proteins in each cluster, one can attempt to infer the role of unidentified proteins. This can be based on the already known functions of the proteins which are identified with high confidence and are found to be close to unidentified proteins in the same cluster.

Protein-protein interactions prediction using association rules

The work by Kotlyar *et al.* [56] are first attempts that use association rules not only to discover protein-protein interactions, but also to predict whether a given pair

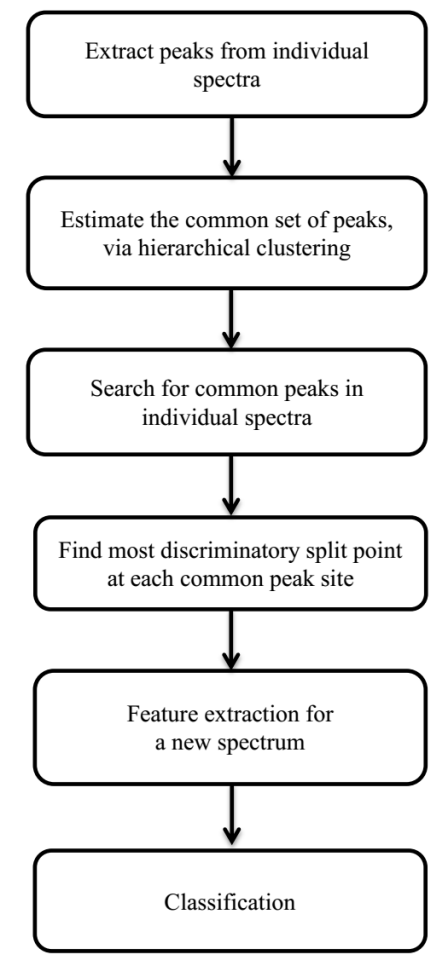

Figure 4: Flow chart of the Peak Probability Contrast classification analysis[46].

of proteins interacts. Predicting interactions with association mining can be viewed as a classification problem where the part of the rule consists of a single item only, the class variable. After the application of association mining, the rules are ranked according to a measure of "interestingness" (e.g., confidence, support) and used for prediction as follows:a given protein pair is predicted to interact if its attributes include the items of any rule.

The presented approach is based on the idea that both direct and indirect evidence (e.g., data coming from experimental and computational methods) could be used to predict interactions reliably and on a proteome-wide scale. In particular, datasets that consist of interacting and non-interacting protein pairs annotated with different types of evidence are first constructed. Then, with the help of association rules, patterns that discriminate the interacting and the non-interacting proteins are detected. Lastly, using these patterns the prediction of interactions is achieved, assigning a confidence level to each interaction[57–59].

To conclude, with this approach, different types of evidence for interaction are integrated in order to create rules that act as a classifier for new interaction pairs. Thus, association mining is used to search thoroughly in large datasets for predictive patterns. However, to evaluate the performance of this method and strengthen its applicability, it is important to incorporate additional evidence, perform testing and validation using already known interactions from specific organisms and compare the results to those of other interaction detection methods[60].

Biomarker discovery

Data mining can also be useful in determining which proteins, from MS data, could be used as biomarkers to differentiate between samples of different classes[61]. Table 5 includes information from investigations on the application of data mining on mass spectrometry data for the identification of the most suitable biomarkers, based on factors such as the ability to test for proteins in a clinical setting; this includes both identified proteins and mass spectral peaks as biomarkers. Further analysis, following identification of peptides or proteins as putative biomarkers, are then required, as it may be that the proteins identified would not actually be suitable for use as biomarkers. For example, body fluids such as urine and serum (blood) are regarded as being most suitable fluids to search for biomarkers because they are easier to obtain for assessment purposes during diagnostic tests and treatments. Also, blood is pumped around the body by the circulatory system and bathes cells, tissues, and organs, thus carrying putative protein biomarkers around the body before being processed by the liver and filtered by the kidneys into urine[62]. Table 5 also shows that the number of possible biomarkers identified varies greatly between studies, due to differing complexities of data, for example, Ralhan *et al.*[63] identified only three m/z values as biomarkers, and Fan and Chen[64] formed a panel of five biomarkers. This is in comparison to Ryberg *et al.*[65] and Bloemen *et al.*[66] who identified 41and 40 putative biomarkers, respectively. Some found biomarkers that had previously been identified; this is both useful as support for the previous investigations, and as some validation to the methods being newly applied to the area. Other investigations identified biomarkers that work specifically well together and so formed panels of markers.

