Peakbin Selection in Mass Spectrometry Data Using a Consensus Approach with Estimation of Distribution Algorithms

Rubén Armañanzas, Yvan Saeys, Iñaki Inza, Miguel García-Torres, Concha Bielza, Yves van de Peer, and Pedro Larrañaga

Abstract—Progress is continuously being made in the quest for stable biomarkers linked to complex diseases. Mass spectrometers are one of the devices for tackling this problem. The data profiles they produce are noisy and unstable. In these profiles, biomarkers are detected as signal regions (peaks), where control and disease samples behave differently. Mass spectrometry (MS) data generally contain a limited number of samples described by a high number of features. In this work, we present a novel class of evolutionary algorithms, estimation of distribution algorithms (EDA), as an efficient peak selector in this MS domain. There is a trade-of f between the reliability of the detected biomarkers and the low number of samples for analysis. For this reason, we introduce a consensus approach, built upon the classical EDA scheme, that improves stability and robustness of the final set of relevant peaks. An entire data workflow is designed to yield unbiased results. Four publicly available MS data sets (two MALDI-TOF and another two SELDI-TOF) are analyzed. The results are compared to the original works, and a new plot (peak frequential plot) for graphically inspecting the relevant peaks is introduced. A complete online supplementary page, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms, includes extended info and results, in addition to Matlab scripts and references.

Index Terms—Mass spectrometry, EDA, feature selection, biomarker discovery.

1 Introduction

Estimation of distribution algorithms (EDAs) are a novel class of evolutionary algorithms that emerged as a natural alternative to classical genetic algorithms (GAs). EDAs turn the population statistics to their advantage and eliminate the need for the crossover and mutation operators used by traditional GAs. EDAs have produced competitive results in many domains [1], [2], and they have already demonstrated this potential for tackling high-dimensional data problems in the field of computational biology [3].

Throughout this work, we propose and explore a population consensus on top of the general EDA scheme to deal with another recent bioinformatics problem, the discovery of biomarkers in mass spectrometry data. Due to the small ratio between samples and MS data readings, this consensus approach enhances the robustness of the results. Originally developed by Karas et al. [4], matrix-assisted laser desorption/ionization (MALDI) technology can simultaneously measure peptide abundances in a given sample (e.g., serum, plasma, or urine samples) by enzymatically digesting the sample and running it through a mass spectrometer device. To the same end, Hutchens and Yip (1993) [5] introduced a variation in the way the sample is attached to the chemical matrix and named it surface-enhanced laser desorption/ionization (SELDI).

Both techniques are usually coupled by a time-of-flight (TOF) detector. This detector measures not only the peptide abundance, but also the time each peptide takes to reach the spectrometer's detector. Samples analyzed by this platform produce what is generally known as SELDI-TOF or MALDI-TOF data spectra. These spectra sort the abundances based on the ratio of each peptide's mass to its charge, known as the mass-to-charge (m/z) ratio.

Petricoin et al. [6] were the first to use this technology to identify proteomic biomarkers in complex diseases. Since then, many authors have followed their example [7], [8], reporting promising results for inducing classification systems and even more interesting findings for further research on the biology of such diseases [9]. However, the analysis of this kind of data is still far from being standardized, and the scientific community is developing robust and novel methodologies [10], [11].

The physics of spectrometer devices biases their outcome, adding chemical noise, signal shifts, and artifacts that the subsequent analysis must deal with. As an initial contribution, this paper presents a full preprocessing pipeline to remediate all these unwanted side effects [12].

- *R. Armañanzas, C. Bielza, and P. Larrañaga are with the Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo, 28.660 Boadilla del Monte, Spain. E-mail: r.armananzas@upm.es, {mcbielza, pedro.larranaga}@fi.upm.es.*
- *Y. Saeys and Y. van de Peer are with the Bioinformatics and Systems Biology Group, VIB—Ghent University, Technologiepark 927, B-9052 Ghent, Belgium. E-mail: {yvan.saeys, yves.vandepeer}@psb.vib-ugent.be.*
- *I. Inza is with the Department of Computer Science and Artificial Intelligence, Facultad de Informática, University of the Basque Country, Paseo Manuel de Lardizabal 1, 20.018 Donostia—San Sebastián, Spain. E-mail: inaki.inza@ehu.es.*
- *M. García-Torres is with the Area of Languages and Computer Systems, Pablo de Olavide University, Edificio 11—Conde de Aranda, Carretera de Utrera Km. 1, 41.013 Sevilla, Spain. E-mail: mgarciat@upo.es.*

Manuscript received 12 Oct. 2009; revised 20 Jan. 2010; accepted 25 Feb. 2010; published online 8 Sept. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2009-10-0184. Digital Object Identifier no. 10.1109/TCBB.2010.18.

The preprocessing ends with a peak profiling algorithm that identifies possible relevant points in each spectrum. These points, commonly known as *peaks* or *peakbins*, are the features whose values are used as the input of a complex feature subset selection procedure. This procedure finds which are the most relevant points.

The high dimensionality of MS data sets favors the use of stochastic search algorithms to look for relevant peaks over unfeasible exhaustive or limited greedy search strategies. Additionally, the usually small number of samples in this domain implies the need to use robust and reliable methods. Within this scenario, our second contribution in this paper is a population consensus on top of an EDA algorithm to find relevant peaks. Consensus approaches have reported good results on high dimensionality and noisy data in the past [13], [14], [15], especially in terms of reliability and low false-positive findings [16]. Specifically, the consensus propose allows an expert to select a confidence threshold and rely on findings above the set level only. The method's output is filtered by a multiobjective criterion that removes repetitive or not competitive results.

Finally, we introduce a new plot, called the *peak frequential plot* or *PF plot*. This plot graphically displays the results of a MS biomarker discovery procedure and aims to improve the interpretability of the results by a domain expert. Combining the phenotypic spectra in the PF plot, an expert can easily inspect the results and graphically identify new unexpected m/z points for further biological validation.

All the experiments have been carried out with meticulous care, embedding the tasks in a running schema, or workflow especially suited to the particularities of the MS data [17]. As spin-off results of this schema, we also discuss the consistency and stability of our results and how the classification estimation accuracies in a feature subset selection problem may overfit the training and test sets in use. We provide the community with a set of Matlab scripts containing an implementation of the proposed techniques.[1]

The paper is divided as follows: Section 2 introduces the set of tasks for preprocessing the original MS data. Section 3 presents the fundamentals of EDA and our consensus proposal for selecting relevant peakbins. A way to compare how consistent different peakbin sets are is also discussed. Section 5 details the data workflow from the initial data sets to the final consensus results. Sections 6 and 7, respectively, describe each of the four analyzed MS data sets and the running parameters used throughout the data analysis. Experimental results and a lengthy discussion are given in Section 8. Lastly, Section 9 sets out the conclusions and future trends of this work.

2 Data Preprocessing Pipeline

Within the MS domain, the preprocessing stage is an elementary and critical part of the design analysis protocol (DAP [17]). The DAP stage converts the data from its raw, initial form into a compact and homogeneous matrix forming the input for subsequent methods, such as machine learning or pattern recognition techniques. Thus, the main objective of the preprocessing task is to clean the data and detect the true signals in the noisy spectra.

MS data pose similar problems to most classical signal processing problems. Additionally, since the sample composition is often unknown or overly complex, the original signal decomposition is unknown. There have been attempts to mathematically model the *true* signal in a MS experiment but with limited or no success. The most accepted formulation is shown in Equation 1:

$$f(t) = B(t) + N \cdot S(t) + \varepsilon(t). \quad (1)$$

The first term $f(t)$ is the observed signal. $B(t)$ is a visually identifiable additive baseline component, and $S(t)$ is the expected true signal, which is modified by a normalization factor N. The last element $\varepsilon(t)$ is an unknown noise component that groups the remaining variations.

Although there is no standard preprocessing pipeline for MS data, the most accepted dataflow core stages are: baseline removal or correction, interspectra normalization, signal noise reduction or smoothing, peak detection, and finally peak alignment. Since there is no standard preprocessing pipeline, we have reviewed the state-of-the-art methods and decided which are the most suited to our data domain. In the cases where selected methods have been modified or augmented, we include a brief description of the changes.

In order to remove the low-range noise, we propose the use of the top-hat filter operator [18], [19] as the baseline correction method. A normalization task converts all the spectra signals to the same intensity ranges, so a fair comparison can be made among them. The use of local estimators over m/z windows with rescaling to the median value of the total ion count (TIC) is suggested for this aim [20]. The next processing step consists of smoothing the signal wave from the input signal to avoid the low signal fluctuations. The most common signal smoothing technique is wavelet denoising proposed by [12], [21], and implemented in the *Cromwell package*.[2]

The following task comprises the identification of peaks in the signal, or peak detection. This detection is individually applied to each separate spectrum, and then, a list of candidate peaks is retrieved for each spectrum. The peak detection algorithm proposed in [22] is borrowed as the starting point. To make the detection more restrictive, we have included two constraints into the algorithm: 1) the signal value of a candidate peak must be higher than a sensitivity threshold, and 2) a candidate peak must have an SNR higher than or equal to 3 within its associated intensity window [23]. The peak SNR is computed as the ratio between the point's height and the median absolute deviation (MAD) in the window under consideration, as suggested in [24].

Lastly, the peak assembly or alignment tries to match similar peaks detected across all spectra. There is no definite order in which this and the former (peak detection) tasks should be performed: peak alignment followed by peak detection [12] or vice versa [25]. To overcome signal shifts and potentially isotopic formations or very close compounds, we propose to assemble peakbins of different widths. Our preprocessing pipeline uses the Pearson linear

1. See *supplementary content* page, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms.

2. http://bioinformatics.mdanderson.org/cromwell.html.

$g \leftarrow 0$.
$D_g \leftarrow$ Generate and evaluate M random individuals (the initial population).
do
  $D_g^S \leftarrow$ Select $N \leqslant M$ individuals from D_g according to a selection method.
  $p_g(\mathbf{x}) = p(\mathbf{x} \mid D_g^S) \leftarrow$ Estimate the joint probability distribution of the selected individuals.
  $D_{g+1} \leftarrow$ Sample and evaluate M individuals from $p_g(\mathbf{x})$ (the new population).
  $g \leftarrow g + 1$.
until A stopping criterion is met.

