Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 4380–4394
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A new genetic algorithm in proteomics: Feature selection for SELDI-TOF data

Christelle Reynès[a,*], Robert Sabatier[a], Nicolas Molinari[b], Sylvain Lehmann[c]

[a] *Laboratoire Physique Industrielle et Traitement de l'Information, EA 2415, Faculté de Pharmacie, 15 av. Charles Flahault, BP 14491, 34093 Montpellier Cedex 5, France*
[b] *Laboratoire de Biostatistique, EA 2415, Institut Universitaire de Recherche Clinique, 641 av. G. Giraud, 34093 Montpellier, France*
[c] *Institut de Génétique Humaine du CNRS, UPR 1142, 141, rue de la Cardonille, 34396 Montpellier Cedex 5, France*

Received 14 May 2007; received in revised form 21 February 2008; accepted 21 February 2008
Available online 29 February 2008

Abstract

Mass spectrometry from clinical specimens is used in order to identify biomarkers in a diagnosis. Thus, a reliable method for both feature selection and classification is required. A novel method is proposed to find biomarkers in SELDI-TOF in order to perform robust classification.The feature selection is based on a new genetic algorithm. Concerning the classification, a method which takes into account the great variability on intensity by using decision stumps has been developed. Moreover, as the samples are often small, it is more appropriate to use the decision stumps simultaneously than building a complete tree. The thresholds of the decision stumps are determined in the same genetic algorithm. Finally, the method was generalized to more than two groups based on pairwise coupling. The obtained algorithm was applied on two data sets: a publicly available one containing two groups allowing a comparison with other methods from the literature and a new one containing three groups.

1. Introduction

In the field of proteomics, mass spectrometry and especially SELDI-TOF (Surface Enhanced Laser Desorption Ionisation-Time Of Flight), is considered as a promising technique and is increasingly used. Indeed, it is noticeable among other spectrometric techniques because of the ability of retaining proteins according to their chemical properties thanks to different chemical surfaces. Huge amount of data are produced quite rapidly and they are expected to contain valuable information.

However, this technique suffers from a relative lack of accuracy, especially in intensity, and consequently from a lack of reproducibility. Actually, it is not a really quantitative technique as most mass spectrometry approaches. That is why it is necessary to use the intensity values very carefully and to conceive statistical methodologies which are likely to extract interesting features using as little information as possible. Now, the simplest way to use quantitative data is to binarize them. To do that, it is possible to define a threshold in intensity and to turn the raw intensities into above/below the threshold.

* Corresponding author. Tel.: +33 4 67 54 80 68; fax: +33 4 67 66 81 91.
E-mail address: creynes@univ-montp1.fr (C. Reynès).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.025

Yet, one of the goals of SELDI-TOF spectrometry is the discovery of biomarkers, that is to say peaks in the spectra which are interesting to perform discrimination between two or more groups of samples. In this context, biomarkers would be peaks allowing discrimination by applying thresholds.

The method developed in this paper is based on Genetic Algorithms (denoted GA) and very simple classification trees (one-level decision trees). The algorithm aims at choosing a set of interesting peaks in the mass spectra and optimal split points in the trees which correspond to threshold intensity. Such a set of peaks and their corresponding thresholds will be denoted a *committee*.

In the first part, the preprocessing of the data will be described. In Section 3, the discrimination step will be introduced, firstly for only two groups, then Section 3.2 will focus on a generalization to more than two groups. The GA used to optimize the parameters will be described in Section 4. At last, the method will be applied to two sets of data containing two and three groups (Section 5).

2. The preprocessing

In mass spectrometry, preprocessing is of prime importance (Chen et al., 2007) to be able to compare spectra. The preprocessing applied here is divided into the following three steps:

- baseline subtraction and normalization: the base level in mass spectra is not homogeneous along the mass-to-charge ratio (m/z) axis. Indeed, it is a kind of exponential curve for the small masses, then it becomes a linear function of the mass (Wagner et al., 2003). In our method, the baseline correction and the normalization (based on total ion current) provided by Ciphergen ProteinChip® software (Ciphergen Biosystems, Freemont, CA, USA) have been used. However, they are easy to implement independently if necessary.
- peak extraction: many methods exist to perform peak extraction (see for example, Coombes et al. (2003), Yasui et al. (2003) and Tibshirani et al. (2004)). Here, this step was performed using an new algorithm which looks for local maxima as potential peaks. The boundaries of peaks are set where the *derivative curve* (based on the first differences) of the spectra becomes null on each side of the potential peak (changes in slope). Finally, peaks considered as noise are identified thanks to their small valley-depth and discarded. By this method, there are few influent parameters to be fixed as the algorithm is adapted to each peak shape. The extraction method provided by Ciphergen seems to be very efficient too but an independent method allows to control all the parameters and to perform comparisons between different technologies (Reynès et al., 2007).
- peak matching: all the peaks detected in all the spectra are gathered on a *super vector* of peak locations and average-link hierarchical-clustering (Duda et al., 2001; Prados et al., 2004) is applied to define which peaks are likely to arise from the same peptide and which are not. The threshold used to cut the tree depends on the location along the m/z axis as it is based on the mass accuracy (0.1%). A cluster of peaks will be referred to as a *class* of peaks so that there is no confusion with the *groups* of spectra which are the discrimination objective. This step of alignment is quite succinct in Ciphergen software. For example, it takes the average location of peaks to be matched and takes the intensity at this location as the peak intensity whereas it is not always the local maximum.