Table 5: List of studies that used data mining on mass spectrometry data for biomarker identification.

Aim of paper and dataset	Methods of quantitation, data mining and evaluation	Identification of biomarker
Classification of prostate cancer samples Serum from 19 patients with bone metastases and 19 without[67].	Mascot and novel spectra analysis implement using Leave-one-out cross-validation	Multiple biomarkers identified.
Classification and identificationof biomarkers of heart failure. Training set - 100 heart failure (HF)&100 healthy control. Test set-32 HF, 20 control[68].	Background subtraction and feature extraction. SVM tested	18 putative biomarkers identified.
Identification of head and neck cancer biomarkers. Five sets of four samples plus control for each set[63].	Bayesian classifiers used for biomarker panel analysis using 3-fold cross validation.	Panel of 3 best biomarkers identified.
Ovarian cancer biomarker discovery and classification. 37 patients with papillary serous ovarian cancer and 35 controls[69].	Quantification using mzMine. SVM tested.	Accuracy of 97.2% using a combination of nonlinear SVM with an SVM-based feature selection method. Average of 38 features identified as putative biomarkers using 4 different methods.
Identification of biomarkers for prostate cancer. 179 adenocarcinoma of the prostate and 74 benign[70].	Feature selection, baseline correction, and normalization. Novel feature selection method: Extended Markov Blanket using 10-fold cross validation.	26 peaks were identified as possible biomarkers.
Biomarker panel development for breast cancer. 40 plasma samples from patients with breast cancer, 40 samples from healthy[64].	Label-free proprietary protein quantification Artificial Neural Network with a test set of 40 plasma samples from patients with breast cancer and 40 samples from healthy controls.	Two best protein panel of biomarkers identified, containing 7 proteins.
Identification of amyotrophic lateral sclerosis (ALS) biomarkers. 100 ALS, 18 multiple sclerosis, 53 Alzheimer's disease, 29 other neurologic disease, and 41 healthy control subjects[65].	Using a biomarker panel of 41 mass peaks between 1.5 and 35kDa. Rule-based Learner algorithm, using 10-fold cross-validation.	Biomarker panel used for classification and a putative biomarker identified.
Classification using exhaled proteins as potential biomarkers for asthma. Exhaled breath condensate. 26 well controlled asthma, 14 partially or not controlled and 30 healthy[66].	SVM used for classification.	100% classification accuracy. This was lowered to 73% when the diagnosed and non-diagnosed asthma samples were treated as separate classes.

Continued on next page

Table 5 continued

Potential urine protein biomarkers for kidney transplantation dysfunction. 264 biobanked urine samples with matched biopsies[71].	Selected reaction monitoring method (SRM)	Optimize and detect urinary peptides from 67 proteins.
Candidate biomarkers for hepatocellular carcinoma. Serums from 205 patients [72].	LC-MS/MSuntargeted proteomic analysis	11 new biomaker candidates discovered
Biomarkers for idiopathic pulmonary fibrosis (IPF). 97 differentially expressed proteins (38 upregulated proteins and 59 downregulated proteins)[73].	STRING software, a regulatory network containing 87 nodes and 244 edges was built,	4 proteins were found as specific IPF biomarkers
A protein biomarker panel has been developed specific for diabetic kidney disease. 572 patients with significant correlations with the current measures of disease[74].	Bayesian classifiersusing 3-fold cross validation.	Five proteins were significantly associated with diabetic kidney disease
Biomarker for early diabetic mellitus (DM) discovery. 942 proteins in healthy volunteer urine and 645 proteins in the DM patient urine were identified with label-free semi-quantitation[75].	Gene ontology and pathway analysis	In total, 344 proteins were significantly associated with DM.
Biomarker for survival in Non-small-cell lung carcinoma (NSCLC)patients with immunotherapy 47 patients with advanced stage NSCLC[76].	Machine learning	Serum proteomic signature may serve as a biomarker for survival outcome in patients with NSCLC, including patients undergoing immunotherapy

In Fan and Chen[64], different panels of biomarkers were compared and those markers that worked best together werei dentified. The development of panels of biomarkers is useful as using multiple biomarkers may reduce false positives as itremoves dependence on individual proteins, and allows proteins that are detected for different diseases to be useful. To discriminate between samples, the majority of the studies applied data mining to only the peaks from the mass spectrometry data that correspond to peptides. To facilitate the development of diagnostic assays and/or inform the underlying biology at a molecular level, peptide biomarkers require further investigation.