Fig. 1. Main scheme of the estimation of distribution algorithm approach.

correlation coefficient to group the peaks into peakbins, as the computation time and memory demands are much lower than the classical clustering approaches[17], [20], [26]. Peakbins are scanned recursively and their signal values are quantified as the maximum value found in the bin [22]. The stopping criterion is met when there is no single peak or peakbin that shows a correlation value greater than a given threshold.

Each chosen method of the preprocessing is in detail discussed and documented in the *supplementary content* page, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms. In addition to the cited tasks and methods, there exists other additional tasks such as outlier detection [18], [27] and raw signal binning [27], [28]. The reader interested in alternative preprocessing can consult the *state-of-the-art* literature [29].

3 Peakbin Selection through Estimation of Distribution Algorithms

3.1 EDA Basics

Estimation of distribution algorithms [1], [2], [30], [31], [32] are a class of stochastic, iterative sampling techniques, related to the well-known field of genetic algorithms [33], [34], [35]. EDAs overcome the problems of having to tune multiple parameters in traditional GAs caused by the recombination and mutation operators. Instead of using these operators on explicit representations of population members, EDAs proceed by estimating the joint probability distribution of the promising solutions. Based on such a distribution, the next population is generated by sampling it. Fig. 1 shows the general running scheme of an EDA.

The initial population D_0 contains M individuals generated at random. The population is then sorted using an evaluation function, and a number N of individuals are chosen. Using the values of those N individuals, an n-dimensional (where n refers to the number of features or variables) probabilistic model $p_g(\mathbf{x})$ is induced. To generate the next population, M new individuals are sampled from the learned probabilistic model and evaluated again. The scheme is repeated until a stopping criterion is met.

The main characteristic that sets apart current EDA procedures is how the probability distribution $p_g(\mathbf{x})$ is learned. It is not affordable to compute all the parameters needed to specify the full probability model. Thus, the different EDA families must assume different factorizations according to a probability model and to the problem dimensionality. Based on these assumptions, EDAs can be divided into univariate, bivariate, or multivariate families. A complete taxonomy of these families can be found in [3].

3.2 Feature Selection Using an UMDA Population Consensus

Of the currently developed factorizations of EDAs, the simplest approach is the univariate marginal distribution algorithm (UMDA) [31]. UMDA factorization is usually suited to high-dimensional problems in which the possible relationships among the problem variables are unclear. In fact, this technique assumes that the probability distribution of each feature is marginal, that is, no dependence between the problem variables is taken into account when learning the factorization. Thus, the n-dimensional joint probability distribution factorizes as a product of n univariate and independent probability distributions:

$$p_g(\mathbf{x}) = \prod_{i=1}^{n} p_g(x_i).$$

This formulation implies that the learning process is fast compared to other more complex models. Moreover, UMDA scalability is one of its best characteristics because it has a running complexity of $n \cdot M$ for the learning process and of $M + \sum_{i=1}^{n}(k_i - 1)$ in memory requirements (k_i is the number of states for feature x_i). A full running example of an UMDA can be found in [36].

Good results have been reported for UMDAs used to address feature subset selection [37], especially within the computational biology field [3], [38]. The UMDA algorithm can be easily adapted to search for relevant features in a supervised classification domain by setting up the following elements:

- *Genotype encoding*: Each individual (or candidate feature subset) is represented as a binary array of size n. Each position of the array maps each problem's variable. A value of 1 implies that the respective variable is selected, whereas a value of 0 denotes that a variable is left out.
- *Evaluation function*: The evaluation function for ranking the merit of each individual is the classification accuracy estimated by a *k-fold cross-validation* process.
- *Stopping criteria*: The stopping criterion is either to achieve a perfect classification (100 percent accuracy estimation) or to have reached a fixed number of generations g.

This scheme is known as *wrapper feature selection* because it includes the classification process [37]. The final output of the algorithm is the best individual in the search, i.e., the feature subset that achieved highest accuracy.

It is very worthwhile to analyze what the selection tendency is over the evolved populations and to investigate if the selected set of features is robust [39]. Especially in problems with great many features, like MS data analysis and other bioinformatic problems [3], it is advisable to enhance the robustness and reliability of the selection of

relevant peakbins. The classical UMDA has to be adapted to achieve higher rates of robustness. Therefore, we propose building a hierarchy of the best solutions found throughout the search, instead of keeping just one best solution. These consensus approaches have already been reported to perform well on similar problems [37].

This improvement to the basic algorithm keeps all the best individuals found in the search and evaluates which are the features that have been flagged as selected throughout those solutions. Formally, given a set of solutions S consisting of r individuals, $S = \{\mathbf{x}^1, \ldots, \mathbf{x}^r\}$, of the form $\mathbf{x}^j = (x_1^j, x_2^j, \ldots, x_n^j)$ with $x_i^j \in \{1, 0\}$, the consensus solution over S with a confidence level T ($T \leq |S|$) is defined in (2):

$$\begin{aligned} \mathbf{x}_T^C(S) = \left(x_{1,T}^C, \ldots, x_{n,T}^C\right) \text{ with } x_{i,T}^C = 1 \iff \\ \iff \sum_{j=1}^{|S|} \delta\left(x_{i,T}^j, true\right) \geq T, \end{aligned} \tag{2}$$

where

$$\delta\left(x_{i,T}^j, true\right) = \begin{cases} 1, & \text{if } x_i^j = 1, \\ 0, & \text{if } x_i^j = 0\ . \end{cases}$$

The consensus solution $\mathbf{x}_T^C(S)$ contains the features in S that are selected at least T times. Obviously, the maximum value for T is $|S|$. This corresponds to the features that have always been flagged as selected in S. By decreasing the value of T, a hierarchy of consensus solutions can be built. This hierarchy fulfills the inclusion property, stating in this case that given S and any $T_0 \leq T$, the following implication $x_{i,T}^C = 1 \Longrightarrow x_{i,T_0}^C = 1$ holds for all $i = 1, \ldots, n$.

As the value of T decreases, the number of features selected in the consensus solution should increase. In fact, the addition of new features helps the search procedure to not get trapped in a local optimum and other parts of the search space are also considered. Thanks to this flexibility, a whole range of subsets can be evaluated instead of just a single one. Therefore, the user can set up a maximum and a minimum value for T, and the procedure will output all the consensus solutions within that range. Thus, the features returned by this consensus approach are expected to be the most reliable and, at the same time, best suited to the classifier used by the wrapper evaluation. Section 4 introduces two different consistency metrics to properly analyze the consistency of the population consensus method against the classical procedure.

4 Consistency Measures and Stability Index

Stability analysis is a recent topic in the feature selection domain [40], [41]. The main aim of stability analysis is to provide a means to state whether the features selected by a given selection approach are robust to changes in the data. In domains where knowledge discovery is a key objective, the stability of the selected features is a highly desirable property. The currently available stability studies rely on the concept of consistency between solutions. A consistency measure between two different subsets of selected features quantifies the degree of (dis)similarity between the two subsets. There exist different ways to measure this consistency, and different interpretations of the measures in terms of the source of the compared solutions: solutions could come from different runs of the same algorithm or from runs of different algorithms.

In stochastic searches, two subsets (A and B) seldom contain an equal number of features. A consistency metric should deal with this effect and be able to analyze subsets of different sizes. In principle, this difference in size should be a penalization term. In this scenario, we present two different metrics to measure consistency and show how to combine them into a stability index.

Let X be the set of available problem features and A and B two subsets of it, $A, B \subset X$. Further, let $n = |X|$ denote the number of features or cardinality of the set X. Let $|A| = k_A$, $|B| = k_B$, and $r = |A \cap B|$ be the cardinalities of the subsets A, B, and $A \cap B$. It is possible to define a consistency index between two subsets A and B of different sizes k_A and k_B by adapting Kuncheva's original metric [41]. The difference in size is taken into account in the index by selecting the highest cardinality between the two subsets, $k_M = \max\{k_A, k_B\}$. Kuncheva's consistency index can then be reformulated as

$$I_K(A, B) = \frac{rn - k_M^2}{k_M(n - k_M)}.$$

Despite the different sizes of A and B, there exists another consistency index able to compare feature subsets of different sizes. Usually known as the Jaccard similarity coefficient, it has already been used to measure consistency in feature selection problems [40]. The Jaccard index is based on comparing the number of common features in A and B and the total number of selected features:

$$I_J(A, B) = \frac{r}{k_A + k_B - r}.$$

The boundary limits of both indices are different: I_K varies between -1 and 1, while I_J ranges from 0 to 1. Therefore, it is not possible to directly compare their values.

The stability index is defined as a metric for comparing the consistency between a set of rather than just two solutions A and B. The mathematical formulation of this metric is straightforward: compute the average of all pairwise consistency measures. Therefore, given a set of solutions $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$, the stability among them can be computed as

$$\Sigma(\mathbf{S}) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} I(S_i, S_j),$$

where $I(S_i, S_j)$ is one of the two possible consistency indices presented above: I_K or I_J. The stability results of the comparisons between the feature sets returned by the classical UMDA and our consensus approach are detailed in Section 8.2.

