OA Bioinformatics
Open Access

Computational Biology

Review

Stratification of biological samples based on proteomics data

D Di Silvestre[1], I Zoppis[2], G Mauri[2], PL Mauri[1]*

Abstract

Introduction

Stratification of biological samples by using high-dimensional data, such as those derived from mass spectrometry-based proteomics approaches, has become a promising strategy to solve biological questions, as well as to classify samples in relation to different phenotypes. In this regard, we have discussed some computational aspects related to the processing of Multidimensional Protein Identification Technology data through a class of algorithms widely used in machine learning community, such as support vector machines. Specifically, after a short presentation of the input data structure, we focused on properties and abilities of feature selection and classification models, indicating useful tools for assisting scientists in these computations. Finally, we concluded this review hinting at new strategies of inference which coupled to mass spectrometry improvement, in instruments and methods, may represent the perspectives of this field.

Conclusion

In this review we have made a well-defined overview of a method that, by combining high-throughput proteomic data and machine learning algorithms, allows the stratification of biological samples. Besides the importance that these procedures can play for diagnostic or prognostic purposes, they are useful also for identifying meaningful expression patterns. Therefore, it represents a valid tool for investigating both clinical and biological aspects.

Introduction

Recent developments in analytical techniques such as mass spectrometry (MS) have created the opportunity to measure proteomes at large-scale, providing a representative snapshot of cells and/or tissues associated with different phenotypes. In this context, new MS instruments are able to reach the limit of detection up to attomole and a dynamic range of 1 × 10[6][1]. As a consequence, MS has become essential for proteomic research, and owing to its powerful activity of discovering it has already been introduced as a tool for clinical applications. In fact, one of the main aims in this field is to use relevant biomarkers for improving current methods of diagnosis (e.g. healthy–diseased), for selecting appropriate therapeutic approaches and for monitoring their effectiveness[2].

The construction of an inference model able to discriminate biological samples (sharing some characteristics, such as m/z ions, peptides or proteins) is a common issue in many areas of life sciences including proteomics. In the last few years, a variety of algorithms have been designed for this purpose. In many of these studies, different authors applied support vector machines (SVMs)[3] to experimental data mainly generated by analysing body fluids through MALDI (matrix-assisted laser desorption/ionisation) and SELDI (surface-enhanced laser desorption/ionisation) technologies, while very few cases investigated the data obtained by liquid chromatography coupled to MS[4]. In a number of publications, discovery of biomarker patterns has been reported with diagnostic sensitivities and specificities approaching 100%. Although these results prefigure a prominent position for diseases diagnosis, to realise the potential of MS-based proteomics in the area of clinical utility, additional requirements, such as reproducibility and standardisation of methods, need to be addressed[5].

Regardless of the analytical methodology used to generate proteomic data, two main interests address the inference on the biological sample discrimination: *the feature selection* and *the classification* problems (Figure 1). For each of them, scientists can apply a wide range of algorithms, hence there is no unique way leading to an adequate inference model. As a consequence, which strategy works best is yet an open issue. To answer this question, some investigators have begun to perform studies for assessing which procedure

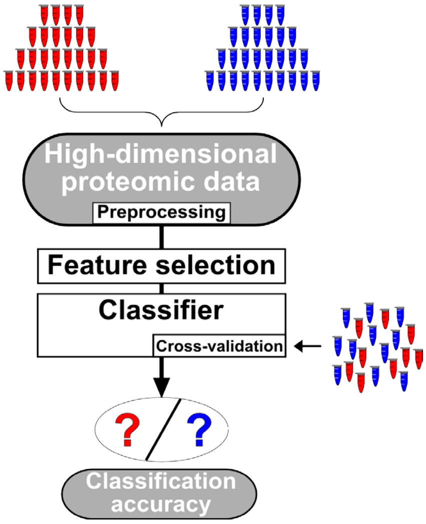


Figure 1: General workflow for sample classification by using high-dimensional proteomic data.

* Corresponding author
Email: pierluigi.mauri@itb.cnr.it

[1] Institute for Biomedical Technologies, Council National Research (ITB-CNR), 20090 Segrate, Milan, Italy

[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20132 Milan, Italy

FOR CITATION PURPOSES: *Di Silvestre D, Zoppis I, Mauri G, Mauri PL.* Stratification of biological samples based on proteomics data. OA Bioinformatics 2013 Aug 01;1(1):1.

*Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

allows the best performances; some of them combined feature selection techniques with statistical and machine learning[6], while others evaluated biomarker discovery of different feature selection methods[7] or the classifier capabilities associated with different data types[8]. Although these comparative studies are of great importance, their comparison is very difficult because they vary in several conditions, ranging from analytical parameters to algorithms used for statistical data analysis. For this reason, in this review, we have focused mainly on the computational aspects to discriminate phenotypes, by a class of algorithms widely used in the machine learning community, such as SVMs, and proteomic data obtained by Multidimensional Protein Identification Technology (MudPIT)[9]. In particular, we first shortly describe the MudPIT approach and the structure of its data, and then we discuss the computational aspects of the two main interests mentioned earlier, that is *feature selection* and *SVM classification*.

Proteomic data

MudPIT approach is a powerful analytical method based on two-dimensional liquid chromatography separation coupled to tandem MS (Figure 2). Except sample preparation, the whole experiment is fully automated and the process, repeated several times, produces thousands of MS and MS/MS spectra. MS precursor ion intensities can be used for peptide quantification, while MS/MS spectra contain sequence information and are processed for obtaining peptide and protein lists. Specifically, MS/MS interpretation is based on the comparison of experimental spectra versus theoretical peptide fragments calculated from a reference database. This job is performed by specific database search engines, commercial or available for free. One of oldest and best recognised algorithms for this purpose is the SEQUEST[10], which is inserted in Bioworks and Discovery software (ThermoFisher Scientific). The latter also provides spectra interpretation by means of Mascot[11], which is probably the most widely used tool for mass spectra interpretation. Typically, a single MudPIT run allows the identification up to a thousand proteins, peptides, and spectra per sample[8]. These data are associated with quantitative sampling parameters, such as peak area intensity or spectral count (SpC), which in the last decade have permitted the development of tools and procedures for characterising biomarkers, by means of both label- and label-free approaches[4].

Both for biomarker discovery and classification inferences, results from multiple MudPIT experiments are conveniently represented in a table, as shown in Figure 3, where the number of variables are usually bigger than the number of analysed samples ($v >> s$). Columns are indexed through some characteristics describing the samples. As we shall see, these are

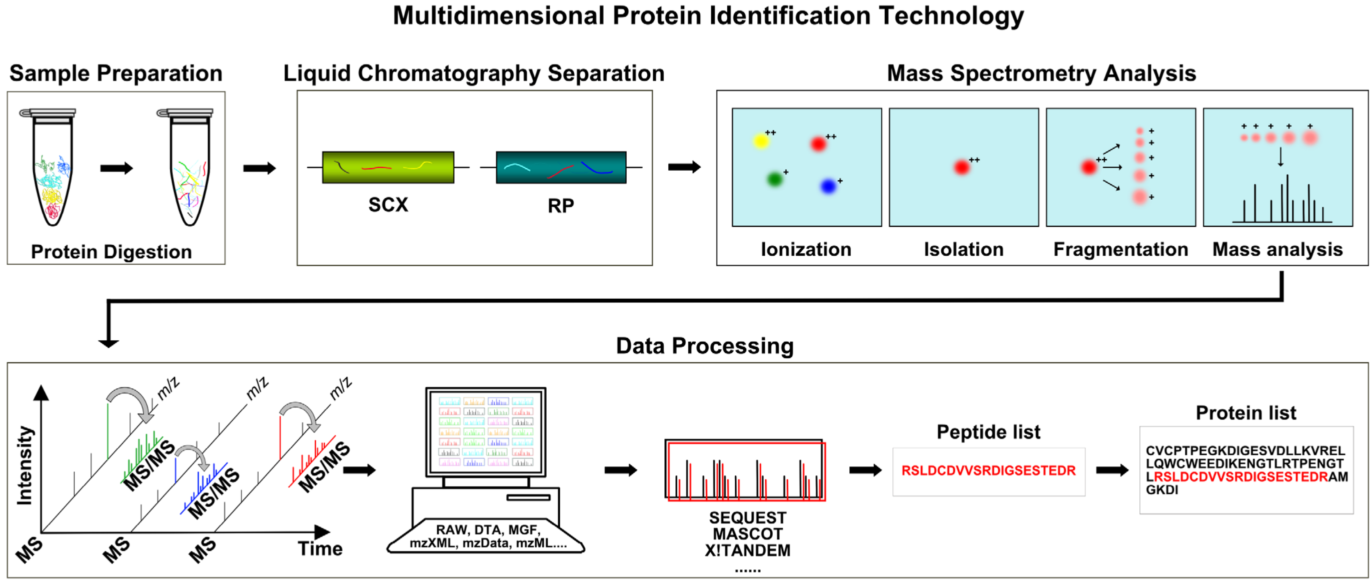


Figure 2: Multidimensional Protein Identification Technology workflow. MudPIT approach is made of four distinct phases consisting of sample preparation, liquid chromatography separation, mass spectrometry analysis and data-processing. Sample preparation usually involves protein extraction and trypsin digestion. Generated peptides are separated by means of strong cation exchange (SCX) followed by C18 reverse phase (RP) chromatography, and directly analysed by mass spectrometer. It isolates ions (MS) of a particular peptide, subjects them to fragmentation and records the produced fragments in a tandem mass spectrum (MS/MS). Finally, MS/MS spectra are processed by means of database search engines for obtaining peptide and protein lists.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

the *features* considered for the task at hand. Rows are the analysed samples (e.g. cell lines, tissues and body fluids). Each cell represents the value (i.e. expression level) assumed by the feature *j* when describing the sample *i*. A specific feature identifying the group membership of each observation is generally given in the case of a classification problem. In figure 3, a vector of target variables *yi*{−1, +1} is coded to identify the group membership (i.e. case/control) of each observation.