Literature mining and pathway analysis

Data mining has been shown to highlight important peptides/MS peaks, however further analysis is required to determine to which proteins they relate. In the case of data mining applied to quantified proteins, literature mining is also useful for understanding the biological relevance of the proteins identified as potential biomarkers.

It may be important to discover more information about interacting proteins and pathways in which they have arole[77]; by doing this, it can be determined whether the identified proteins may become useful biomarkers and which processes would be measured. Pathway analysis can be used to narrow down, or provide a focus to, the search for biomarkers by determining which pathways they participate in[78] . Literature mining is also essential in discovering more information after data mining has been applied to M Speaks, however identification of the proteins the peaks relate to is first required[70]. Tools such as Ingenuity Pathway Analysis (http://www.ingenuity.com) and DAVID[79] can be used to facilitate literature mining and pathway analysis, or information can be mined directly using such article databases as PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

Limitation

The use of MS in proteomics studies has opened up a number of opportunities; however, there are also technical and conceptual challenges that need to be overcome and these will vary from study to study. First, it is often impractical to produce large numbers of samples due to time and financial constraints. Furthermore, a high-through-put approach is not always required[80]. There is also some difficulty in finding proteins of interest if they are at low abundance, when compared to other proteins within the sample, which is often the case for proteins that may be suitable as disease biomarkers[81]. Moreover, compared to genome studies, current protein studies often involve several cases, or represent discoveries that are only intended to prove the principles. Furthermore, although modern machine learning methods are available, their integration into proteomics analysis is rarely performed[82].

Another limitation of proteomics experiments is the experimental design. Appropriate experimental design is often important for a successful study, including the amount and type of repetition, as well as the randomized principles[83]. A weak test design can not even determine whether the observed difference between samples is due to biotransformation or simply to a technical factor. The high cost of proteomics experiments often leads to poor experimental design, which includes small amounts of inadequate duplication and control, and therefore poor reproducibility[84].

Future Perspectives

Data mining has been successfully applied to proteomics studies, yet it can still be used for other purposes. For example, rule-based learners, as well as being used for classification, are suitable for the identification of biomarkers, as the attributes that are used frequently in rules are those that are better at discriminating between classes. Rule-based machine learning has also been applied to microarray data to develop gene interaction networks based on genes that are used together in rules[85]. This method could be applied to MS data in the same way, generating networks from groups of proteins that appear together in rules.

There are also other methods that were originally developed for transcriptomic data, such as gene set analysis[86], that could be modified for application to proteomics. Furthermore, machine learning could be combined with literature data to include background knowledge, which is not necessary for machine learning to be applied, but could improve the data analysis process[87].

Deep learning is a recent and fast-growing field of machine learning. It attempts to model abstraction from large-scale data by employing multi-layered deep neural networks (DNNs), thus making sense of data such as images, sounds, and texts. The early framework for deep learning was built on artificial neural networks (ANNs) in the 1980s, while the real impact of deep learning became apparent in 2006. Since then, deep learning has been applied to a wide range of fields, including automatic speech recognition, image recognition, natural language processing, drug discovery, and bioinformatics[88].

Peptide identification by fragmentation is a fundamental part of bottom-up mass-spectrometry-based proteomics. Peptide molecules are fragmented with the aid of one of several technique, including collision-induced dissociation (CID), higher energy collisional dissociation (HCD) and electron transfer dissociation, producing a pattern of fragments that is indicative of the amino acid sequence[89]. The frequency with which a peptide backbone bond breaks determines the relative signal intensities in a fragmentation spectrum. Theoretically, the intensities can be calculated by quantum chemistry. However, for molecules as large as peptides, this is too computationally expensive to be practical. Hence, the intensity information contained in fragmentation spectra remains underused in many peptide identification strategies. This problem is an ideal situation to employ deep learning. It can learn the relationship between sequence and fragment abundances based on a large dataset of training examples, without explicit knowledge of the physical mechanisms behind it. Furthermore, the predictive models do not have to remain black boxes, but can be examined with specialized methods that identify features or combinations thereof that are most relevant for making a prediction. While fragment intensity prediction has been attempted before using a variety of methods, they have had limited success[90,91]. Very recently, Tiwary *et al.*[92] present a deep learning method called DeepMass whose accuracy is close to the theoretical limitation. Furthermore, they demonstrate its utility by integrating it into data-dependent acquisition(DDA) and data-independent acquisition (DIA) computational proteomics workflows, and the results suggest that both can benefit from the improved spectrum prediction. With the applications of deep learning in the field of mass spectrometry, we can successfully demonstrated a more accurate method that significantly increases our ability to identify and characterize known biomarkers in a sample.