5 Data Analysis Workflow

A data analysis workflow (DAW) refers to the whole pipeline of tasks that a database under research follows [17]. This workflow is sometimes designed carelessly and with out much concern about the side effects it could have on the

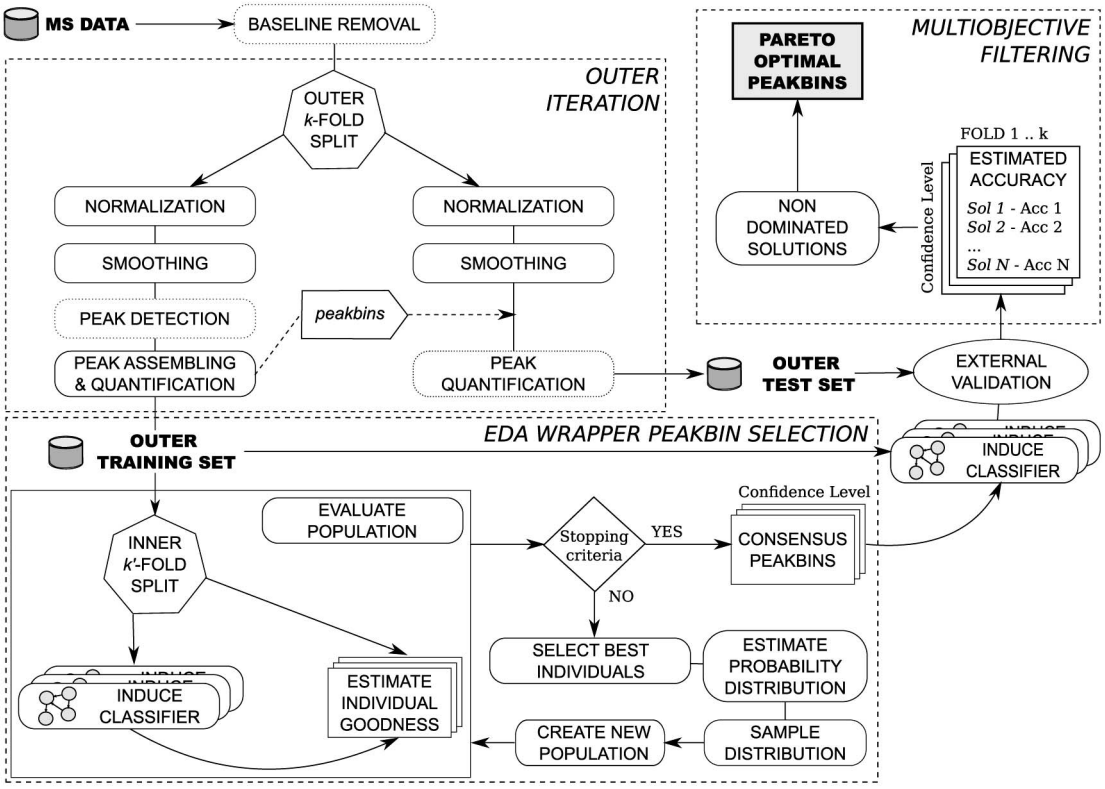

Fig. 2. Data analysis workflow. Tasks performed individually are included in boxes with dotted lines.

final results [42]. Critical DAW aspects that could potentially bias the results have been identified in the machine learning field. The most important include:

- Performance of any preprocessing task on the whole data set instead of first splitting the training from the test sets. In fact, if a workflow were to imitate a real scenario, the new cases would arrive at the end of the analysis [37].
- Setting the learning parameter values. This is especially tricky in wrapper schemes where the estimations are done on the same data that are afterwards used to train the model [43]. In a wrapper approach to feature selection, the feature selector accuracy must be estimated with a set of previously unseen instances [44].
- Previously setting a number of features to be kept could lead to overfitting. If we set the number of features to be retained, the feature selection algorithm is forced to look not only for the relevant features, but also for the features that achieve the highest accuracies when classifying phenotypes. The consequence is that the classification model is accurate in data sets with not many instances, but generalizability will be lacking when a new set of instances is provided [45].
- Procedures that include stochastic elements in their formulations should be run on different multistarts. Since stochasticity drifts apart from deterministic behaviors, a single run of such techniques does not guarantee the reliability of the outputs. This effect is usually coupled with the internal variance shown by different shufflings of the instances in a k-fold cross validation estimation [46].

Bearing in mind all the mentioned drawbacks, Fig. 2 introduces the data analysis workflow for the whole MS profiling experiment. It is designed to overcome the above issues and can be divided into three main parts. Before doing anything, the MS database should be baseline corrected. Since this task is independent for each spectrum (see Section 2), it can be considered as a separate task.

The first main part, namely the *outer iteration*, in Fig. 2 corresponds to the first k-fold split. To proceed with a fair estimation in the subsequent validations, the training and test sets should be completely separated from the very beginning [37]. Therefore, the *outer iteration* splits $k-1$ folds as the training set and keeps the remaining fold as the *outer test set*. After this division, the remaining preprocessing tasks are applied separately to the outer training and test set. Peakbins are detected and assembled only in the training set (left workflow branch), and the resulting peakbins are then quantified in the test set (right side). This external loop is repeated k times.

The second major part comprises wrapper peakbin selection using the proposed UMDA consensus approach. The main search scheme for a relevant subset of peakbins was introduced in Section 3. We refer to this part as the internal loop or *inner k′-fold split* or *validation*. The wrapper peakbin selector uses the classification accuracy estimation as the evaluation function to measure the merit of each individual (subset of peakbins) over the search. In the case that any other parameter of the classifier should be estimated, that estimation must be done within this inner fold [43]. After the inner search, a classifier is induced

taking into account only the values of the outcome peakbins and the outer training set. This classifier is then fairly evaluated with the outer test to output what we call the outer or external accuracy.

A comparison between the outer and the inner validations was performed. Large differences are observed between both, sometimes as much as a 5 percent in accuracy estimation. In addition and as expected, results show that the inner estimations have a low variance, whereas the variance in the outer estimation is up to an order of magnitude greater. This increase in the variance is explained by the fact that the inner models overfit to that fold's training set and because their generalization power considerably degrades when unseen instances are tested. Extended results and discussion are included in the *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms.

The values of k (outer) and k' (inner) could be different, but we recommend a low value for k' because one inner cross validation procedure is performed for each individual and each evaluated population. In the internal search, once the stopping criterion is met, we can keep just the solution produced by the classical UMDA approach, or we can apply the consensus peakbin approach. A confidence range $T_1 < T_2 < \cdots < T_t$ is then set, and a group of different solutions $\Phi_{T_1,T_t}(S)$ is collected from S: $\Phi_{T_1,T_t}(S) = \{\mathbf{x}^C_{T_1}(S), \mathbf{x}^C_{T_2}(S), \ldots, \mathbf{x}^C_{T_t}(S)\}$. This set of solutions is thus formed by different consensus solutions at different confidence thresholds (see Section 3.2). Each consensus solution is evaluated afterwards using the outer test set.

The input for the last part of the data workflow after all the external folds have been completed is all the accuracy estimations and the set of consensus solutions for each fold $k, \Phi_{T_1,T_t}(S_k)$. All this collected information can then be sorted by the confidence threshold values. To this end, for each confidence threshold $T_i \in \{T_1, \ldots, T_t\}$, we have k accuracy estimations achieved by k consensus solutions $\mathbf{x}^C_{T_i}(S_l)$ with $l = 1, \ldots, k$ and $i = 1, \ldots, t$. Note that the number of peakbins included in each solution is variable due to the intrinsic stochasticity of the UMDA approach. Thus, for a given confidence degree, there can exist two solutions with the same mean accuracy over the k folds but with a different number of peakbins.

The results of this consensus approach suggest using a multiobjective filter rather than forcing the selection of a single threshold or solution. It is very worthwhile studying how each confidence level solution behaves. To this end, there are four different objectives: the mean accuracy, its associated standard deviation, the average number of peakbins and also its standard deviation. Since there could be many solutions, we should keep only the really profitable ones. The next section presents the multiobjective dominance criterion used as the filter.

5.1 Multiobjective Sifter

The proposed DAW gives the expert the chance to study a full range of solutions instead of just one. Moreover, these solutions are the result of two conflicting criteria: the accuracy estimation and the size of the peakbin set (feature set). Previous studies on feature selection explored how the accuracy of the classification models evolves when the number of features increases or decreases. In general, these tendencies are dependent on the problem, however, it is generally accepted that the accuracy increases with the addition of features from an empty set until a size is reached where the accuracy no longer improves or even decreases.

Therefore, instead of using a single criterion to assess the goodness of a solution, we propose four different ones:

1. How large is the mean estimated accuracy?
2. How small is the average peakbin set size?
3. How low is the standard deviation associated with the estimated accuracy?
4. How low is the standard deviation associated with the peakbin set size?

Of two solutions with the same average size and mean accuracy, the one with the lowest variance for one or both of the objectives should be kept. All the above perfectly fits the concept of dominance [47]. Formally, a solution $\mathbf{u}$ can be expressed in terms of all the o objectives to be evaluated, $\mathbf{u} = (u_1, \ldots, u_o)$, where each u_i is the evaluation of the elements that form $\mathbf{u}$ in the ith objective. Within minimization, the dominance criterion states that a solution $\mathbf{u} = (u_1, \ldots, u_o)$ dominates another solution $\mathbf{v} = (v_1, \ldots, v_o)$, $\mathbf{u} \prec \mathbf{v}$, if

$$\mathbf{u} \prec \mathbf{v} \Longleftrightarrow \forall i \in \{1, \ldots, o\},\ u_i \leq v_i \text{ and } \exists j \in \{1, \ldots, o\} \mid u_j < v_j.$$

The set of nondominated solutions is known in operational research as the Pareto frontier or Pareto set [47], [48]. The Pareto frontier will only include the set of solutions $\mathbf{v}$ that are not dominated by any other solution $\mathbf{u}$. This Pareto set thus comprises all the solutions that cannot be improved for any objective without simultaneously degrading some other objective value.

Table 1 contains an example with six different solutions from a solution set S. Each solution is retrieved at a different confidence level T_i, and the four objectives are included. The first four solutions are nondominated, and they will be output as valid consensus approach solutions, while the last two will be removed because $\mathbf{x}^C_{T_5}(S)$ is dominated by $\mathbf{x}^C_{T_3}(S)$ and $\mathbf{x}^C_{T_6}(S)$ by the above four.

TABLE 1
Example of the Multiobjective Sifter for a Set of Six Solutions at Different Confidence Thresholds $T_1, \ldots, T_6$

	Accuracy	Acc. Std	Peakbins	Peak. Std	Pareto
$\mathbf{x}^C_{T_1}(S)$	94.5	2.3	15	4	✓
$\mathbf{x}^C_{T_2}(S)$	80.6	10.1	5	3	✓
$\mathbf{x}^C_{T_3}(S)$	96.0	1.8	30	10	✓
$\mathbf{x}^C_{T_4}(S)$	80.1	9.0	6	2	✓
$\mathbf{x}^C_{T_5}(S)$	96.0	1.8	31	10	×
$\mathbf{x}^C_{T_6}(S)$	80.0	11.0	40	15	×

The last column indicates which solutions are not dominated by any other and belong to the Pareto front.