Let us notice that in the last two steps, the mass accuracy is taken into account. It is generally considered to be 0.1% (Yasui et al., 2003). Indeed, we use it to define the width of the window used to calculate the first differences in the peak finding algorithm and the cut point in the hierarchical clustering. This allows to take into consideration the heterogeneity of SELDI data along the m/z axis. Actually, the data are denser at the lower end of the signal and peak shape evolves: from narrow and very high at the lower end, they become broader and lower as we progress along the axis (Tibshirani et al., 2004). Then, taking it into account significantly increases the relevance of the preprocessing.

¿From now on, the data will be collected into a $(n \times p)$ matrix, X. The element of the i-th row and j-th column, denoted $x(i, j)$ is the intensity of the peak in the j-th class found in the i-th spectrum. If a peak could not be found in a spectrum, its intensity is set to zero or to the minimum intensity in the corresponding zone.

3. The discrimination step

Ciphergen software and most of published papers concerning biomarker discovery using SELDI-TOF are based on t-tests or F-tests to find discriminant peaks. However, a single peak is unlikely to be able to make a perfect discrimination between groups. Using multivariate classifiers seems much more efficient. This approach has already

been explored by several authors. Some of them are mentioned in the next sections and especially in paragraph 5.1.1. However, several drawbacks can be noticed in existing methods: some of them cannot be easily generalized to more than two groups, other ones are based on raw intensity which seems to be irrelevant as SELDI-TOF is known to have a great variation in intensity measure. That is why we developed a method which uses several peaks simultaneously, using binary data for intensity (thanks to a threshold) and which has been generalized to more than two groups. This method is introduced in the next sections.

3.1. Case of $k = 2$ groups

The discrimination step will be introduced in this section for $k = 2$ groups. It will be generalized to $k > 2$ groups in the next paragraph. Attention is drawn on the way of performing the discrimination and obviously to compute the rate of good classification, once a *committee* (peaks and thresholds) has been chosen. The way of choosing the committee will be considered in Section 4.

In this part of the analysis, it is necessary to take into account the intensities of the peaks but it is well known that the variability in intensities, in SELDI-TOF technology, can reach 50% (Yasui et al., 2003). Hence, it is more suitable to use minimum information about those data. The simplest way to use quantitative variables is to binarize the data and to use thresholds like in classification trees (Breiman et al., 1984). However, in classification trees, the sub-populations which are used to take the successive decisions (at the different nodes) are smaller and smaller and so, less and less representative of the whole population. This phenomenon is likely to lead to overfit the training data set and the application of such a model on new samples would probably not be very accurate. It will particularly arise in small samples.

Then, using an one-level decision tree, which is called a *decision stump*, could be an interesting solution. Such a tree would split the whole population into two subsamples thanks to one threshold concerning one variable and applied on the whole training population. However, such discrimination is only efficient if one of the features is able to perform by itself a satisfying discrimination between groups and this is generally not the case. Therefore, several decision stumps will be used simultaneously. This approach was chosen by Qu et al. (2002) but it was introduced in a boosted way where the different decision stumps were chosen successively, each new stump aiming at correcting the first members mistakes. In our case, the decision stumps are chosen simultaneously. It can appear weaker but as GA will be used to choose them and as parsimony will be taken into account, sets of classifiers having a few but complementary peaks will naturally be favoured. It leads to a really multivariate method with each decision taking into account the whole population. The set of decision stumps makes up the committee. Each decision stump corresponds to the *vote* of one peak. A committee will contain at most N_{mp} peaks. Thus, we consider simultaneously between 1 and N_{mp} one-level tree(s) to perform the discrimination.

To determine the decision rules of the committee, one part of the samples, L, containing n_g spectra from the g-th group ($g = \{1, 2\}$), is used. L_g will denote the set of the spectra from the g-th group which are included in L. Let us consider a committee, C, containing N_C peaks and their corresponding thresholds. For each peak m_i^C ($i = 1, \ldots, N_C$) of the committee, one has to decide to which group g, a spectrum, t ($t \in L$), will be affected depending on whether its peak, m_i^C, has an intensity $x(t, m_i^C)$ above or below the threshold, $z_{m_i^C}$. The decision taken when $x(t, m_i^C) > z_{m_i^C}$ will be denoted $V^+_{m_i^C}$ ($V^+_{m_i^C} \in \{1, 2\}$) and $V^-_{m_i^C}$ when $x(t, m_i^C) \leq z_{m_i^C}$ ($V^-_{m_i^C} \in \{1, 2\}$). These votes are computed as follows:

$$V^+_{m_i^C} = \arg \max_{g=\{1,2\}} \left(\frac{1}{n_g} \sum_{t \in L_g} \mathbf{1}\{x(t, m_i^C) > z_{m_i^C}\} \right), \tag{1}$$

$$V^-_{m_i^C} = \arg \max_{g=\{1,2\}} \left(\frac{1}{n_g} \sum_{t \in L_g} \mathbf{1}\{x(t, m_i^C) \leq z_{m_i^C}\} \right), \tag{2}$$

where, $\mathbf{1}\{B\}$ is the indicator of the set B. Then, one has to compute these two votes for the N_C peaks. Once, the decision rules have been defined, the rate of good classification is computed by classifying the remaining samples, T. In that goal, for each spectrum $t \in T$, its allocation to the groups, $A(t)$ ($A(t) \in \{1, 2\}$), is defined as follows:

$$A(t) = \arg \max_{g=\{1,2\}} \left(\sum_{i=1}^{N_C} \mathbf{1}\{v_{m_i^C}(t) = g\} \right), \tag{3}$$

where,

$$v_{m_i^C}(t) = \mathbf{1}\{x(t, m_i^C) > z_{m_i^C}\} V^+_{m_i^C} + \mathbf{1}\{x(t, m_i^C) \leq z_{m_i^C}\} V^-_{m_i^C}. \tag{4}$$

If two groups tie, the spectrum is considered as misclassified. Because of this constraint, odd numbers of decision stumps will often be favoured.