CONCLUSION

Data mining is a data-driven process where the results obtained largely depend on the analyzed data. The methods employed for feature selection, classification, data sampling, and performance evaluation drive the process and alter final results. Thus, it is recommended to explore more than one technique to make comparisons and better understand the problem in hand. Furthermore, standardized and optimized methodology is essential for achieving accurate measurement and meaningful analysis. This includes all involved steps extending from experimental design, specimen collection, storage and handling, throughout all methods used in the analytical chemistry and MS signal processing. Proper bioinformatics including analytical tools, data storage and sharing are required for data mining and validation.

As proteins are critical biomarkers of disease development and progression - the more we know about them and their relationship to specific diseases, the earlier and more precisely we can intervene. We hope that data mining will enable researchers to characterize disease-relevant protein profiles to build new diagnostic tools and therapeutics. We look forward to continuing the application of machine learning and deep learning to proteomics and other fields, to fulfill the mission of making health data useful.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

AUTHOR CONTRIBUTIONS

All authors contributed extensively to the work presented in this paper. Le Anh Vu and Phan Thi Cam Quyen contributed to acquisition and interpretation of the data. Le Anh Vu drafted the work and Nguyen Thuy Huong contributed critical appraisal and revision for important intellectual content. All authors provide final approval of the revision submitted for publication and agree to be accountable for all aspects of the work as presented.

ACKNOWLEDGEMENTS

This paper is funded by Ho Chi Minh City University of Technology, VNU-HCM, under grant number TNCS-KTHH-2017-12.

REFERENCES

1. Li H, Han J, Pan J, Liu T, Parker CE, Borchers CH. Current trends in quantitative proteomics - anupdate. J Mass Spectrom. 2017;52(3):319–341.
2. Gupta S, Venkatesh A, Ray S, Srivastava S. Challenges and prospects for biomarker research: a current perspective from the developing world. Biochim Biophys Acta. 2014;1844:899–908.
3. Rosenqvist H, Ye J, Jensen ON. Analytical Strategies in Mass Spectrometry-Based Phosphoproteomics. In: Gevaert K, Vandekerckhove J, editors. Gel-Free Proteomics. Methods in Molecular Biology (Methods and Protocols). vol. 753. Humana Press; 2011.
4. Kuhlmann L, Cummins E, Samudio I, Kislinger T. Cell-surface proteomics for the identification of novel therapeutic targets in cancer. Expert Rev Proteomics. 2018;15(3):259–275.
5. Mann M. Origins of mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2016;17(11):678.
6. Nguyen T, Nahavandi S, Creighton D, Khosravi A. Mass spectrometry cancer data classification using wavelets and genetic algorithm. FEBS Lett. 2015;589(24):3879–3886.
7. Wehr T. Top-Down versus Bottom-Up approaches in proteomics. LCGC North America. 2006;24(9):1004–1010.