Before applying classical label-free quantification approaches, as well as future selection and classification algorithms, standard procedures involve data pre-processing to remove instrumental noise and to make measurements comparable. For instance, mass spectral profiles may be affected by baseline effects, shifts in mass-to-charge ratio, alignment problem or differences in signal intensities, which may be corrected by software, such as MZmine[12]. In the same way, variation of sampling parameters associated with identified proteins, such as SpC, could be due to different amounts of analysed sample. These differences are adjusted by using strategies of data normalisation[13]. For example, we report one of the simplest procedures, called Total Signal method: given two or more samples, $S1, S2 \ldots S_i$, and with $T_1, T_2 \ldots T_i$ the respective sum of SpC for all the proteins, this method aims to obtain $T_1 = T_2 \ldots = T_i$. Therefore, normalisation of SpC_{ij} is obtained by dividing SpC_{ij} for the value obtained by summing SpC values of all proteins belonging to the same list (T_i).

$$\mathrm{SpC}_{ij} = \mathrm{SpC}_{ij}/\mathrm{T}_i$$

Feature selection

Recent proteomic technologies provide great amount of data, which need to be processed through sophisticated tools. As mentioned previously, when handling such data there are two general interests.

1. the construction of a model which is able to discriminate between case and control samples (i.e. *classification problem*).
2. the need to perceive which proteins (peptide, signals etc.) are associated with specific factors of interest (e.g. differentially expressed proteins between case and control groups), thus suggesting potential biomarkers for future investigation (i.e. *feature selection problem*).

In general, the first interest does not imply the second; however, in some cases they cannot be considered as independent issues[14]. For instance, proteomic profiles consist of a wide range of measurements, such as peak area intensity or SpC, evaluated for both biomarker discovery and classification inference. Hence, to assure good inference accuracy one has to search for a robust combination of feature selection methods and classification models.

Features (or attributes) are characteristics describing the samples. In many applications, identifying the most characteristic features is critical, for example when one tries to minimise the inference accuracy obtainable in the subsequent classification problem. The feature selection[15–19] refers to the task of identifying the useful subset of attributes to be used for representing patterns of a larger set of often mutually redundant or irrelevant attributes. This process is fundamental for proteomic data sets due to the abundance of noisy (e.g. chemical) or misleading features. Depending on the characteristics of the classification model, irrelevant and redundant features could worsen the prediction rate for the classification problem. Moreover, reducing the number of features gives less computationally intensive models. To provide the most accurate subset of features, we would ideally have to test all the subsets of the original n features. This exhaustive enumeration is infeasible in most cases as it results in 2^n subsets to be tested. Finally, in many applications (e.g. diagnosis), it is clear that a reduced number of features avoid risks (e.g. invasive exploratory surgery) and save costs during their utilisation.

Generally, feature selection algorithms can be classified into three main categories based on whether or not feature selection is done independent of the learning algorithm used, for example to construct the classifier. These categories include filter, wrapper and embedded methods.

Filter methods

The fastest way for feature selection is probably ranking the features. In this case, *filters* do not take into account feature interaction, but they assess the relevance of a feature by looking only at the intrinsic characteristics of the data. In most cases, a feature relevance score is calculated, and low-scoring features are removed. The features can be ranked with some statistical test. Several methods use simple correlation coefficients similar to Fisher's discriminant criterion; others adopt mutual information[19] or statistical *t*-test, or *F*-test. Advantages of filter techniques are that they easily scale to very high-dimensional data sets, and they are computationally simple and fast. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated. As mentioned earlier, a limitation of filter methods is that they ignore both interaction with the classifier and the feature dependencies. To overcome this last problem, a number of *multivariate* filter techniques were introduced[18].