3.2. Generalization to $k > 2$ groups

When there are more than two groups, it is still possible to use a tree-based approach but, as previously noticed, it works on smaller and smaller samples which is not suitable for small populations. In other respects, it would be possible to use other approaches such as discriminant analysis which takes into account the real intensity value but as this value is not reliable, we want to keep our threshold approach. For $k = 3$ groups, it is possible to set two thresholds for each peak if necessary. Then, one peak is able to discriminate the three groups by itself. But for $k > 3$, it becomes very difficult. That is why, the chosen approach was pairwise coupling (Wu et al., 2004). It consists in performing all the pairwise comparisons, to deduce class probabilities from each comparison and to estimate global class probabilities by combining these pairwise probabilities.

3.2.1. Computation of the pairwise class probabilities

To perform pairwise classification, a committee has to be chosen for each comparison. There are $k(k-1)/2$ pairwise comparisons for k groups, so $k(k-1)/2$ committees have to be chosen. The goal of this step is to compute, for each spectrum $t \in T$:

$$r_{ij}(t) = P(t \in i | t \in \{i, j\}, X, C^{ij}), \quad i = 1, 2, \ldots, k, i < j \leq k, \tag{5}$$

where, X is the matrix containing the intensities of the peaks (see Section 2), C^{ij} denotes the committee chosen to compare groups i and j, it is made up of $N_{C^{ij}}$ peaks and thresholds. One obtains a $(k \times k)$ matrix $R(t)$, where the element from the i-th row and j-th column, $r_{ij}(t)$ is estimated by $\hat{r}_{ij}(t)$ as follows:

$$\hat{r}_{ij}(t) = \frac{1}{N_{C^{ij}}} \sum_{n=1}^{N_{C^{ij}}} \mathbf{1}\{v^{ij}_{m_n}(t) = i\} \quad \text{if } i < j, \tag{6}$$

$$= \frac{1}{N_{C^{ij}}} \sum_{n=1}^{N_{C^{ij}}} \mathbf{1}\{v^{ij}_{m_n}(t) = j\} \quad \text{if } i > j, \tag{7}$$

$$= 0 \quad \text{if } i = j, \tag{8}$$

where, $v^{ij}_{m_n}(t)$ is the vote of the n-th peak of the committee chosen to compare groups i and j (the subscript, C^{ij} for m has been omitted to lighten the writing).

3.2.2. Computation of the global class probabilities

Once the pairwise probabilities are computed, there are several ways of estimating the global class probabilities $p = (p_1, \ldots, p_k)$ (Wu et al., 2004). The chosen approach was developed by Wu et al. (2004). They performed several simulations and compared the performance of different probability estimates. Their method seems to be really robust, it gives good results for various structures of probabilities. It consists in computing the vector of probabilities p so that it minimizes

$$\min_p \sum_{i=1}^{k} \sum_{j:j\neq i} (r_{ji} p_i - r_{ij} p_j)^2, \quad \text{subject to } \sum_{i=1}^{k} p_i = 1, \, p_i \geq 0, \forall i. \tag{9}$$

All the details are provided in Wu et al. (2004).

Table 1
Example of encoding of one solution containing three peaks: the peak labels are followed by their corresponding thresholds, at the end, as many NAs as necessary have been added to obtain length $2 \times N_{mp}$

110	11.47	188	6.21	180	1.53	NA	...

4. Choice of the committees by Genetic Algorithm

Genetic Algorithms (GA) are inspired by nature and especially by natural selection (Goldberg, 1989; Chatterjee et al., 1996), they are very useful in complex optimization issues. Here, the GA will be used to find up optimal committees. Thus, it can be considered as a feature selection issue, a field where GAs are widely used (Kapetanios, 2007; Ambrogi et al., 2007). The algorithm begins with a population constituted of several individuals which correspond to potential solutions in the optimization problem. Thus, in our context, the individuals will be committees. Then, this population evolves according to three operators described in the next paragraphs: crossover, mutation and selection. Selection (see Section 4.4) is a crucial step as it allows to keep the best individuals with regard to the issue. It is based on a fitness value (see Section 4.3) which quantifies the quality of each solution. Other steps (see Sections 4.5 and 4.6) are independent from the problem. Then, any GA can be described as follows:

Main steps of a GA:
1. construction of the first generation
2. selection
while stopping criteria not met **do**
   3. crossover
   4. mutation
   5. selection
end

In a first time, the GA used will be precisely described for two groups and the necessary adaptations for the case of more groups will be studied in Section 4.8.

4.1. Encoding solutions

The first step to perform GA is the encoding of solutions. Indeed, in GA, each potential solution has to be fully described by a numerical vector. The historical encoding is bit strings but real encoding is more and more used and has certain advantages (Salomon, 1996). In our case, each solution has to contain the labels of peaks and the corresponding thresholds values. In order to make the operators easier to apply, all the solutions will be of the same length. As a maximum number of peaks, N_{mp}, is set, a solution vector will always have length $2 \times N_{mp}$. For the solutions containing less than N_{mp} peaks, the vector will be filled with NA values. An example is given in Table 1.