8. Sun CS, Markey MK. Recent advances in computational analysis of mass spectrometry for proteomics profiling. J Mass Spectrom. 2011;46:443–456.
9. Daniel J, Jürgen R, C. Computational Methods for Understanding Mass Spectrometry-Based Shotgun Proteomics Data. Annu Rev Biomed Data Sci. 2018;1(1):207–234.
10. Tsai TH, Wang M, Ressom HW. Preprocessing and analysis of LC-MS-based proteomics data. Methods Mol Biol. 2016;1362:63–76.
11. Sellers KF, Miecznikowski JC. Feature detection techniques for preprocessing proteomics data. Int J Biomed Imaging. 2010;2010:896718.
12. Christin C, Hoefslooth HC, Smilde AK, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Mol Cell Proteomics. 2013;12(1):263–276.
13. Edwards N, X W, Tseng CW. An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. Clin Proteom. 2009;5:23.
14. Nagaraja K, Braga-Netou U. Bayesian classification of proteomics biomarkers from selected reaction monitoring data using an approximate bayesian computation-markov chain monte carlo approach. Cancer Inform. 2018;17. 1176935118786927.
15. Swaney DL, Mcalistergc GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat Methods. 2008;5(11):959–964.
16. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013;14(3):315–326.
17. Webb-Robertson BJM. Support Vector Machines for Improved Peptide Identification from Tandem Mass Spectrometry Database Search. In: Lipton MS, Paša-Tolic L, editors. Mass Spectrometry of Proteins and Peptides. Methods In Molecular Biology. vol. 492. Humana Press; 2009.
18. Raczynski L, Rubel T, Zaremba K. Neural Network-Based Method for Peptide Identification in Proteomics. In: Pięţka E, Kawa J, editors. Information Technologies in Biomedicine. Lecture Notes in Computer Science. vol. 7339. Berlin, Heidelberg: Springer; 2012.
19. Daniel B. Cross-Validation. In: Shoba R, G M, N K, S C, editors. Encyclopedia of Bioinformatics and Computational Biology. Academic Press; 2019. p. 542–545.
20. Fan Z, Kong F, Zhou Y, et al. Intelligence algorithms forprotein classification by mass spectrometry. Biomed Res Int. 2018;2018:2862458.
21. Manes NP, Nita-Lazar A. Application of targeted mass spectrometry in bottom-up proteomics for systems biology research. J Proteomics. 2018;189:75–90.
22. Wilson R. Sensitivity and specificity: twin goals of proteomics assays. Can they be combined? Expert Rev Proteomics. 2013;10(2):135–49.
23. Deutsch EW, Albar JP, Binz PA, et al. Development of data representation standards by the human proteome organization proteomics standards initiative. J Am Med Inform Assoc. 2015;22(3):495–506.
24. Martens L, Chambers M, Sturm M, et al. mzML - a community standard for mass spectrometry data. Mol Cell Proteomics. 2010;10(1). R110.000133.
25. Jones AR, Eisenacherm M, Mayer G, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics. 2012;11(7). M111.014381.
26. Walzer M, Qi D, Mayer G, et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics. 2013;12(8):2332–40.
27. Vizcaíno JA, Csordasa A, del Toro N, et al. 2016 update ofthe PRIDE database and its related tools. Nucleic Acids Res. 2016;44(D1):D–447–D–456.
28. Craig R, Cortens JP, Beavis RC. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. J Proteome Res. 2004;3(6):1234–1242.
29. Desiere F, Deutsch EW, King NL, et al. The PeptideAtlas Project. Nucleic Acids Res. 2006;34(Suppl 1):D–655–D–658.