Wrapper methods

Wrapper techniques perform a search in the space of feature subsets by incorporating the classification algorithm within the process. In other

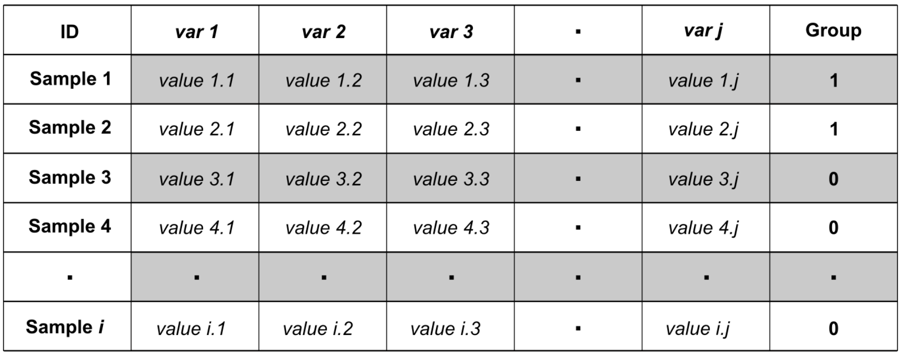
Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

words, *wrappers* utilise the classifier as a 'black box' to score the subsets of features based on their predictive power. This way, the evaluation of a specific subset is obtained by *training* and *testing* the classification model, rendering this approach tailored to the specific classification algorithm. As the number of all feature combinations is exponential in the number of the considered features, the search for the subset which provides the most accurate classification accuracy is often critical for its practical acceptance. To overcome this problem, many heuristic methods, deterministic and randomised, are used to guide the search of 'sub-optimal' subsets[15,16].

Embedded methods
Embedded methods search among different feature subsets, however, unlike wrappers, the process is tied closely to a certain classification model and takes advantage of its characteristics and structure. In other words, the learning part and the feature selection problem cannot be separated, for example Weston et al.[20] measured the importance of a feature using a bound that is valid only for SVMs.

Classification with SVM

SVMs have become a state-of-the-art technique in solving classification and regression problems[3,21,22]. The reason for this success is not only because of their sound theoretical foundation but due to their good generalisation performance in many real applications, also. Here, we focus on SVMs for two-class classification problems (e.g. cases/control). In this case, samples can be suitably represented geometrically through their feature values (rows of the table in Figure 3) as set of points in R^n. Figure 4a gives a simple example of representation for the Euclidean plane. By associating with the points of their group membership variable, we obtain the representation in Figure 4b.



ID	*var 1*	*var 2*	*var 3*	.	*var j*	Group
Sample 1	*value 1.1*	*value 1.2*	*value 1.3*	.	*value 1.j*	**1**
Sample 2	*value 2.1*	*value 2.2*	*value 2.3*	.	*value 2.j*	**1**
Sample 3	*value 3.1*	*value 3.2*	*value 3.3*	.	*value 3.j*	**0**
Sample 4	*value 4.1*	*value 4.2*	*value 4.3*	.	*value 4.j*	**0**
.	**.**	**.**	**.**	**.**	**.**	**.**
Sample *i*	*value i.1*	*value i.2*	*value i.3*	.	*value i.j*	**0**

Figure 3: Example of profile data from case and control groups. Rows represent samples, while columns indicate their features (e.g. *m/z*, peptides or proteins). In each cell, a value corresponding to the parameter associated with feature is reported. For instance, peak area intensity (AUC) may be used for *m/z* mass points or peptides, while SEQUEST score or spectral count (SpC) for proteins.