4.2. Initial population

Like in any step by step optimization problem, the knowledge of good starting parameters benefits the convergence speed of the algorithm. But such knowledge is rarely available. This leads to generate a random initial population which covers a very large part of the solution space. Therefore, a very heterogeneous initial population is suitable to make the exploration of the solution space easier.

In our case, for each of the S_{pop} individuals of the initial population, the parameters to be set are the following ones: the number of peaks a, the set of peaks used and their corresponding thresholds. Moreover, a peak having large amplitude among the different spectra is likely to be more interesting for discrimination. That is why the probability to select one peak in the initial population is proportional to the range of its intensity values among all the spectra. Then, for each individual, a random number of peaks is computed and the peaks are chosen according to their amplitude. This selection also leads to disadvantage very small peaks which might be remaining noise.

Thresholds could be uniformly chosen between zero and the maximum intensity for each class of peaks, but it is not necessary to pass through this entire interval. Indeed, considering two different values between two successive

observed intensities will lead to the same discrimination result. That is why, the list of potential thresholds will only contain one value between two successively observed intensities. Therefore, if there are n spectra, a maximum of $n-1$ threshold values have to be proposed per peak. It considerably decreases the number of possibilities to be explored and hence, the number of iterations needed to converge.

Moreover, if only one peak is considered, an optimal threshold can be found as follows. Let us consider two groups of spectra of size n_1 and n_2. Then, the proportion $\pi_{mg}(j)$ of spectra in group g ($g = \{1, 2\}$) whose peak m has a larger intensity than the j-th threshold (z_m^j) (the superscript j is only added in this section to find the initial optimum) for this class is defined as

$$\pi_{mg}(j) = \frac{1}{n_g} \sum_{i \in g} \mathbf{1}\{x(i, m) > z_m^j\}. \tag{10}$$

The optimal threshold, $\hat{z}_m^{\hat{j}}$ is chosen among the c_m possible values for the m-th peak so that:

$$\hat{j} = \arg \max_{j=1,\ldots,c_m} |\pi_{m1}(j) - \pi_{m2}(j)|. \tag{11}$$

This optimal threshold is automatically applied when a new peak is introduced. This aims at avoiding a systematic decrease of the fitness value when a new peak is added to a possible solution. Obviously, the optimal threshold in a particular committee is not necessarily the same as the one defined peak by peak but along the generations this threshold will evolve.

4.3. Fitness value

As discrimination of the different spectra is the objective, the fitness value will have to take into account the good classification rate, τ, achieved by each potential solution. Moreover, generally, a committee made up of numerous peaks is more likely to perform a good discrimination than a small one. Nevertheless, using too many decision stumps may lead to overfit the training set and lose generality. So, a parsimony term concerning the number of peaks, a, in the committee is added. As we have no *a priori* knowledge of the optimal number of decision stumps needed, the parsimony term, $\rho(a)$, will be defined as a linear function of the committee size (small sizes are favoured):

$$\rho(a) = \alpha a + \beta. \tag{12}$$

The real numbers α and β are computed so that $\rho(1) = 1$ and $\rho(N_{mp}) = 0$. A scalar, c, is used to make the balance between the model accuracy (good classification rate) and the parsimony leading to the following expression of the fitness value for each solution:

$$\text{fitness} = \tau + c \times \rho(a). \tag{13}$$

In practice, τ (the good classification rate) is the first goal of the optimization, so it will be favoured. As τ and $\rho(a)$ belong to [0, 1], c has to be a real number in]0, 1[. In fact, small values of c quickly lead to a maximum number of peaks in committees (N_{mp}), on the opposite, for high values of c, the number of peaks is jammed to one. Only a narrow range of values for c allows intermediate number of peaks. In this range, c influences the convergence rate more than the final result. We used $c = 0.7$.

4.4. Selection step

This step, based on the fitness values, is defined as in Reeves and Rowe (2003). The individuals are ranked according to their fitness value, the best one having the highest rank. Then, the probability to keep one solution in the next generation is

$$P(\text{selecting } k\text{th ranked solution}) = \delta + \mu \times k, \tag{14}$$

where δ and μ are chosen so that

$$\sum_{k=1}^{S_{\text{pop}}} \delta + \mu \times k = 1, \tag{15}$$

and

$$P(\text{selecting best solution}) = 2 \times P(\text{selecting median ranked solution}). \tag{16}$$

As theoretically proved in Bhandari et al. (1996), two conditions are necessary and sufficient for GA to converge as the number of iterations goes to infinity:

- The best solution in the present population has a fitness value no less than the fitness values of the optimal strings from the previous populations.
- Each solution has a positive probability of going to an optimal string within any given iteration.

Thus, the best solution of each generation is compulsory included in the next population, which is called *elitism*. It takes the place of any randomly chosen other solution so that the population always keeps the same size, S_{pop}.

4.5. Crossover step

The objective of this step is to make combinations of the previously retained solutions in order to gather interesting features (peaks and thresholds) of several solutions in new individuals. It is important to notice that this step in independent from the optimization, that is to say, a crossover is likely to produce better and worse solutions equally. Only the selection step is likely to eliminate bad solutions.