6 Mass Spectrometry Data Sets

Four different data sets have been used to illustrate the presented peakbin selection processing. Two are from a SELDI, whereas the other two are from a MALDI spectrometer. The number of samples, phenotypes and available m/z readings varies noticeably. All these data sets are

TABLE 2
Data Sets Summary Details

	No. spectra	No. readings	Values range	Reference
Ovarian cancer profiling (OVA)	200 (121+79)	45,200	700.116 – 12,000	[6]
Detection of drug-induced toxicity (TOX)	62 (28+34)	45,200	799.115 – 12,000	[49]
Hepatocellular carcinoma (HCC)	150 (78+72)	36,802	799.725 – 9,999.975	[50]
Detection of glycan biomarkers (DGB)	128 (78+25+25)	16,075	1,499.8 – 5,518.3	[9]

Column No. spectra *includes the total number and number of spectra for each phenotype The rest of the columns illustrate the number of readings coupled with the minimum and maximum m/z values. The original publication is also referred*

available at the sites of their respective authors [6], [9], [49], [50]. Unfortunately, there is currently no uniform storage protocol for this kind of data. Thus, we had to use a parsing algorithm to adapt the original raw data files. None of the provided plain text files for all the data sets share the same m/z axis, even within the same data set. So, we had to set a resolution of 0.025 and average all points over their maximum and minimum values using bins of this width. When there were no values available for an interval, a null value was assigned. A summary of each data set's characteristics is presented in Table 2. For a full description and the Matlab data files, the reader can consult the *supplementary content* page, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms.

7 Running Parameters

For reproducibility purposes, we detail all the parameter settings used in our experiments. The full DAW can be divided into two main steps: the preprocessing stage and the UMDA peakbin consensus selection. While many of the preprocessing parameters are common for all the data sets, there are some of them that need to be tuned individually though. All these preprocessing parameters can be fully reviewed in the *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms.

Regarding the rest of the running parameters, note that the number of times the outer loop k is performed produces a large increase in the total computational time and, even more so when the number of inner folds k' is high. Our running scheme uses $k = 5$ folds on the outer loop and $k' = 5$ inner folds on each individual accuracy estimation. For this wrapper accuracy estimation, the classification model is a continuous naïve Bayes [51] with conditional normal density distribution for the features. This classification model is used to estimate the goodness of the features in the EDA population search and so other classification paradigms could be used for the same purpose. For clarity, we only refer to the results provided by the naïve Bayes model, as similar results were obtained with other models (see *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms).

The initial population of each UMDA selection is randomly drawn from a Bernouilli distribution with a success probability $p = 0.1$ of each peakbin being initially selected. This was found to be the best value for reaching a compromise between the number of selected peakbins and their performance in a wrapper selection. There are two stopping criteria: either to achieve a 100 percent accuracy estimation, or to reach a hundred generations. Each population is formed by 100 individuals and the truncation threshold is set at 50 percent as usual. The UMDA consensus was implemented using the Matlab toolbox for Estimation of Distribution Algorithms (MATEDA-2.0), available online [52].

To output the set of consensus solutions, the confidence range is set up with confidence levels of $T_1 = 10\%$ and $T_t = 100\%$ with a 1 percent step. This means that, for the set of best individuals, features selected less than 10 percent of the times are rejected. In the outermost case, 100 different best individuals are collected (one per population), and the total number of consensus solutions on each outer fold also reach 100. Since all these solutions are sifted by the multiobjective filter, only those belonging to the Pareto front will be retained as valid results.

8 Results and Discussion

Stochastic approaches take advantage of their random search policies to inspect the search space. However, this is a drawback rather than an advantage when a precise result is required: the need to repeat the search approach several times [45]. Usually known as multistart or rerun, the aim of a systematically run repetition is to find a stable outcome.

In our case, this random component is present in several stages of the DAW, namely, in each outer and inner fold, in each initial population and in the stochastic behavior of the UMDA itself. Thus, the multistart run is a must rather than a choice. To avoid this intrinsic variance, all the results presented throughout this section are extracted from a set of 500 multistart runs for each of the analyzed MS data sets.

8.1 Multistart Nondominated Solutions

Like most search procedures, the UMDA-wrapper peakbin selection only outputs a single solution. On many occasions, the search procedure could have explored parts of the search space with good solutions that, however, do slightly worse on the evaluation and are, thus, discarded. The retrieval of this useful information is the aim of the population consensus proposed in Section 3.2. Its first advantage is that whereas the classical scheme only retrieves one solution, the consensus approach may produce as many solutions as populations have been generated in a single run. Many of these solutions may be similar or even equal. Hence a filtering process is required to output the really interesting solutions. In our case, this sift is the nondominance criterion with respect to the four objectives defined earlier.

The first row of Table 3 presents the total number of nondominated solutions that have been reported throughout the whole set of multistart runs. This is tens of thousands for

TABLE 3
Descriptive Overview of the Multistart Results Produced by the Population Consensus Proposal

	OVA	TOX	HCC	DGB
Total number of solutions throughout 500 runs	72,744	66,277	62,597	38,735
Mean number of solutions on each Pareto front	35.15	29.25	27.16	16.26
Mean number of peakbins per Pareto solution	97.64	155.34	25.28	11.74
Maximum accuracy	100±0	93.84±6.43	97.33±2.79	95.29±5.06
Peakbins	39.80±17.06	114.60±60.08	20.60±19.03	8±4

The first row indicates the number of nondominated solutions collected for all the runs. The mean values represent the mean number of nondominated solutions per run and the mean number of peakbins in each solution. The last two rows show the accuracy and mean number of peakbins associated with the best solution found by the consensus.

all data sets, whereas the classical UMDA approach reports only a total of 500 solutions (one per run). The second and third row show the mean number of solutions per run, as well as the mean number of peakbins in each solution, of the Pareto front.

The estimated accuracy of every one of the nondominated solutions is also computed by the validation on the outer test sets. The difference from the estimation output by the classical UMDA approach is clear. The consensus approach is able to find solutions that outperform the UMDA approach accuracy estimations for all data sets (see the outer accuracies in the *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms for comparison). For the OVA data set in particular, the estimated accuracy reaches a value of 100 percent for the fivefold estimator in use. For the other three data sets, the mean estimator also reaches competitive percentages: 93.84 percent, 97.33 percent, and 95.29 percent. Nevertheless, if we take the standard deviation into consideration, there are some folds for which accuracy is also 100 percent. Notice that this variance is numerically similar to the values reported by the classical UMDA approach.

The *peakbins* row in Table 3 shows the average number of peakbins included in the best consensus solution (*maximum accuracy*). An in-depth analysis of this characteristic illustrates another interesting effect: the parsimonious behavior of our consensus approach. Fig. 3 presents, for each multistart run and for each nondominated solution of each of these runs, the average number of peakbins. The side color map adds a fourth component to the plot: the mean accuracy achieved by each solution. The first conclusion from the charts is that the solutions with fewest peakbins do not achieve a good classification accuracy. However, when more peaks are added, the solutions achieve significant accuracy levels. It is when this number of newly added points increases that the parsimonious behavior [53] is observed: accuracy does not improve as a consequence. Since all the new peakbins are relevant for the problem, classification accuracy is not harmed. In terms of phenotype separability, however, the new points add no new information. The different number of nondominated solutions in each run is clearly explained by the stochasticity discussed at the beginning of this section.

8.2 Peakbin Stability Comparison between the Consensus and the Classical UMDA Approach

Apart from classification power, it is worthwhile analyzing how stable the nondominated solutions are compared to the regular solutions output by the classical UMDA. A general stability index Σ was introduced in Section 4, and two different consistency measures (I_K and I_J) were also presented.

A consistency measure is used to quantify the (dis)similarity degree between two subsets of features in a feature subset selection problem. High levels of consistency between both subsets suggest that the feature selection approach is highly stable, a desirable behavior in knowledge discovery tasks.

When there are more than two subsets (solutions), we can compute the stability degree Σ among all of the subsets as the average of all pairwise consistency comparisons. Notice that when there is a relatively high number of solutions, the combinatorial number of comparisons could lead to an unfeasible computational time.

The number of solutions that the consensus approach outputs prevents us from computing the global stability value by inspecting all possible combinations (see the total number of solutions in Table 3). Therefore, we propose analyzing stability by averaging the stability values of each multistart run rather than mixing the solutions from different runs.

As the multistart runs are based on five external folds, the classical UMDA solution selects five different peakbin sets for each i-th run, $\mathbf{S^i} = \{S_1^i, \ldots, S_5^i\}$. After choosing one of the two consistency measures, we can then compute the stability of this solution set in the ith run, using the stability index, $\Sigma(\mathbf{S^i})$. Assuming B different multistart runs, the mean stability value $\Im$ is calculated straightforwardly as

$$\Im = \sum_{i=1}^{B} \frac{1}{B} \Sigma(\mathbf{S^i}).$$

In the case of the consensus approach, there is a variable number of nondominated solutions per fold and run. To compare all these solutions fairly, we first need to select a representative solution from each Pareto set. We have chosen two criteria: 1) the solution that achieves the highest accuracy in each fold (if there is a draw, the one with fewer peakbins is selected) and 2) the solution that includes the maximum number of peakbins. Once the solutions are retrieved, the mean stability value is computed.

Results of all the mean stability values are set out in Table 4. Rows under $\Im_{I_K}$ and $\Im_{I_J}$ refer, respectively, to the average stability values using I_K and I_J consistency measures. Since all these values are based on averages, it is possible to statistically compare their differences using a t-test of equal means. All the comparisons between the classical UMDA and the consensus values (the highest accuracy or the maximum number of peakbins) are significant at a significance level of $\alpha = 0.01$.