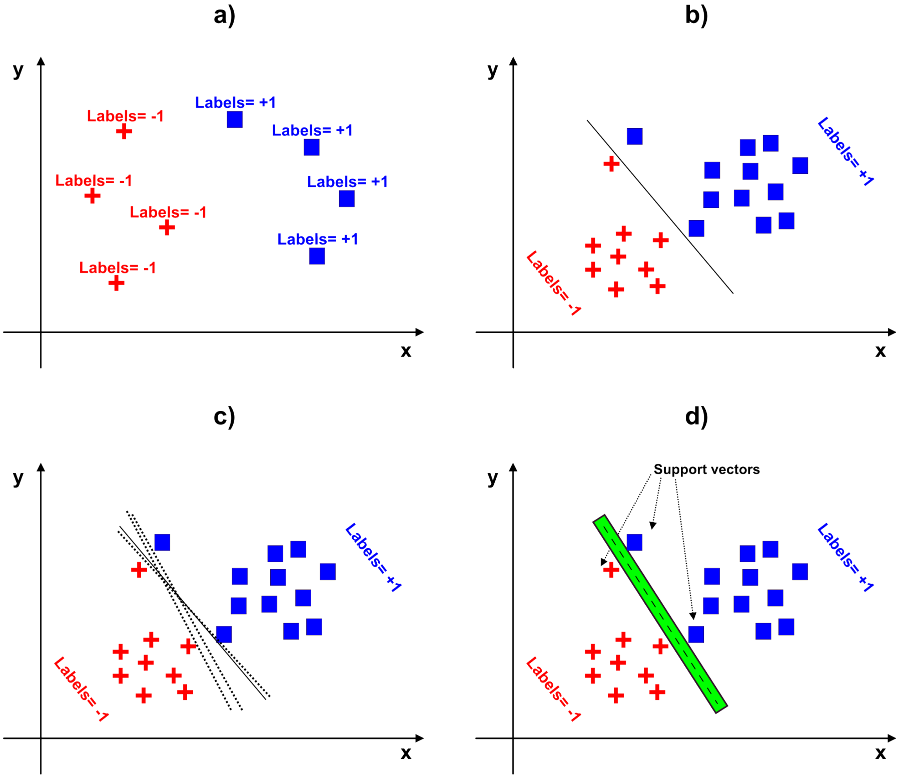


Figure 4: Linear SVM classification. (a) Labelled points are represented in R^2. (b) A decision surface (line) separates the two classes of points. (c) There are many possible hyperplanes separating the two classes of points. (d) The optimal hyperplane separates positive and negative examples with the maximal margin. The position of the optimal hyperplane is determined by the examples that are closest to the hyperplane (support vectors.).

In the figure, we illustrate a set of *linearly separable* points labelled by −1 for disease patients and +1 for control subjects. We point out that a set of points are linearly separable when they can be completely separated by a line (in R^2) or, for much higher dimensions, by a hyperplane. All

FOR CITATION PURPOSES: *Di Silvestre D, Zoppis I, Mauri G, Mauri PL.* Stratification of biological samples based on proteomics data. OA Bioinformatics 2013 Aug 01;1(1):1.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

samples of one class lie on one side of the line and all samples of the other class lie on the other side. In this situation, the class of *linear* SVMs can be easily applied, for example for the classification problem.

Linear SVMs search for the optimal hyperplane that is equidistant from the two considered classes of samples. Generally, there are most likely many possible hyperplanes that separate the classes (Figure 4c). For this reason, the main issue of the SVMs is to find the separating hyperplane with the largest distance (i.e. *margin*) between border-line samples (i.e. *support vectors*) from the two classes (Figure 4d). In other words, the hyperplane is a decision boundary between the two classes of points. Once this hyperplane has been obtained, the class label (i.e. membership group) of new samples can be predicted by testing which side of the hyperplane they appear.

Now, we give a brief mathematical summary of this approach. Let K be the collection of points $\{(x_i, y_i), x_i \in \mathrm{R}^n, y_i \in \{-1, 1\}\}$, where y_i indicates the class label for $\mathbf{x}_i$, $1 \le i \le N$. The goal is to find the *maximum-margin hyperplane* dividing the points having $y_i = 1$ from those having $y_i = -1$. We can express any hyperplane as the set of points $\mathbf{x}_i$ satisfying, $\mathbf{w}\cdot\mathbf{x}_i - b = 0$, where · denotes the dot product and $\mathbf{w}$ is a normal vector perpendicular to the hyperplane, the value $b/||\mathbf{w}||$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. Therefore, we can choose $\mathbf{w}$ and b to maximise the distance (margin) between the parallel hyperplanes in such a way that the data are separated. The two equations describing these hyperplanes are $\mathbf{w}\cdot\mathbf{x}_i - b = 1$ and $\mathbf{w}\cdot\mathbf{x}_i - b = -1$. For linearly separable points, the distance between these two hyperplanes is $2/||\mathbf{w}||$; hence, the goal is to minimise $||\mathbf{w}||$. Furthermore, to prevent the data points from falling into the margin, the following constraints are also needed: for each $\mathbf{x}_i$ either $(\mathbf{w}\cdot\mathbf{x}_i - b \ge 1)$ for the first class or $(\mathbf{w}\cdot\mathbf{x}_i - b \le -1)$ for the second, and more compactly y_i $(\mathbf{w}\cdot\mathbf{x}_i$ - b ≥ 1) for all $1 \le i \le n$. Putting all of this together, we obtain the following optimisation problem:

$$\min_{w,b} \|w\|$$
$$s.t. y_i\left(w x_i - b \ge 1\right), 1 \le i \le n.$$

The optimisation problem in Equation (1) is difficult to solve as it depends on $||\mathbf{w}||$ (the norm of $\mathbf{w}$) that involves a square root. Note that by substituting $||\mathbf{w}||$ in the objective with $1/2||\mathbf{w}||^2$ the solution remains unchanged as the minimum of the original and the modified equation both have the same $\mathbf{w}$ and b (the factor of 1/2 being used for mathematical convenience). Hence, we obtain the equivalent quadratic programming (QP) optimisation problem as follows:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2}$$
$$s.t. y_i\left(w x_i - b \ge 1\right), 1 \le i \le n$$

This problem can be solved using standard QP optimisation techniques[3]. The optimal solution $(\tilde{w}, \tilde{b})$ for the problem in Equation (2) enables the classification of a new point $\mathbf{z}$ according to the following expression: class($\mathbf{z}$) = sgn($\tilde{w} \cdot \mathbf{z} + \tilde{b}$), where sgn is the signum function. This way, the label of $\mathbf{z}$ is +1 if the vector $\mathbf{z}$ is greater than or equal to zero and −1 if it is less than zero. When the linear decision surface does not exist (i.e. points are not linearly separable) the data can be mapped into a much higher-dimensional space where the separating hyperplane can be found[23].