Crossover is defined as in Jeffries (2004). A proportion π_c of the population is randomly chosen to undergo crossover. As explained in the previous paragraph, the convergence is not dependent on the crossover. Hence, π_c only influence the convergence rate. A few tests led us to choose $\pi_c = 0.7$. The selected solutions are grouped in pairs. We consider a pair containing one solution with nbp_1 peaks and one with nbp_2 peaks. The number of peaks, N_{cp}, to be crossed is uniformly chosen between 1 and $\max(nbp_1, nbp_2)$. Then, the beginning of the crossover has to be randomly chosen in $\{1, 2, \ldots, \min[(N_{mp} - Ncp + 1), \max(nbp_1, nbp_2)]\}$. Finally, the corresponding peaks are crossed giving two new solutions. The same process is repeatedly applied to each pair of solutions.

4.6. Mutation step

This step brings the necessary hazard to efficiently explore the solution space. It assures that any point of this space can be reached. Moreover, if a local optimum is obtained, mutations will avoid a too quick convergence to this local optimum. Thus, efficient convergence of GA is highly dependent upon this step. The mutation rate (proportion of the solutions which will undergo mutation), π_m is defined for each generation. It is set to a maximum value at the beginning then decreases to allow convergence and finally, it increases again to avoid local optima. In practice, at the beginning of the algorithm, it is usual to have in probability one mutation per individual (in this stage, the proportion of the different mutations is more important), hence it is set to 0.9. At the minimum of mutation rate (75% of the total number of generations) it has to be very low, we chose $\pi_m = 0.1$. Between these two values the decrease is linear (generally the slope is very small as the number of generations is high) and the same slope is kept at the end of the algorithm to increase π_m.

In practice, mutation consists in changing values in the vectors corresponding to the solutions that have been chosen to undergo mutation. In our context, mutations can be divided into three different types:

- peak elimination: a peak is randomly chosen and removed from the solution (*i.e.* the committee),
- peak addition: a new peak is chosen, added and the optimal threshold is associated,
- threshold relocation: one of the thresholds is randomly removed and replaced by another one.

Firstly, according to the mutation rate, we choose the individuals which will undergo mutation. Then, the mutation to apply has to be chosen. In this context, the probability, p_e, to eliminate a peak, p_a, to add a peak, and p_r, to relocate a threshold for a solution containing k peaks are defined as follows (Green, 1995) and computed:

$$p_e = \lambda \times \min\left(1, \frac{\text{pois}(k-1)}{\text{pois}(k)}\right), \tag{17}$$

where $\text{pois}(k)$ is the probability for the number of peaks to equal k thanks to a Poisson distribution of parameter $N_{mp}/2$ (DiMatteo et al., 2001). The parameter of the Poisson distribution is chosen so that parsimony is favoured.

The λ coefficient is used so that $p_r \geq 0$ (see Eq. (19)). As our maximum number of peaks, N_{mp}, was ten, $\lambda = 0.4$ performed well.

In the same way,

$$p_a = \lambda \times \min\left(1, \frac{\text{pois}(k+1)}{\text{pois}(k)}\right), \tag{18}$$

and

$$p_r = 1 - (p_e + p_a). \tag{19}$$

Thanks to these probabilities, a kind of mutation is chosen and applied.

4.7. End of the GA

Concerning the stopping criterion, several approaches can be chosen. Firstly, it is possible to set a maximum number of generations, N_{gene}, or (equivalently) a maximum computation time. On the other hand, another class of stopping rule is to study the evolution of the population. It can consist in a nonsignificative evolution in the average fitness in the population since a given number of generations. It can also be based on the population diversity. This kind of approaches is not easy to implement as it is compulsory to determine thresholds which are dependent upon the fitness function shape. Then, it is necessary to study the behaviour of the algorithm for each new application. In this context, it is easier to study the number of generations, it is the choice we made in the following applications.

When the final population is obtained, the committee which will be chosen is the one which occurs the most frequently in the S_{pop} potential solutions.

4.8. Adaptation to more than two groups

Globally, the unfolding will be the same for $k > 2$ groups but there will be as many populations as possible pairwise comparisons, that is to say, $k(k-1)/2$, for k groups.

Concerning the initialization, initial optimal thresholds are computed for each comparison and each initial population is built as explained in Section 4.2. It is necessary that the $k(k-1)/2$ populations have the same size, S_{pop}.

Crossover and mutation steps are performed independently for each population as described in Sections 4.5 and 4.6.

It is less obvious with the selection step which has to take into account the solutions of the $k(k-1)/2$ populations simultaneously. Indeed, we are interested in the global class probabilities which are obtained by using all the committees. The split between learning L and test T samples is randomly chosen for each generation. So, the solutions of one generation are evaluated on the same subsample but it is not the same subsample in the next generation. First, the pairwise class probabilities are computed for the q-th solution ($q \in \{1, \ldots, S_{\text{pop}}\}$) of each population. Then, for each $k(k-1)/2$-uplet of pairwise probabilities, the global class probabilities $p_g(t)$ for $g = 1, \ldots, k$ are computed as explained in Section 3.2.2. Thanks to these probabilities, the spectra in T are affected to the groups and the rate of good classification is calculated by comparison to the real groups. Finally, the selection step follows exactly the unfolding described in Section 4.4 with the fitness function defined in Section 4.3 where τ is the rate of good classification previously computed.

5. Results

Two datasets will be studied. The first one is a publicly available one which has been analysed by several different authors. It will allow comparing our method with other ones which have been developed especially for such data. The second data set is a new one, involving three groups of neurological diseases. It will be used to compare our approach with more established methods (Linear Discriminant Analysis, K-Nearest Neighbours, Classification Trees and Random Forests).

In the two applications, the same parameters will be used in the GA. Their values are gathered in Table 2.