30. Perez-Riverol Y, Alpie E, Wang R, Hermjakob H, Vizcaíno JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015;15(5-6):930–979.
31. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–81.
32. Karimpour-Fard A, Epperson LE, Hunter LE. A survey of computational tools for downstream analysis of proteomics and other omic datasets. Hum Genomics. 2015;9:28.
33. Chisanga D, Keerthikumar S, Mathivanan S, Chilamkurti N. Network Tools for the Analysis of Proteomics Data. In: Keerthikumar S, Mathivanan S, editors. Proteome Bioinformatics. Methods in Molecular Biology. vol. 1549. New York, NY: Humana Press; 2017.
34. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J. BioServices: a common Python packageto access biological Web Services programmatically. Bioinform. 2013;29(24):3241–2.
35. Prlić A, Yates A, Bliven SE, et al. BioJava: an open-source framework for bioinformatics in 2012. Bioinform. 2012;28(20):2693–5.
36. Hernández Y, Bernsteinr R, Pagan P, et al. BpWrapper: BioPerl-based sequence and tree utilities for rapid prototyping of bioinformatics pipelines. BMC Bioinform. 2018;19(1):76.
37. Chawade A, Sandin M, Teleman J, et al. Data processing has major impact on the outcome of quantitative label-free LC-MS analysis. J Proteome Res. 2015;14:676–87.
38. Navarro P, Kuharev J, Gillet LC, et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 2016;34:1130–6.
39. Runxuan Z, Barton A, Brittenden J, et al. Evaluation for computational platforms of LC-MS based label-free quantitative proteomics: a global view. J Proteomics Bioinform. 2010;3:260–5.
40. Smith R, Ventura D, Prince JT. LC-MS alignment in theory and practice: a comprehensive algorithmic review. Brief Bioinform. 2015;16:104–17.
41. Lange E, Tautenhahn R, Neumann S. Critical assessment of alignment proceduresfor LC-MS proteomics and metabolomics measurements. BMC Bioinform. 2008;9:375.
42. Zhang J, Gonzalez E, Hestilow T, et al. Review of peak detection algorithms inliquid-chromatography-mass spectrometry. Curr Genomics. 2009;10:388–401.
43. Tyanova S, Temu T, Carlson A, et al. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics. 2015;15:1453–1456.
44. Röst H, Sachsenberg T, Aiche S, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016;13:741–748.
45. Zhang J, Xin L, Shan B, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics. 2012;11(4). M111.010587.
46. Tibshirani R, Hastiet T, Narasimhan B, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics. 2004;20(17):3034–3044.
47. Ushijima M, Miyata S, Eguchi S, et al. Common peak approach using mass spectrometry data sets for predicting the effects of anticancer drugs on breast cancer. Cancer Inform. 2007;3:285–293.
48. Antoniadis A, Bigot J, Lambert-Lacroix S. Peaks detection and alignment for mass spectrometry data. J Société Française Stat. 2010;151(1):17–37.
49. Gibb S, Strimmer K. Differential protein expression and peak selection in mass spectrometry data bybinary discriminant analysis. Bioinformatics. 2015;31(19):3156–3162.
50. Brochu F, Plante P, Drouin A, et al. Mass spectra alignment using virtual lock-masses. Sci Rep. 2019;9:8469.
51. Ventoura S, Giannopoulou EG, Manolakos ES. ProtCV: A Tool for Extracting, Visualizing and Validating Protein Clusters Using Mass Spectra Peak-Lists. In: 21st IEEE International Symposium on Computer-Based Medical Systems, Jyvaskyla; 2008. p. 221–223.