Data mining workflows

Many *open source* systems support users with graphical interface for rapid prototyping of machine learning and knowledge discovery (KD) processes, for example Weka[24], Taverna[25], KNIME[26] and RapidMiner[27]. The KD process is modelled by a complex nested chain of objects, called operators, which can be dropped as nodes onto a *working pane*. In this way, data-flows are specified by connecting the operator nodes in such a way that one is able to represent the conceptual sequence of operational steps (i.e. workflows) applied for different data mining experiments. The workflows are essentially executable visual representations of complex procedures, and can realise easily classification, clustering, feature selection and even data integration tasks[28]. For example, Figure 5 shows the RapidMiner workflow designed for a feature selection process. Basically, it implements standard SVM algorithms to forecast the patient membership group.

As mentioned above, SVMs are used as 'black box' inference processes to score each set of features according to the inference performance of the algorithm. After learning is completed, adequacy of features or a classifier is evaluated to insure that it provides a universal model able to generalise to new data the relationships learned on the training set. This capability may be tested through different methods (e.g. *k*-fold cross-validation, bootstrapping and hold-out)[3] and by using validation sets, previously unseen. One of the most common procedures to measure the classifier performance is based on the confusion matrix. It assesses phenotype prediction through standard indices (Table 1). However, other methods, such as receiver-operating characteristic curve, are available[29].

Discussion

Traditional inference tasks, such as classification, feature selection or clustering, attempt to find patterns in a data set characterised by a collection of independent instances of a single 'table'. Numerous algorithms have been designed to work on such a standard approach, where instances can be easily represented