Table 2
Values of the parameters introduced in the previous sections and used in the GA for the two applications on datasets

Parameter	N_{mp}	S_{pop}	c	π_c	λ	N_{gene}
Value	10	200	0.7	0.8	0.4	1000

π_m value is not reported because it evolves as described in Section 4.6.

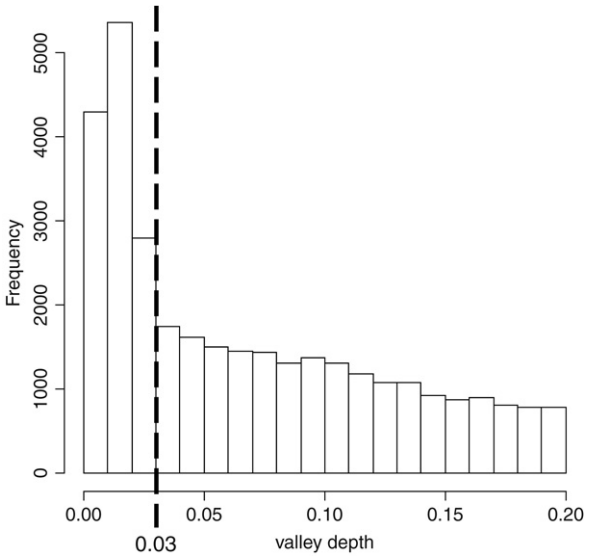

Fig. 1. Distribution of the valley depths among the peaks of the 253 spectra: the dotted line represents the threshold used.

5.1. Ovarian data

5.1.1. The data

The first studied data are a publicly available SELDI-TOF dataset. It consists of ovarian cancer patients and healthy controls. We used the *low resolution* mass spectrometry data from a Ciphergen instrument which is identified on the NCI-FDA website (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp) as the 8-7-02 data. These data are made up of 162 ovarian cancer samples and 91 control samples. Each spectrum contains 15 154 points with mass-to-charge ratio (m/z) ranging approximately from 0 to 20 000 Da. From the 253 samples, 46 control and 81 cancer spectra are randomly chosen to form the training set, the remaining samples (45 control and 81 cancer spectra) will be used to test the solution quality.

Major drawbacks due to nonbiological bias (Baggerly et al., 2003; Sorace and Zhan, 2003) have been identified for these data. However, our purpose here is essentially to be able to compare with other methods. Indeed, this data set has already been used by numerous authors and analysed by several methods: k-nearest neighbours (Zhu et al., 2003), nonparametric statistics and stepwise discriminant analysis (Sorace and Zhan, 2003), genetic algorithm (Jeffries, 2004) based on the one described by Petricoin et al. (2002), SVM (Jong et al., 2004) and logical analysis (Alexe et al., 2004). Moreover, Liu et al. (2002) compared several methods for feature selection and discrimination on it. Hence, a comparison of the results will be possible. We will compare their peak selections with the one obtained by applying the method described in the previous sections. So, even if the results can not be used to draw medical conclusions, it will allow assessing the efficiency of our method compared with other ones in the literature.

5.1.2. Preprocessing

As we described in Section 2, the first step is the identification and superposition of the peaks in the different spectra. After applying the peak finding algorithm, about 350 potential peaks were identified per spectrum. In order to eliminate too small peaks considered as noise, the valley-depths are computed and a threshold is chosen according to their distribution (Fig. 1). It can be easily visually identified as 0.03 (*breaking* in the distribution). Thanks to this threshold, a total number of 12 500 peaks are discarded (50 peaks in each spectrum in average).

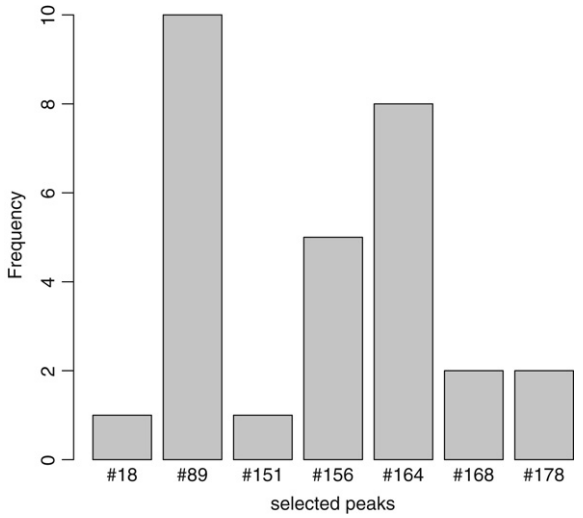

Fig. 2. Number of occurrences of the different peaks found in the committee for the ten training samples.

Then, the next step is the matching of identified peaks. Average-link hierarchical clustering is applied and leads to the identification of 1097 different classes of peaks. As we retained only groups containing peaks from more than 46 spectra (half the number of samples in the smaller group), 670 groups are eliminated. Finally, the matrix X containing peak intensities will have dimension 253×427. In order to simplify the descriptions, peak classes will be referred to by numbers and the real location of interesting peaks will be given in the end.

5.1.3. GA results and discrimination

The discrimination model is based on the training set and will be tested on the test sample once the final committee is chosen. However, one particular distribution between training and test samples is not enough to accurately evaluate the quality and stability of our method. That is why, ten different random splits were performed (defining ten train/test sets) and analysed by genetic algorithm.