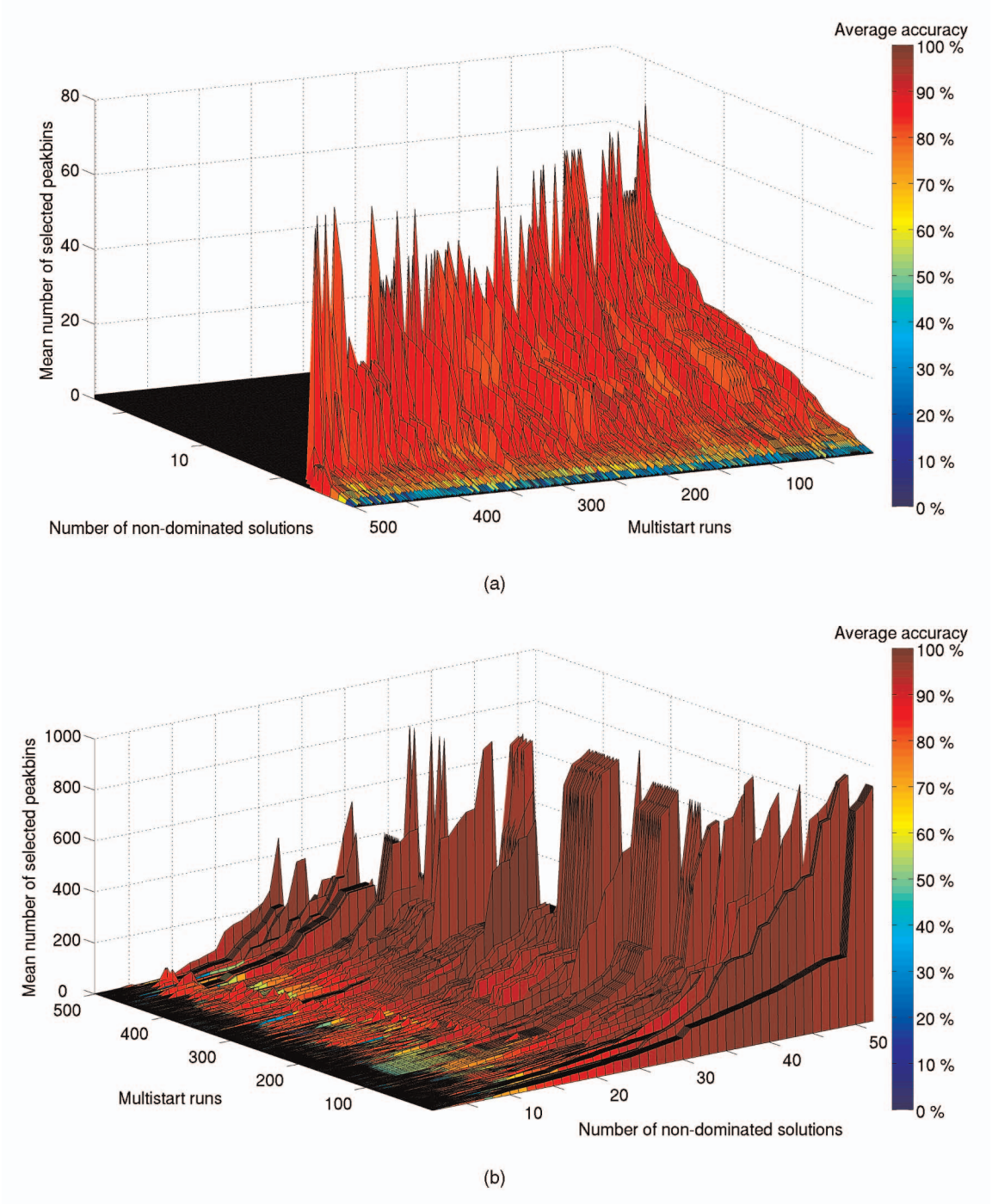

Fig. 3. Graphical representation of the nondominated solutions collected in the multistart run process for two of the data sets. (a) The DGB data set results, whereas (b) the results for the OVA data set. In (a) and (b), the Z axis presents the mean number of peakbins that each solution contains. The X and Y axis have been switched between (a) and (b) for clarity. The color map of the surface represents the average accuracy estimated for each one of the points (solutions).

From the results, the first observation is the difference in the stability index between the consensus solutions in the highest accuracy and the largest number of peakbins. The low stability for most accurate consensus solutions is a consequence of the high variance in the number of peakbins in each solution. As discussed previously, the most accurate solutions only include the features necessary to tackle the classification problem (parsimonious tendency or Occam's razor). As a consequence, solutions from different folds can differ significantly.

Looking at the largest solutions (max size), however, the stability index improves significantly. Notice that the largest solutions are nondominated, so their accuracy performance is mostly expected to be as good as the most accurate solutions (see Fig. 3 and discussion on Section 8.1). The stability improvement is due to the fact that the addition of

TABLE 4
Mean Stability Values $\Im$ Computed in Terms of the I_K and I_J Consistency Measures

	OVA	TOX	HCC	DGB
$\Im_{I_K}$				
$\mathbf{S}$ = consensus (best acc.)	0.0721	0.0087	0.2114	0.3678
$\mathbf{S}$ = consensus (max size)	0.2181	0.1466	0.3696	0.3040
$\mathbf{S}$ = classical	0.1217	0.0647	0.3045	0.2721
$\Im_{I_J}$				
$\mathbf{S}$ = consensus (best acc.)	0.0662	0.0251	0.1722	0.3362
$\mathbf{S}$ = consensus (max size)	0.1720	0.1406	0.2680	0.2243
$\mathbf{S}$ = classical	0.0888	0.0680	0.2127	0.1945

Classical *rows present the values for the classical UMDA scheme. Values in* consensus *rows show the respective values for the most accurate set of solutions and for the solutions with the largest number of peakbins in each run, respectively.*

new peakbins in our consensus approach focuses on increasing the robustness of the selected set of variables.

Comparing the classical UMDA and the consensus stability results, we find that, in three out of four data sets, the classical UMDA solutions show higher consistency values compared to the more accurate but smaller and more diverse consensus solutions. However, when the classical UMDA solutions are compared to the largest nondominated consensus solutions, they are defeated in terms of stability for all the four data sets. In the cases of OVA and TOX, the stability gauged by both Kuncheva's and Jaccard's consistencies is doubled by population consensus, and, as previously pointed out, they always show statistically significant differences.

8.3 Peakbin Concurrences

Despite the quantitative performance in stability, the qualitative performance of the results should be also inspected. In this sense, the consensus approach brings us with a great number of selected peakbins. Therefore, graphical tools are needed to take advantage of all these results.

Let the term concurrence refer to the together inclusion of peakbins in a given run. In concrete for the consensus results, the total times a m/z point has been selected is computed by adding all its occurrences along each single nondominated solution and over all the external folds. These frequentials are then normalized and treated as percentages. Since the preprocessing of each fold retrieves slightly different m/z windows for each peakbin, the intersection of them may produce at times higher windows or disjoint ones.

There are peakbins with a high degree of concurrence which implies that they are included as relevant in the majority of the nondominated solutions. By decreasing the associated percentage of concurrence, such set of bins diversifies and not so frequent bins appear. In terms of stability, a stable set of peakbins will be the set formed by the bins many times and jointly included at the same time. In order to check these configurations, Fig. 4 displays which m/z points have been simultaneously included at least in 90 percent of the nondominated solutions.

From Fig. 4, it is easily identifiable some m/z windows or bands that are presented along almost all the runs. In the case of the OVA data set, there are three m/z bands, namely [1,034-1,036], [7,052-7,061], and [10,259-10,267], clearly represented. It is also noticeable the band in [3,961-4,012] but its width suggest the presence of several peakbins. The reader may identify other two bands, but their concurrences are not so important. The subfigure related to the TOX data set illustrates the high degree of variability this data set includes. We can find several bands in the figure but none of them seem to be important enough. This fact relates with the differences of stability indexes between this data set solutions and the rest (see Table 4 of previous Section 8.2).

Regarding HCC data set, two close bands significantly appear: [1,864.225-1,869.225] and [1,907.475-1,910.475]. A third one could be extracted from the window [934.225-937.475], although its concurrence level varies through all the runs. Also note that there are other selected m/z positions in the spectra but with a more sparse behavior. For the last data set, DGB subplot of Fig. 4 shows an undoubtful conformation of three bands. The first two are in [2,039.7615-2,041.7615] and [2,243.7615-2,244.7615]. The third one is located in the point 2,355.0115 but it could be confused with other points of the close bin [2,391.0115-2,392.0115].

All the mentioned m/z bands match some of the peakbins with highest percentages of occurrence throughout all the solutions and runs (see Section 8.4). In addition, the concurrence levels presented in Fig. 4 state that they almost always come in a joint manner. Therefore, not only is important the own set of peakbins, but, also, is the fact of happening together.

8.4 Knowledge Discovery Using the Consensus Results

The data mining and machine learning disciplines provide computational biology with powerful tools to help in the analysis, diagnosis, prognosis, and new knowledge discovery within data produced by high-throughput biological devices [54]. Analysis, diagnosis, and prognosis form what is known as personalized medicine. Although they are all in constant evolution, knowledge discovery is the topic that is most likely to enrich basic research and propose new hypotheses about complex biological problems or diseases.

Therefore, we consider that an optimization process, like the search for relevant or discriminative peaks in mass spectra data, must also comply with the proposal of new biological hypotheses for validation. Throughout this section, all the consensus multistart results will be graphically presented and discussed with respect to the original author findings.

For quick reference, we have designed a combined plot. This new plot, referred to as the *peak frequential plot* or *PF plot*, is formed by two overlapped subplots. The first subplot illustrates the absolute intensity differences between the mean spectrum of the different phenotypes. The second subplot includes the percentage of occurrence of each m/z position being selected as relevant. Applied to our approach, this percentage shows how many times each m/z position, in all the nondominated solutions, is added as a new relevant peakbin. Figs. 5 and 6 are examples of this peak frequential plot. When there are more than two phenotypes (as is the case of DGB), the top subplot is computed as the sum of all the pairwise differences between the mean spectrum of each phenotype.