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

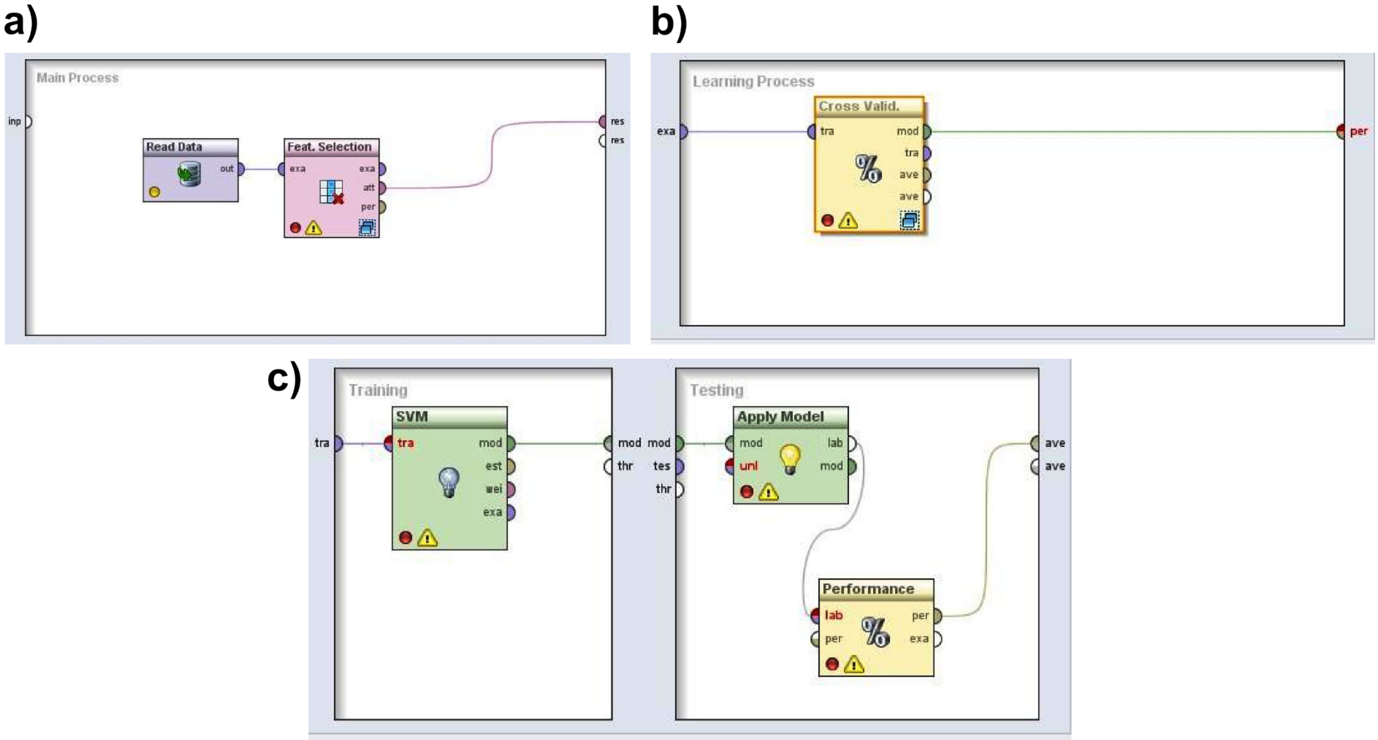


Figure 5: RapidMiner workflow. (a) Data are retrieved by the '*Read data*' operator and the feature selection is performed ('*Feature Selection*' operator). (b) Feature selection encapsulates a cross-validation process ('*Cross-validation*' operator) to select the most performing set of features. (c) *Cross-validation* operator encapsulates a *k-fold cross validation process*. Cross-validation is a two-step process: in the first step a classifier is built describing a predetermined set of data classes. In the second step, the model (a trained SVM) is used for testing new classification examples. The first inner operator ('*SVM*') realises the first step described earlier (*Training*). The second inner operator ('*Apply Model*') realises the second step. Finally, the predictive accuracy of the classifier is estimated by the '*Performance*' operator (*Testing*).

as fixed-length vectors of attribute values. Unfortunately, many studies still do not consider that many real problems are best described by structured data where instances of multiple types are related to each other in complex ways. For this reason, data sets to be analysed may be described by a relational database or semi-structured representations, such as XML. In this case, features of one entity are often correlated with the features of related entities. It may happen that, just as some features are not helpful for mining data sets, some relations might provide information for clustering or classification algorithms. For instance, when it comes to analyse differentially expressed MS peaks (or proteins) in a case/control classification problem, comparisons are generally performed between profiles of different groups or between statistics summarising the peaks' property of a group[30]. Actually, different neighbourhoods in the *m/z* spectra can be (anti)correlated to each other, and this property, in turn, may change from group to group. In such a situation, the incorporation of relational information can give powerful discrimination ability. This has been proved useful in many fields[31–33], and represents a promising approach also in relation to both MudPIT data structure and MS improvement, in instruments and methods, such as targeted proteomics or data-independent analysis[34,35]. In fact, the improved quality of data has the potential to optimise classifier capability and address the increasing demand of systems biology studies for correlating molecular expression to biological processes.

Conclusion

Improvement in mass spectrometry, coupled to advanced statistical analysis, represents a good starting point for developing procedures of investigation more and more accurately and precise. It may have important effects on understanding biological questions, and represent a crucial aspect for developing efficient methods of diagnosis, prognosis and therapeutic follow-up of human diseases. However, both for analytical and statistical parts, some questions like

FOR CITATION PURPOSES: *Di Silvestre D, Zoppis I, Mauri G, Mauri PL.* Stratification of biological samples based on proteomics data. OA Bioinformatics 2013 Aug 01;1(1):1.

Table 1: Indices for evaluating classifier performance.
TP and TN stand for true positives and true negatives and represent correct classifications. FP and FN indicate false positive and false negative, respectively. FP is when the outcome is incorrectly classified as positive, while FN is when the outcome is incorrectly classified as negative.

Index	Formula
Sensitiviy	$\frac{TP}{TP+FN}$
Specificity	$\frac{TP}{TN+FP}$
Positive predicted value (PPV)	$\frac{TP}{TP+FP}$
Negative predicted value (NPV)	$\frac{TN}{TN+FN}$
Overall classification accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Balanced accuracy	$\frac{Sensitivity+Specificity}{2}$
F-score	$2*\frac{PPV^{*}Sensitivity}{PPV+Sensitivity}$
Informedness	*Sensitivity* + *Specificity* −1
Matthews correlation coefficient	$\frac{(TP^{*}TN)-(FP^{*}FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

measurement reproducibility and lack of standardisation, remain open and further studies are obviously needed. In fact, although MS-based proteomics have become an important field for clinical applications, its potential have not yet been extensively developed and applied.

Abbreviations list

KD, knowledge discovery; MudPIT, Multidimensional Protein Identification Technology; MS, mass spectrometry; QP, quadratic programming; SpC, spectral count; SVM, support vector machine.

Acknowledgment

This study was supported by the Italian Ministry of Economy and Finance to the CNR for the Project 'FaReBio di Qualita', by PON AAA (01_02388) and by Cariplo Foundation (2009-2532).