52. Chen L, Liu Y, Liu C, et al. A two-phase clustering approach for peak alignment in mining mass spectrometry data. IEEE International Conference on Bioinformatics and Biomedicine Workshop. 2009;p. 226–230.
53. Liu YC, Chen LC, Liu CW, et al. Effective peak alignment for mass spectrometry data analysis using two-phase clustering approach. Int J Data Min Bioinform. 2014;9(1):52–66.
54. Kilgour D, Hughes S, Kilgour SL, et al. Autopiquer - a robust and reliable peak detection algorithm for mass spectrometry. J Am Soc Mass Spectrom. 2017;28:253–262.
55. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45(D1):D331–D338.
56. Kotlyar M, Jurisica I. Predicting protein-protein interactions by association mining. Inf Syst Front. 2006;8:37–47.
57. Kaake RM, Wang X, Huang L. Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry. Mol Cell Proteomics. 2010;9(8):1650–1665.
58. Miteva YV, Budayeva HG, Cristea IM. Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions. Anal Chem. 2013;85(2):749–768.
59. Schoenrock A, Samanfar B, Pitre S, et al. Efficient prediction of human protein-protein interactions at a global scale. BMC Bioinformatics. 2014;15(1):383.
60. Yugandhar K, Gupta S, Yu H. Inferring protein-protein interaction networks from mass spectrometry-based proteomic approaches: a mini-review. Comput Struct Biotechnol J. 2019;17:805–811.
61. Saeys Y, Abeel T, Peer Y. Robust feature selection using ensemble feature selection techniques. ProcEur Conf Machine Learning Knowledge Discovery Databases. 2008;Part II:313–325.
62. Veenstra TD. Globaland targeted quantitative proteomics for biomarker discovery. J Chromatog B. 2007;847:3–11.
63. Ralhan R, Desouza LV, Matta A, et al. Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. Mol Cell Proteomics. 2008;7:1162–1173.
64. Fan Z. A neuralnetwork approach to multi-biomarker panel development based on LC/MS/MSproteomics profiles: A case study in breast cancer. In: Jake YC, editor. 22nd IEEE International Symposium on Computer-Based Medical Systems; 2009.
65. Ryberg H, J A, Darko S, et al. Discovery and verificatio nof amyotrophic lateral sclerosis biomarkers by proteomics. Muscle Nerve. 2010;42:104–111.
66. Bloemen K, Denheuvel RV, Govarts E, et al. A new approach to study exhaled proteins as potential biomarkers for asthma. Clin Exp Allergy. 2011;41:346–356.
67. Le L, K C, Tyldesley S, et al. Identification of serum amyloid A as a biomarker to distinguish prostate cancer patients with bone lesions. Clin Chem. 2005;51:695–707.
68. Willingale R, Jones DJL, Lamb JH, et al. Searching for biomarkers of heart failure in the mass spectra of blood plasma. Proteomics. 2006;6:5903–5914.
69. Guan W, Zhou M, Hampton C, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformat. 2009;10:259.
70. Oh JH, Lotan Y, Gurnani P, et al. Prostate cancer biomarker discovery using high performance mass spectral serum profiling. Computer Methods Programs Biomed. 2009;96:33–41.
71. Sigdel T, Salomonis N, Nicora C, et al. Potential urine protein biomarkers for kidney transplantation dysfunction through quantitative proteomics. Am J Transplant. 2015;15(suppl 3).
72. Tsai TH, Song E, Zhu R, et al. LC-MS/MS-based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma. Proteomics. 2015;15:2369–2381.
73. Niu R, Liu Y, Zhang Y, et al. iTRAQ-based proteomics reveals novel biomarkers for idiopathic pulmonary fibrosis. PLoS ONE. 2017;12(1):e0170741.
74. Bringans SD, Ito J, Stoll T, et al. Comprehensive mass spectrometry based biomarker discovery and validation platform as applied to diabetic kidney disease. EuPA Open Proteom. 2017;14:1–10.
75. Hirao Y, Saito S, Fujinaka H, et al. Proteome profiling of diabetic mellitus patient urine for discovery of biomarkers by comprehensive MS-based proteomics. Proteomes. 2018;6(1):9.
76. Chae Y. Mass spectrometry-based serum proteomic signature as a potential biomarker for survival in NSCLC patients with immunotherapy. J Thorac Oncology. 2018;13(10):S529–S530.
77. Tsuruoka Y, Miwa M, Hamamoto K, et al. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011;27(13):i111–i119.
78. Lawlor K, Nazarian A, Lacomis L, et al. Pathway-based biomarker search by high-throughput proteomics profiling of secretomes. J Proteome Res. 2009;8:1489–1503.
79. Huang DW, Lempicki RA, Sherman BT. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
80. Matthiesen R, Bunkenborg J. Introduction to Mass Spectrometry-Based Proteomics. In: R M, editor. Mass Spectrometry Data Analysis in Proteomics. Methods in Molecular Biology (Methods and Protocols). vol. 1007. Totowa, NJ: Humana Press; 2013.
81. Ray S, Reddy PJ, Jain R, et al. Proteomics technologies forthe identification of disease biomarkers in serum: Advances and challenges ahead. Proteomics. 2011;11:2139–2161.
82. Vaudel M, Verheggen K, Csordas A, et al. Exploring the potential of public proteomics data. Proteomics. 2016;16:214–225.
83. Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic sexperiments. J Proteome Res. 2009;8:2144–2156.
84. Tabb DL, Vega-Montoto L, Rudnick PA, et al. Repeatability and reproducibility in proteomics identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res. 2010;9:761–776.
85. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PLoS ONE. 2012;7:e39932.
86. Luo W, Friedman M, Shedden K, Hankenson K, Woolf P. GAGE: Generally applicable gene set enrichment for pathway analysis. BMC Bioinformat. 2009;10:161.
87. Koo CL, Liew MJ, Mohamad MS, Salleh AH. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. Biomed Res Int. 2013;2013:432375.
88. Wang S, Fu L, Yao J, et al. The Application of Deep Learningin Biomedical Informatics. In: International Conference on Robots & Intelligent System (ICRIS), Changsha; 2018. p. 391–394.
89. Quan L, Liu M. CID, ETD and HCD fragmentation to study protein post-translational modifications. Mod Chem. 2013;appl 1:e102.
90. Arnold RJ, Jayasankarn N, Aggarwal D, et al. A machine learning approach to predicting peptide fragmentation spectra. Pac Symp Biocomput. 2006;p. 219–230.
91. Dong NP, Liang YZ, Xu QS, et al. Prediction of peptide fragment ion mass spectra by data mining techniques. Anal Chem. 2014;86:7446–7454.
92. Tiwary S, Levy R, Gutenbrunner P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods. 2019;16(6):519–525.