The top subplot presents what could be considered as the simplest peak selector, whereas the second subplot sets out the results of the peak selection method. As we discuss

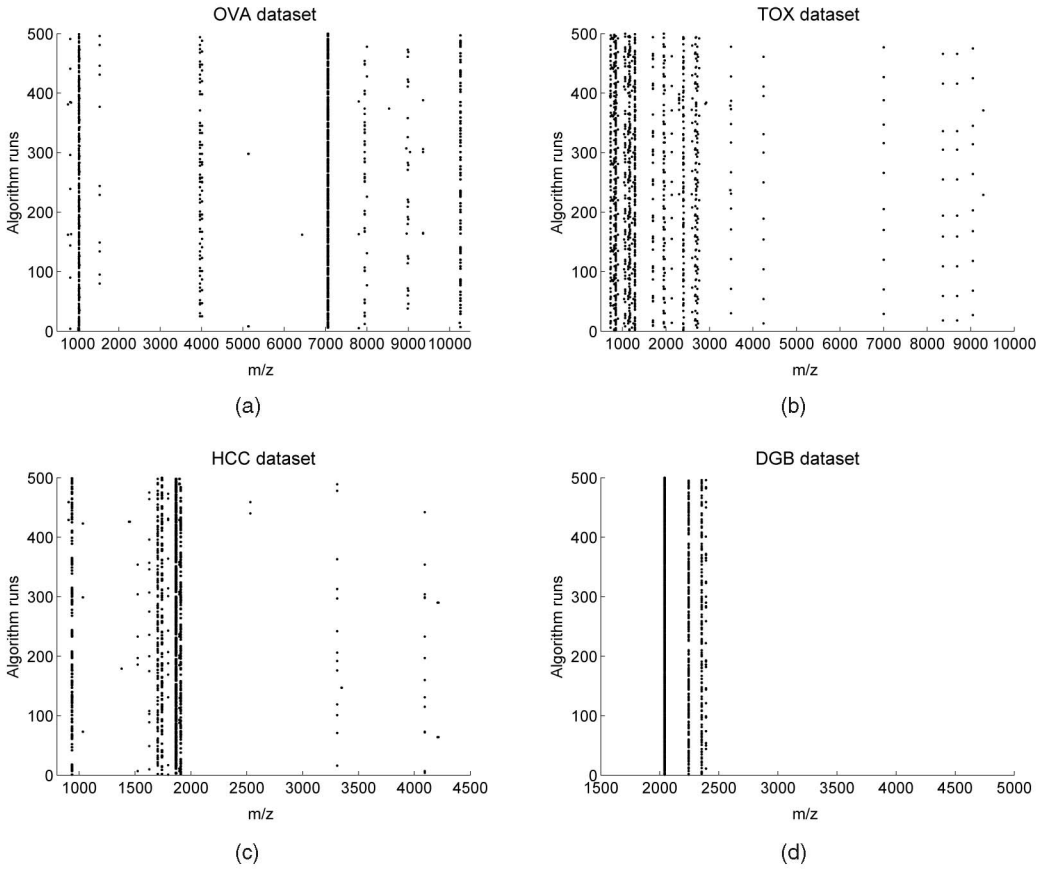

Fig. 4. Scatter plots of the most times included m/z points throughout all the experiments. The points reflects the bins that have been included at least in the 90 percent in all the consensus solutions. To simplify the computation, the solutions from different outer folds were taken together for each run.

later, the positions showing the largest differences are expected to be relevant for both methods. Even so, an expert may find a peakbin that is selected many times but whose average behaviors differ little more interesting.

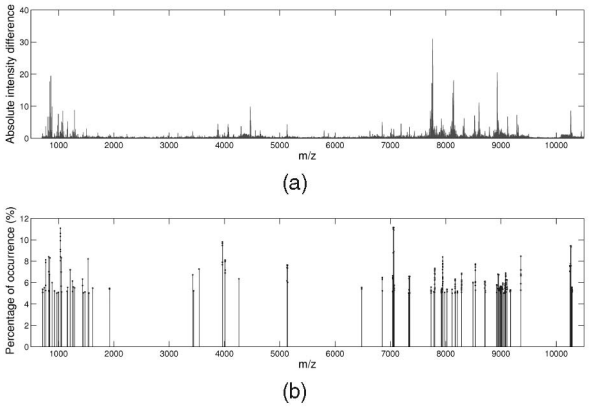

Fig. 5. Peak frequential plot for the OVA data set. (a) The absolute differences among the average spectra of each phenotype. (b) Sets out the results of the multistart consensus approach. It shows the percentage of occurrence of each m/z position being selected throughout the whole process (occurrences below 5 percent are not shown).

The original paper on the ovarian cancer profiling data set [6] reports a discriminative rule of five peaks that provided an almost perfect classification. Fig. 5 presents the PF plot of our consensus approach for the OVA data set. Already shown in Section 8.1, we are able to achieve the highest accuracy value in terms of spectra separability. However, our peakbin set did not compare with the set originally reported. The results reported by Petricoin et al. have been previously said to contain artifacts supposedly from an unfit denoising [42], [55].

Looking at Fig. 5, the [7,052-7,061] peakbin has the highest occurrence level, and its width could suggest a possible isotopic configuration. Other interesting values with large occurrences are [1,034-1,036], [3,961-3,963] and [1,025.9116-1,026.7366]. Lastly, notice that for the peakbin configuration at [5,131-5,142.6], the associated difference is small, whereas the bin is often selected.

The authors also aimed for a panel of only five predictive peaks for the TOX data set [49]. The original results are calculated based on a different sample distribution, so outcome of comparing their panel and our results might be slightly different. Nevertheless, our results are able to identify four out of five members of the suggested peakbin panel.

Since the preprocessing proposals are not equal, the observed intervals for the m/z axis do not exactly match in

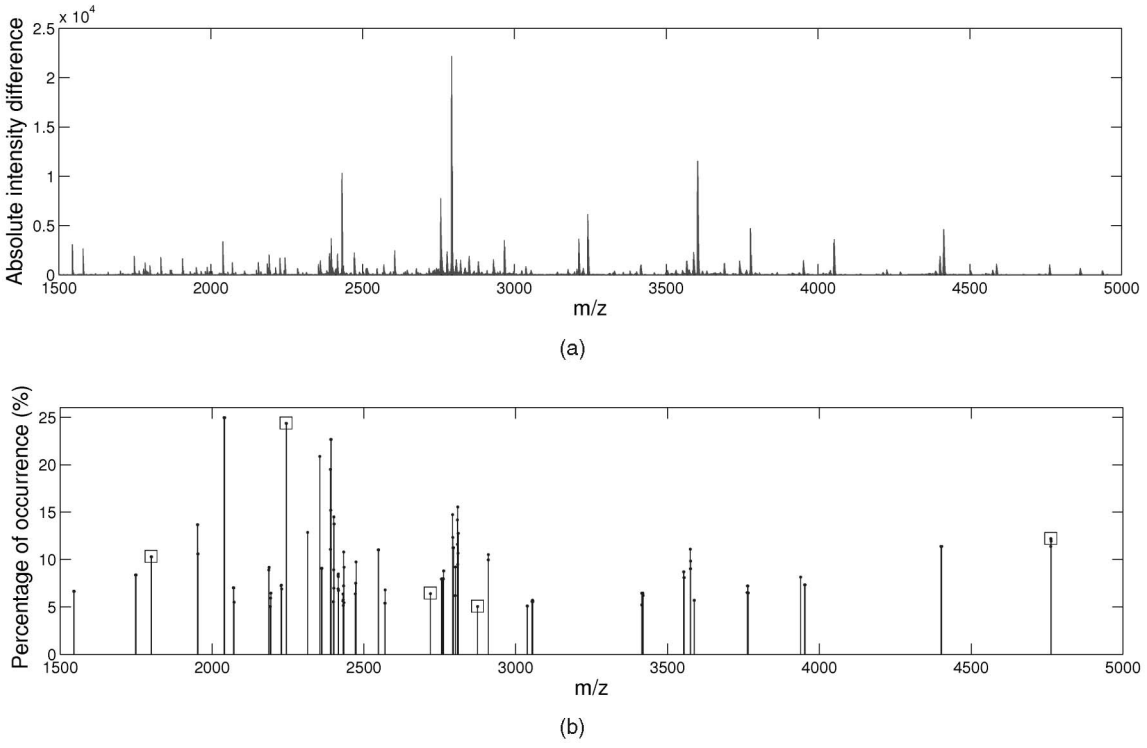

Fig. 6. Peak frequential plot for the DGB data set. (a) The absolute differences among the average spectra of each phenotype. (b) Sets out the results of the multistart consensus approach. It shows the percentage of occurrence of each m/z position being selected throughout the process (occurrences below 5 percent are not shown).

both studies. For instance, the peak at 810.33765 maps in our data to the bin [810.115-810.365] with an occurrence of 4.20 percent. Similarly, the peak at 981.8242 matches the [981.615-981.865] bin and has an occurrence of 3.50 percent. The original peaks at 1,987.9727 and 2,013.5771 are also detected as relevant by the multistart process but with an insignificant occurrence level.

The PF plot for this TOX data set is attached as *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms. The phenotype spectra have a high variance in this data set. As a consequence, the estimators in the classification have a high associated standard deviation: the classifier is able to achieve up to 100 percent accuracy in some folds, whereas, on the same run, accuracy for other folds is only 88 percent. Either way, the peak frequential plot shows other m/z values that seem to be of biological interest.

Results for the HCC hepatocellular carcinoma show a significant match. Ressom et al. presented a MS biomarker panel of six peakbins in the study of hepatocellular carcinomas triggered by viral infections. Our results reported a full coincidence with this six peakbin panel. Table 5 presents the original m/z bins, our corresponding m/z bin and the percentage of occurrence of each bin. Not only are all six values found to be relevant by our consensus approach, but the percentage of occurrence for these six is also remarkably high. Three of them present the highest occurrence values in the multistart process with a value of around 15 percent.

Apart from the above six relevant bins, the PF plot (see *supplementary content*, which can be found at http://www.sc.ehu.es/ccwbayes/members/ruben/ms) suggests that there are other relevant peakbins that may merit an in-depth analysis. A closer look at these peakbins suggests that, comparing the absolute difference and the percentage of occurrence, three, namely [1,445.725-1,454.475] with 10.50 percent, [3,307.975-3,309.725] with a 12.95 percent, and [4,206.975-4,216.225] with a 11.80 percent of occurrence, respectively, are noticeable. The density of bins surrounding the latter peakbin also may suggest a possible isotopic effect on that m/z position.