References

1. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11: 49–79.

2. Baker ES, Liu T, Petyuk VA, Burnum-Johnson KE, Ibrahim YM, Anderson GA, et al. Mass spectrometry for translational proteomics: progress and clinical implications. Genome Med. 2012 Aug;4(8):63.

3. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.

4. Di Silvestre D, Brunetti P, Mauri PL. Processing of mass spectrometry data in clinical applications. In: Wang X, editor. Translational bioinformatics, vol. 3. Bioinformatics of human proteomics; 2013. p207–33.

5. Palmblad M, Tiss A, Cramer R. Mass spectrometry in clinical proteomics – from the present to the future. Proteomics Clin Appl. 2009 Jan;3(1):6–17.

6. Sampson DL, Parker TJ, Upton Z, Hurst CP. A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. PLoS One. 2011;6(9):e24973.

7. Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Mol Cell Proteomics. 2013 Jan;12(1):263–76.

8. Di Silvestre D, Zoppis I, Brambilla F, Bellettato V, Mauri G, Mauri P. Availability of MudPIT data for classification of biological samples. J Clin Bioinforma. 2013 Jan;3(1):1.

9. Mauri P, Scigelova M. Multidimensional protein identification technology for clinical proteomic analysis. Clin Chem Lab Med. 2009;47(6):636–46.

10. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, et al. Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol. 1999 Jul;17(7):676–82.

11. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 'Probability-based protein identification by searching sequence databases using mass spectrometry data'. Electrophoresis. 1999;20(18): 3551–67.

12. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395.

13. Carvalho PC, Fischer JSG, Chen EI, Yates 3rd JR, Barbosa VC. PatternLab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics. 2008;9:316.

14. Kohavi, R. and John, G. The wrapper approach. In: Liu H, Motoda, H., editors, Feature selection for knowledge discovery and data mining. Norwell, MA, USA: Kluwer Academic Publishers; 1998. p.33–50.

For citation purposes: *Di Silvestre D, Zoppis I, Mauri G, Mauri PL.* Stratification of biological samples based on proteomics data. OA Bioinformatics 2013 Aug 01;1(1):1.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

15. Blum A, Langley P. Selection of relevant features and examples in machine learning. Art Intell. 1997;(1–2):245–71.

16. Guyon I, Elissee A. An introduction to variable and feature selection. J Machine Learning Res. 2003;3:1157–82.

17. Webb A. Statistical pattern recognition. Arnold; 1999.

18. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Trans Pattern Analysis Machine Intelligence. 2000;22(1):4–37.

19. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Analysis Machine Intelligence. 2005;27(8):1226–38.

20. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. In: Solla SA, Leen TK, Müller K-R, editors. Advances in neural information processing systems, vol 12. Cambridge, MA, USA: MIT Press; 2000.p.526–32.

21. Vapnik V. Statistical learning theory. John Wiley & Sons; 1998.

22. Mitchell T. Machine learning. McGraw Hill; 1997.

23. Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA, USA: MIT Press; 2001.

24. Weka. http://www.cs.waikato.ac.nz/ml/weka/.2012

25. Taverna. http://www.taverna.org.uk/. 2012

26. Knime. http://www.knime.org/.2013

27. RapidMiner. http://rapid-i.com/.2012

28. Zoppis I, Gianazza E, Borsani M, Chinello C, Mainini V, Galbusera C, et al. Mutual information optimization for mass spectra data alignment. IEEE/ACM Trans Comput Biol Bioinf. 2012;9(3): 934–9.

29. Alonzo TA, Pepe MS. Development and evaluation of classifiers. Methods Mol Biol. 2007;404:89–116.

30. Solasso J, Jacot W, Lhermitte L, Boulle N, Maudelonde T, Mang A. Clinical proteomics and mass spectrometry profiling for cancer detection. Expert Rev. Proteomics. 2006 Jun;3(3):311–20.

31. Zoppis I, Borsani M, Gianazza E, Chinello C, Rocco F, Albo G, et al. Analysis of correlation structures in renal cell carcinoma patient data. In Bioinformatics. SciTePress, 2012. p.251–6.

32. Kolaczyk ED. Statistical analysis of network data: methods and models. Springer; 2009.

33. Zoppis I, Mauri G. Clustering dependencies with support vectors. In: Oscar C, et al., editors. Trends in intelligent systems and computer engineering. Lecture Notes Electrical Eng 6; Springer; 2008. p.155–65.

34. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem. 2010 Feb;82(3):833–41.

35. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 012;11(6):O111.016717.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

For citation purposes: *Di Silvestre D, Zoppis I, Mauri G, Mauri PL.* Stratification of biological samples based on proteomics data. OA Bioinformatics 2013 Aug 01;1(1):1.