The convergence of the GA can be studied by observing the final populations. Indeed, the first thing to look at is the number of peaks in the committee. In this application, three peaks were found in about 90% of the final population for each of the ten runs. In the solutions containing three peaks, there was a perfect convergence of the set of peaks: we always found the same three peaks for the whole final population of each run (but they may be different between runs as explained later). With regard to threshold intensities, the convergence is less clear as the most obtained combination represents in average 40% of the final population.

Then, we can study the results for the discrimination performed by using the final committees. For the training data, the obtained rate of good classification is 100% apart from one split which made one misclassification of a control sample as cancer. For the test set, a global good classification rate of 98% was reached (95% for the control group and 99% for the cancer group). Concerning the peaks used in the genetic algorithm, the committee was always made up of three different peaks. Along the ten splits, seven different peaks appeared (Fig. 2); the peaks #89 was always found in the committees, peak #164 appeared in eight out of ten runs and peak #156 was used in half the committees. The other used peaks were more anecdotal and can be attributed to sample particularities. Moreover, it is interesting to notice that the intensity threshold values used to perform the discrimination don't vary much between the different random splits. For instance, for peak #89 it is between 11.13 and 12.71 whereas the intensity varies from 2 to 27 for this peak. Thus, the discrimination is very efficient with stable results when the training set changes.

Now, let us study the three peaks which have been frequently identified, that is to say #89, #164 and #156. Their respective locations are 245 Da, 434 Da and 649 Da. Their intensities for each group are presented in Fig. 3. We can notice that for the peak #164 most of the spectra from the second group don't have this peak, that is why, most of the intensity values are zero. For the three peaks, the intensity values are really different between groups, this shows the efficiency of the GA in finding discriminative peaks.

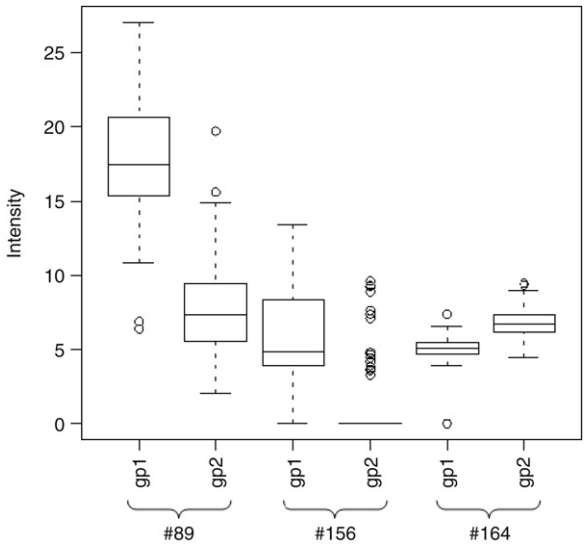

Fig. 3. Boxplots for the intensities of the peaks #89, #164 and #156 with one boxplot per group.

5.1.4. Comparison with other methods results

At last, it is encouraging to notice that those peaks have all been already selected in other published studies using the methods quoted in Section 5.1.1. Indeed, the peak #89 was identified by Alexe et al. (2004) and Sorace and Zhan (2003), the peak #164 by Alexe et al. (2004), Sorace and Zhan (2003), Zhu et al. (2003) and Jeffries (2004) and the peak #156 was found by Jeffries (2004). As those authors all used different methods of peak extraction and selection, finding similar results tends to show that the selected peaks are really meaningful and that our method performs well for two groups, providing a new combination.

5.2. Neurological data

5.2.1. Description of the population

In the framework of the French Creutzfeldt-Jakob Disease (CJD) Surveillance System the Western blotting CSF 14-3-3 analysis and the collection of blood samples have been performed for around 8000 patients initially referred to us as suspected CJD (surveillance period ranging from 01/01/1996 to 01/04/2005). Biological samples were collected through the routine use of the marker in clinical settings. Samples were collected and centrifuged according to classical methods and frozen until analysed. Definite CJD patients ($n = 10$) and patients in whom the diagnosis of sporadic CJD was finally excluded ($n = 27$) were included in the present study. For all the patients, a retrospective analysis of the clinical and eventually neuropathological data has been performed by the French Surveillance System, in order to determine the final diagnosis. Among the 27 patients initial suspected of CJD, 10 ones had a probable or certain Alzheimer disease while 17 had a psychiatric disorder. The latter were initially selected as psychiatric symptoms and may present early signs of CJD. Thus, here, there are $k = 3$ groups and discrimination will use pairwise comparisons.

5.2.2. The data

To generate the SELDI data, 5 μl of each serum sample were diluted into 7.5 μl of 8 M Urea, 1% CHAPS and shaken. Then, 5 μl of this denatured sample were mixed in 195 μl of binding buffer depending to the ProteinChip array (ref). In this study, we used four types of ProteinChip arrays: Q10 (strong anion exchange) at pH9, CM10 (weak cation exchange) at pH4, H50 (hydrophobic interaction chromatography) at 10% of acetonitrile and IMAC30 (immobilized metal affinity capture) loaded with nickel. The arrays were assembled into a bioprocessor and two time 100 μl of the diluted denatured samples were incubated for 1 h on a plate shaker at room temperature. The arrays were washed three times with the proper washing buffer and followed by a final brief water rinse. The arrays were removed from the bioprocessor and arrays air-dried. 0.8 μl of saturated sinapinic acid solution were then applied twice to each spot and air-dried.