The last data set was produced with the aim of selecting glycan structures able to distinguish subjects from prelabeled groups. The authors [9] identify a panel of 10 markers with different frequencies in their results. When compared with our results, we find that 7 out of the 10 markers are also identified by the consensus proposal. Moreover, the percentage of occurrence in our multistart is also high for

TABLE 5
Original Relevant Peakbins Reported by [50] for the HCC Data Set

Original m/z bin	Current m/z bin	Occurrence (%)
933.6 - 938.2	933.475 - 938.225	14.77%
1,378.9 - 1,381.2	1,378.975 - 1,381.225	10.32%
1,737.1 - 1,744.6	1,737.225 - 1,744.475	15.13%
1,863.4 - 1,871.3	1,863.975 - 1,870.225	15.27%
2,528.7 - 2,535.5	2,528.725 - 2,535.475	12.07%
4,085.6 - 4,097.9	4,085.725 - 4,097.975	6.78%

For each bin, the second and third column map, respectively, our correspondent m/z relevant peakbins and the occurrence percentage of each bin in the multistart.

the most important ones. Fig. 6 shows the PF plot of our results. Notice that, of the seven bins in common, five are highlighted by boxes in the figure.

A careful analysis of Fig. 6 draws the attention to three more peakbins that, either because of their high occurrence, or because of a large difference in intensities, may merit further research. The bin at [2,039.7615-2,041.7615] has the highest percentage of occurrence with 25 percent, whereas the bin located at [2,792.0115-2,795.0115] is associated with the largest absolute difference. We would also like to point out bin [4,400.2615-4,4035.1149], which has both a large percentage of occurrence and a visible intensity difference among phenotypes.

All the above peakbins could be of interest for further biological examination. The peak frequential or PF plots are thus a general and powerful proposal for graphically identifying relevant peaks. An expert can easily check or point out some point(s) of interest when inspecting these figures. A PF plot gives a broader view of the results, and opens up prospects for subsequent wet-lab research.

9 Conclusion

On the basis of biomarker discovery in MS data, we find three important advantages. One is the fact that the sample in use is most of the times serum, plasma, or urine. Consequently the test for collecting the sample is almost noninvasive for the patient. Another issue is that the economic cost of a MS run is much cheaper than, for example, a classical cDNA microarray. Yet another good point is the possibility of looking for early stages metabolic markers, an unfeasible search in the microarray field. Nevertheless, there is a big pitfall still to be overcome. This is the fact that the MS profiling results are intrinsically noisy, nonconstant, and difficult to analyze.

As a first step in the search for relevant peaks in MS data, the user encounters the problem of preprocessing the raw data to minimize all the noisy and variance-related behaviors. To this end, we suggest the use of a full pipeline of tasks, including baseline correction, spectra normalization, smoothing, peak detection, and quantification. The preprocessing part of the analysis should be viewed as separate from the subsequent search for relevant peaks. Therefore, other preprocessing pipelines could be used.

Once the data is ready for a relevant peak selection task, the classical feature selectors are confronted with the so-called *curse of dimensionality*. In this context, we propose the use of stochastic policies that are suited to dealing with the high number of features for evaluation. The low number of samples implies that the search is not always as robust as it should be. To improve the reliability of the output relevant peaks, we propose a consensus scheme over the search population. One straightforward advance in robustness is that an expert can set a confidence threshold and rely just on findings above this limit. A multiobjective filter of the solutions outputs only those sets of peaks that are better in terms of phenotype separability power, small set sizes, or low variance in these two terms.

In addition, all the analysis is embedded into a workflow that imitates how all the tasks would behave when dealing with new and unseen samples. If there is no such workflow, results could be overfitted to the available data and may lose generalizability. Moreover, the results of this workflow also behave parsimoniously like supervised classification within feature subset selection procedures: small sets of features achieve good accuracy values, and these values are not improved when adding more predictive features.

Furthermore, our consensus approach allows us to study how stable the selection is. Stability results quantitatively illustrate how the consensus approach is able to retrieve significantly more stable solutions than the classical UMDA approach. As expected, if the practitioner decides to rely on only the most accurate subsets of peaks (usually of small size) then the variability component is large and, thus, stability is penalized.

Finding locations of interesting masses should be coupled to a subsequent knowledge discovery stage in which those peaks are studied and evaluated. To this end, this work presents a novel plot, the PF plot, to display the results of a peak selection method in supervised MS data problems. The new plot enables an expert to graphically explore the results and identify peaks of special interest. Although the presented PF plots include the results from the multistart runs of our consensus UMDA approach, they can be used by any kind of selection method. The relevant peaks found by our consensus approach closely match the peaks reported by the original works. By inspecting the PF plots of our results, we extended the original findings with a series of peaks that could be of interest for a more in-depth biological analysis.

Until this new EDA approach reaches wet-lab routine, some drawbacks need to be overcome. Although being independent from the preprocessing task, it should be crucial that the community finally reaches a consensus about what should be the pipeline of tasks a set of spectra must undergo. In terms of peakbin selection methods, there is a work niche in comparing all the methodologies presented throughout the state-of-the-art literature. As in other computational biology disciplines, it is unsure that a single method can prove it as being the best to accomplish this task. However, an honest comparison and different contrasts using the measures presented in here could shed light to a laboratory implementation of all these techniques.

Acknowledgments

The authors would like to thank Roberto Santana for his help with the EDA implementation. They would also like to thank the anonymous reviewers for their interesting comments and useful suggestions that have significantly improved the quality of this work. This work has been partially supported by the personal grant AE-BFI-05/430, the 2007-2012 Etortek, Saiotek, and Research Group (IT-242-07) programs (Basque Government), TIN2010-20900-C04-04, TIN2010-14931, TIN2008-68084-C02-00, Consolider Ingenio 2010-CSD2007-00018 projects (Spanish Ministry of Science and Innovation), the COMBIOMED network in computational biomedicine (Carlos III Health Institute), and the Cajal Blue Brain project (Universidad Politécnica de Madrid). R. Armañanzas is supported by a Juán de la Cierva postdoctoral fellowship (MICINN) and Y. Saeys would like to thank the Fund for Scientific Research Flanders (FWO) for funding his research.

REFERENCES

[1] *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation,* P. Larrañaga and J. Lozano, eds. Kluwer Academic Publishers, 2002.

[2] *Towards a New Evolutionary Computation: Advances on Estimation of Distribution Algorithms,* J.A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, eds. Springer-Verlag, 2006.

[3] R. Armañanzas, I. Inza, R. Santana, Y. Saeys, J.L. Flores, J.A. Lozano, Y. Van de Peer, R. Blanco, V. Robles, C. Bielza, and P. Larrañaga, "A Review of Estimation of Distribution Algorithms in Bioinformatics," *BioData Mining,* vol. 1, no. 6, 2008.

[4] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, "Matrix-Assisted Ultraviolet Laser Desorption of Non-Volatile Compounds," *Int'l J. Mass Spectrometry and Ion Processes,* vol. 78, pp. 53-68, 1987.

[5] T.W. Hutchens and T. Yip, "New Desorption Strategies for the Mass Spectrometric Analysis of Macromolecules," *Rapid Comm. Mass Spectrometry,* vol. 7, no. 7, pp. 576-580, 1993.

[6] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *Lancet,* vol. 359, no. 9306, pp. 572-577, 2002.

[7] M. Hilario, A. Kalousis, C. Pellegrini, and M. Müller, "Processing and Classification of Protein Mass Spectra," *Mass Spectrometry Rev.,* vol. 25, pp. 409-449, 2006.

[8] H. Shin and M.K. Markey, "A Machine Learning Perspective on the Development of Clinical Decision Support Systems Utilizing Mass Spectra of Blood Samples," *J. Biomedical Informatics,* vol. 39, no. 2, pp. 227-248, 2006.

[9] H.W. Ressom, R.S. Varghese, L. Goldman, C.A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny, and R. Goldman, "Analysis of MALDI-TOF Mass Spectrometry Data for Detection of Glycan Biomarkers," *Proc. Pacific Symp. Biocomputing,* pp. 216-227, 2008.

[10] E. Marchiori, C.R. Jimenez, M. West-Nielsen, and N.H. Heegaard, "Robust SVM-Based Biomarker Selection with Noisy Mass Spectrometric Proteomic Data," *Applications of Evolutionary Computing,* pp. 79-90, Springer, 2006.

[11] P. Bougioukos, D. Glotsos, D. Cavouras, A. Daskalakis, I. Kalatzis, S. Kostopoulos, G. Nikiforidis, and A. Bezerianos, "An Intensity-Region Driven Multi-Classifier Scheme for Improving the Classification Accuracy of Proteomic MS-Spectra," *Computer Methods and Programs in Biomedicine,* vol. 99, no. 2, pp. 147-153, 2010.

[12] K.R. Coombes, K.A. Baggerly, and J.S. Morris, "Pre-Processing Mass Spectrometry Data," *Fundamentals of Data Mining in Genomics and Proteomics,* W. Dubitzky, M. Granzow, and D. Berrar, eds., pp. 79-102, Springer, 2007.

[13] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam, "Consensus Clustering and Functional Interpretation of Gene-Expression Data," *Genome Biology,* vol. 5, no. 11, pp. R94.1-R94.16, 2004.

[14] D. Valkenborg, S. Van Sanden, D. Lin, A. Kasim, Q. Zhu, P. Haldermans, I. Jansen, Z. Shkedy, and T. Burzykowski, "A Cross-Validation Study to Select a Classification Procedure for Clinical Diagnosis Based on Proteomic Mass Spectrometry," *Statistical Applications in Genetics and Molecular Biology,* vol. 7, no. 2, p. 12, 2008.

[15] J. Dutkowski and A. Gambin, "On Consensus Biomarker Selection," *BMC Bioinformatics,* vol. 8 (Suppl 5), no. S5, 2007.

[16] R. Armañanzas, B. Calvo, I. Inza, M. López-Hoyos, V. Martínez-Taboada, E. Ucar, I. Bernales, A. Fullaondo, P. Larrañaga, and A.M. Zubiaga, "Microarray Analysis of Autoimmune Diseases by Machine Learning Procedures," *IEEE Trans. Information Technology in Biomedicine,* vol. 13, no. 3, pp. 341-350, May 2009.

[17] A. Barla, G. Jurman, S. Riccadonna, S. Merler, M. Chierici, and C. Furlanello, "Machine Learning Methods for Predictive Proteomics," *Briefings in Bioinformatics,* vol. 9, no. 2, pp. 119-128, 2008.

[18] A.C. Sauve and T.P. Speed, "Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data," *Proc. Workshop Genomic Signal Processing and Statistics,* 2004.