Tạp chí Phát triển Khoa học và Công nghệ – Kĩ thuật và Công nghệ, 2(4):258-276

Open Access Full Text Article

Tổng quan

Khai thác dữ liệu trong nghiên cứu proteomic dựa trên kỹ thuật khối phổ

Lê Anh Vũ[1,*], Phan Thị Cẩm Quyên[2]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Các nghiên cứu thời kỳ hậu genomic bao gồm nhiều thiết kế thử nghiệm và tính toán để làm rõ và hiểu chức năng của gen cũng như các sản phẩm của chúng. Các nghiên cứu về proteomic đóng một vai trò quan trọng trong xu hướng này bằng cách bổ sung cho các phương pháp tiếp cận chức năng bộ gen khác, bao gồm phân tích quy mô lớn các hỗn hợp phức tạp, xác định và định lượng protein được biểu hiện trong các điều kiện khác nhau, xác định tính chất, biến đổi và chức năng tương ứng của chúng. Việc hiểu làm thế nào các quá trình sinh học được điều chỉnh ở mức độ protein là rất quan trọng để hiểu cơ sở phân tử của bệnh và thường có nhiều tiềm năng ứng dụng việc phòng ngừa, chẩn đoán và điều trị bệnh. Các kỹ thuật thông lượng cao được sử dụng rộng rãi trong nghiên cứu proteomic để thực hiện việc phân tích hàng ngàn protein cùng lúc. Cụ thể, khối phổ (mass spectrometry - MS) là một kỹ thuật thường dùng trong phân tích đặc điểm các mẫu sinh học và ngày càng được sử dụng nhiều trong các nghiên cứu về protein vì khả năng nhắm mục tiêu, không nhắm mục tiêu và hiệu suất cao của nó. Tuy nhiên, khi các tập dữ liệu lớn được tạo, các phương pháp tính toán như kỹ thuật khai thác dữ liệu là cần thiết để phân tích và giải thích dữ liệu liên quan. Cụ thể hơn, việc áp dụng các kỹ thuật khai thác dữ liệu trong các bộ dữ liệu proteomic lớn có thể hỗ trợ nhiều cách hiểu về dữ liệu; nó có thể làm sáng tỏ các tương tác protein-protein, cải thiện nhận dạng protein, đánh giá các phương pháp thí nghiệm được sử dụng và tạo điều kiện thuận lợi cho chẩn đoán và phát hiện dấu ấn sinh học. Với những tiến bộ nhanh chóng trong thiết kế các thiết bị quang phổ khối và phương pháp thí nghiệm phù hợp, proteomic dựa trên MS đã trở thành một công cụ đáng tin cậy và cần thiết để làm sáng tỏ các quá trình sinh học ở cấp độ protein. Trong những thập kỷ qua, chúng ta đã chứng kiến sự mở rộng kiến thức về các bệnh của con người với việc áp dụng các công nghệ proteomic dựa trên MS, dẫn đến nhiều khám phá quan trọng. Trong tổng quan này, chúng tôi trình bày những tiến bộ gần đây của lĩnh vực khai thác dữ liệu proteomic dựa trên MS trong nghiên cứu y sinh. Những nghiên cứu gần đây trong nhiều lĩnh vực cho thấy rằng proteomic đã vượt ra ngoài việc phân loại protein đơn giản trong các hệ thống sinh học và cuối cùng đạt được tiềm năng như một công cụ thiết yếu để hỗ trợ các ngành liên quan, đặc biệt là nghiên cứu về sức khỏe. Do đó, khai thác dữ liệu proteomic dựa trên MS sẽ có tiềm năng lớn để vượt ra ngoài những nghiên cứu cơ bản, ứng dụng vào nghiên cứu lâm sàng và chẩn đoán y sinh.

Từ khoá: khai thác dữ liệu, khối phổ, nghiên cứu y sinh, proteomic, tin sinh học

[1]*Khoa Kỹ thuật Hóa học, Trường Đại học Bách Khoa, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam*

[2]*Phòng Công nghệ Sinh học, Trung tâm Giống Kiên Giang, Việt Nam*

Liên hệ

Lê Anh Vũ, Khoa Kỹ thuật Hóa học, Trường Đại học Bách Khoa, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Email: lavu68@gmail.com

Lịch sử

- Ngày nhận: 18-3-2019
- Ngày chấp nhận: 18-12-2019
- Ngày đăng: 31-12-2019

DOI : 10.32508/stdjet.v2i4.483

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Anh Vũ L, Thị Cẩm Quyên P. **Khai thác dữ liệu trong nghiên cứu proteomic dựa trên kỹ thuật khối phổ**. *Sci. Tech. Dev. J. - Eng. Tech.;* 2(4):258-276.