[19] J.B. Breen, F.G. Hopwood, K.L. Williams, and M.R. Wilkins, "Automatic Poisson Peak Harvesting for High Throughput Protein Identification," *Electrophoresis,* vol. 21, pp. 2243-2251, 2000.

[20] W. Meuleman, J.Y. Engwegen, M.W. Gast, J.H. Beijnen, M.J. Reinders, and L.F. Wessels, "Comparison of Normalisation Methods for Surface-Enhanced Laser Desorption and Ionisation (SELDI) Time-of-Flight (TOF) Mass Spectrometry Data," *BMC Bioinformatics,* vol. 9, article no. 88, 2008.

[21] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.-C. Hung, and H.M. Kuerer, "Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform," *Proteomics,* vol. 5, no. 16, pp. 4107-4117, 2005.

[22] J. Prados, A. Kalousis, and M. Hilario, "On Preprocessing of SELDI-MS Data and Its Evaluation," *Proc. 19th IEEE Symp. Computer-Based Medical Systems,* pp. 953-958, 2006.

[23] Y. Chen, V. Kamat, E.R. Dougherty, M.L. Bittner, P.S. Meltzer, and J.M. Trent, "Ratio Statistics of Gene Expression Levels and Applications to Microarray Data Analysis," *Bioinformatics,* vol. 18, no. 9, pp. 1207-1215, 2002.

[24] H. Shin, B. Sheub, M. Joseph, and M.K. Markey, "Guilt-by-Association Feature Selection: Identifying Biomarkers from Proteomic Profiles," *J. Biomedical Informatics,* vol. 41, no. 1, pp. 124-136, 2008.

[25] D.J. Slotta, L.S. Heath, N. Ramakrishnan, R. Helm, and M. Potts, "Clustering Mass Spectrometry Data Using Order Statistics," *Proteomics,* vol. 95, pp. 1687-1691, 2003.

[26] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.-T. Le, "Sample Classification from Protein Mass Spectrometry, by 'Peak Probability Contrasts,'" *Bioinformatics,* vol. 20, no. 17, pp. 3034-3044, 2004.

[27] H.W. Ressom, R.S. Varghese, S.K. Drake, G.L. Hortin, M. Abdel-Hamid, C.A. Loffredo, and R. Goldman, "Peak Selection from MALDI-TOF Mass Spectra Using Ant Colony Optimization," *Bioinformatics,* vol. 23, no. 5, pp. 619-626, 2007.

[28] H.W. Ressom, R.S. Varghese, M. Abdel-Hamid, S.A. Eissa, D. Saha, L. Goldman, E.F. Petricoin, T.P. Conrads, T.D. Veenstra, C.A. Loffredo, and R. Goldman, "Analysis of Mass Spectral Serum Profiles for Biomarker Selection," *Bioinformatics,* vol. 21, no. 21, pp. 4039-4045, 2005.

[29] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher, "OpenMS—An Open-Source Software Framework for Mass Spectrometry," *BMC Bioinformatics,* vol. 9, article no. 163, 2008.

[30] P.A. Bosman and D. Thierens, "Linkage Information Processing in Distribution Estimation Algorithms," *Proc. Genetic and Evolutionary Computation Conf. (GECCO '99),* pp. 60-67, 1999.

[31] H. Mühlenbein and G. Paaß, "From Recombination of Genes to the Estimation of Distributions. Binary Parameters." *Proc. Fourth Int'l Conf. Parallel Problem Solving from Nature (PPSN),* pp. 178-187, 1996.

[32] M. Pelikan, *Hierarchical Bayesian Optimization Algorithm. Toward a New Generation of Evolutionary Algorithms.* Springer, 2005.

[33] J.H. Holland, *Adaptation in Natural and Artificial Systems.* Univ. of Michigan Press, 1975.

[34] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning.* Kluwer Academic Publishers, 1989.

[35] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, 1992.

[36] P. Larrañaga, "A Review on Estimation of Distribution Algorithms," *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation,* pp. 55-98, Kluwer Academic Publishers, 2002.

[37] Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[38] Y. Saeys, D. Degroeve, D. Aeyels, P. Rouzé, and Y. Van de Peer, "Feature Selection for Splice Site Prediction: A New Method Using EDA-Based Feature Ranking," *BMC Bioinformatics,* vol. 5, article no. 64, 2004.

[39] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," *Machine Learning and Knowledge Discovery in Databases,* pp. 313-325, Springer, 2008.

[40] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms," *Proc. Fifth IEEE Int'l Conf. Data Mining,* pp. 218-225, 2005.

[41] L.I. Kuncheva, "A Stability Index for Feature Selection," *Proc. 25th IASTED Int'l Multi-Conf. Artificial Intelligence and Applications,* pp. 390-395, 2007.

[42] K.A. Baggerly, J.S. Morris, and K.R. Coombes, "Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Data Sets from Different Experiments," *Bioinformatics,* vol. 20, pp. 777-785, 2004.

[43] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis," *Bioinformatics,* vol. 21, no. 5, pp. 631-643, 2005.
[44] J. Reunanen, "Overfitting in Making Comparisons between Variable Selection Methods," *J. Machine Learning Research,* vol. 3, pp. 1371-1382, 2003.
[45] *Computational Methods of Feature Selection,* H. Liu and H. Motoda, eds. Chapman and Hall/CRC Press, 2008.
[46] B. Efron, "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *J. Am. Statistical Assoc.,* vol. 78, no. 382, pp. 316-331, 1983.
[47] J. Handi, D.B. Kell, and J. Knowles, "Multiobjective Optimization in Bioinformatics and Computational Biology," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 2 pp. 279-292, Apr.-June 2007.
[48] V. Pareto, *Cours D' économie Politique,* Univ. of Lausanne, vols. I and II. 1896.
[49] E.F. Petricoin, V. Rajapaske, E.H. Herman, A.M. Arekani, S. Ross, D. Johann, A. Knapton, J. Zhang, B.A. Hitt, T.P. Conrads, T.D. Veenstra, L.A. Liotta, and F.D. Sistare, "Toxicoproteomics: Serum Proteomic Pattern Diagnostics for Early Detection of Drug Induced Cardiac Toxicities and Cardioprotection," *Toxicologic Pathology,* vol. 32, no. 1, pp. 122-130, 2004.
[50] H.W. Ressom, R.S. Varghese, E. Orvisky, S.K. Drake, G.L. Hortin, M. Abdel-Hamid, C.A. Loffredo, and R. Goldman, "Ant Colony Optimization for Biomarker Identification from MALDI-TOF Mass Spectra," *Proc. 28th Int'l Conf. IEEE Eng. in Medicine and Biology Soc.,* pp. 4560-4563, 2006.
[51] G.H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *Proc. 11th Conf. Uncertainty in Artificial Intelligence,* pp. 338-345, 1995.
[52] R. Santana, C. Bielza, P. Larrañaga, J.A. Lozano, C. Echegoyen, A. Mendiburu, R. Armañanzas, and S.K. Shakya, "MATEDA: A Matlab Package for the Implementation and Analysis of Estimation of Distribution Algorithms," *J. Statistical Software,* vol. 35, no. 7, pp. 1-30, 2010.
[53] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* Wiley Interscience, 2001.
[54] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine Learning in Bioinformatics," *Briefings in Bioinformatics,* vol. 7, no. 1, pp. 86-112, 2006.
[55] K.A. Baggerly, J.S. Morris, S.R. Edmonson, and K.R. Coombes, "Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer," *J. Nat'l Cancer Inst.,* vol. 97, no. 4, pp. 307-309, 2005.

Rubén Armañanzas received the MS and PhD degrees in computer science from the University of the Basque Country, Spain, in 2004 and 2009, respectively. He is currently a postdoctoral researcher at the Technical University of Madrid, Spain, working for the international Cajal Blue Brain project. His research interests include computational intelligence approaches to feature selection, classification and knowledge discovery in computational biology, biomedicine, and neuroinformatics.

Yvan Saeys received the PhD degree from Ghent University, Belgium, in 2004. Currently, he is an FWO postdoctoral researcher in the Bioinformatics and Systems Biology Group, where he leads the Data Mining Group. His research interests include data mining and machine learning approaches, and their applications in bioinformatics and systems biology. He has published more than 45 papers in international journals and conferences.

Iñaki Inza received the PhD degree in computer science in 2002. Currently, he is a senior researcher enrolled in the Intelligent Systems Group, University of the Basque Country, San Sebastian, north of Spain. His main methodological research interests include classification and evolutionary computation by means of Bayesian networks, and feature selection. He has taken part in application projects in bioinformatics, web mining, and oceanographic domains.

Miguel García-Torres received the BS degree in physics and the PhD degree in computer science from the Universidad de La Laguna, Tenerife, Spain, in 2001 and 2007, respectively. Currently, he is a lecturer in the Escuela Politécnica Superior of the Universidad Pablo de Olavide, Seville. His research interests include data mining, machine learning, data reduction, bioinformatics, and astrostatistics.

Concha Bielza received the MS degree in mathematics from the Complutense University of Madrid, Spain, in 1989, and the PhD degree in computer science from the Universidad Politécnica de Madrid in 1996. She is currently an associate professor of statistics and operations research with the Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid. Her research interests are primarily in the areas of probabilistic graphical models, decision analysis, metaheuristics for optimization, data mining, classification models, and real applications. She has published more than 25 papers in widely renowned journals.

Yves van de Peer received the PhD degree from the University of Antwerp, Belgium, in 1995. Currently, he is working as a professor in bioinformatics and genome biology at Ghent University, Belgium, where he leads the Bioinformatics and Systems Biology Group. His research interests include using bioinformatics approaches to study the evolution of organisms, genes, and genomes. He has published more than 240 papers in widely renowned journals, and is an editorial board member of seven international journals.

Pedro Larrañaga received the diploma degree in mathematics from the University of Valladolid, Spain, in 1981, and the PhD degree in computer science from the University of the Basque Country, Spain, in 1995, where he became an associate professor in 1998 and full professor in 2004. In 2007, he joined the Technical University of Madrid as full professor in the Department of Artificial Intelligence, where he leads the Computational Intelligence Group. His research interests include probabilistic graphical models and heuristic optimization. He has coauthored two edited books on estimation of distribution algorithms, as well as more than 300 scientific papers in different areas. He has participated in more than 70 research projects at national, European, and international levels.