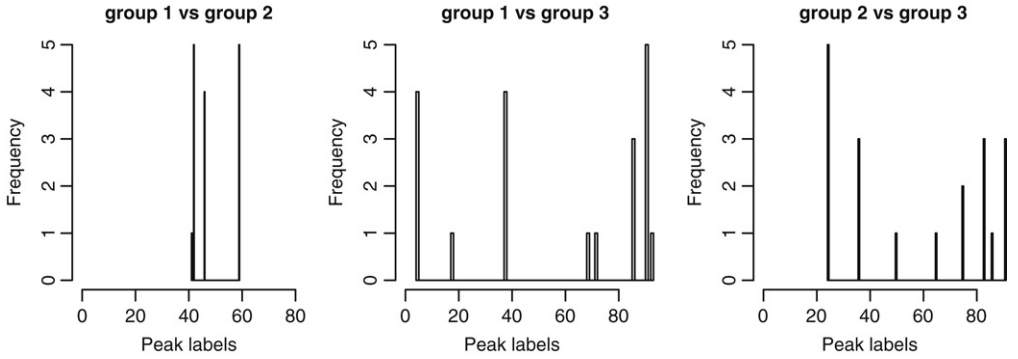

Fig. 4. Histograms for the frequency of selection of the peaks.

SELDI analysis was performed in a PBS-II ProteinChip reader (Ciphergen Biosystems) according to automated setting. Within a single comparative experiment, all conditions were kept the same for data collection (calibration, focusing mass, laser intensity and detector sensitivity). Each spectrum was an average of at least 65 laser shots and externally calibrated with the All-in-1 Protein Standard II (Ciphergen Biosystems). Spectra analysis was carried out using the ProteinChip software version 3.2 (Ciphergen Biosystems). The background was subtracted using the default software settings. All mass spectra were normalized using total ion current that averages the intensity and adjusts the intensity scales for all the spots. SELDI profiles were saved as raw data for bioinformatics analyses.

5.2.3. Preprocessing

The preprocessing steps as previously described are applied. First, the peaks are extracted in each spectrum and then the threshold for the valley depth is chosen and applied. On average, 65 peaks are found in each spectrum. Finally, the peaks are matched between spectra leading to retain 93 classes of peaks. Thus, the matrix X containing the intensities is a 37×93 matrix.

5.2.4. GA results and discrimination

There are three groups in these data, thus, there are three populations in the GA which evolve at the same time. At the end of the GA, three populations are obtained, that is why it is possible to compute a global rate of good classification while having one committee per comparison. Then, we can know which peaks make the discrimination between the first and the second groups,... Moreover, the same peak is allowed to appear in several committees if it is able to make the discrimination between more than two groups.

A GA is a stochastic method, so it is interesting to study the robustness of the results. The GA was run five times on these data and the results are really close. Indeed, the selected peaks for each comparison for the five runs are shown in Fig. 4. For the first comparison (Alzheimer vs Creutzfeldt-Jakob), the committee was always made up of three peaks: peaks #42 and #59 were always found and #46 four out of five times (replaced by #41 in the last run). Concerning the second comparison (Alzheimer vs Psychiatric), the committee consisted in four peaks: peak #91 was always selected and peaks #4 and #38 were found four out of five times. The last comparison (Creutzfeldt-Jakob vs Psychiatric) required four peaks: peak #24 was always selected and three peaks (#83, #91 and #36) were found three times.

Thus, the best results are obtained for the first comparison with very stable committees, the second comparison also gives good results and it is a bit less precise for the last comparison. In fact, other results have shown that the first group is easier to distinguish from the other ones, especially from the second one. This explains the better stability of the committee concerning this group and especially, the one discriminating the first group from the second one. This may be due to the fact that the first group contains more spectra than the other ones leading to more reliable (thus robust) results.

The small size of the sample did not allow to perform *external* cross-validation. Indeed, as explained in Section 3, in the discrimination step (included in the selection step of the GA), the decision rules are built on one part of the data (here 70%) and the rate of good classification is computed on the remaining samples (thus 30%). This *internal* cross-validation is necessary to avoid overfitting. But, as the two last groups contain only ten spectra, it is not possible to move aside spectra before applying the GA.

Table 3
Number of misclassifications by 5-FCV for each of the 5 runs of the GA

# of the run	Number of bad affectations
1	1
2	0
3	2
4	0
5	0

Table 4
Summary of the different methods applied on the neurological data

Method	Nb var	Group 1	Group 2	Group 3	Global
LDA	93	94.12	90.00	80.00	89.19
LDA CV	93	52.94	30.00	50.00	45.95
KNN	93	100.00	100.00	100.00	100.00
KNN CV	93	76.47	20.00	**60.00**	56.76
CT	4	82.35	80.00	90.00	83.78
CT CV	3.59	70.59	30.00	50.00	54.05
RF	7.6	94.12	40.00	20.00	54.46
RF CV	6.53	**100.00**	20.00	20.00	56.76
GA CV	**3**	82.23	**50.00**	40.00	**62.16**

In the first column the method names are found with CV for the results concerning leave-one-out and without CV for the results on the whole training sample. The second is the number of selected variables. For RF, the CV results consist in the average of 500 trees. The other columns correspond to the rate of good classification for the three groups and for all the spectra (last column). The bold numbers are the best ones in each column for leave-one-out results.

In a first attempt to make the results more reliable, the committees (peaks and thresholds) were selected by GA using all the spectra (always with internal cross-validation). Then, the decision rules were made and tested by external 5-fold cross-validation (5-FCV). Indeed, even if the peaks in the committee are the same, as the spectra used are different for each run, the proportions of each group above and below the thresholds are different too, leading to different decision rules. The 5-FCV was performed five times (one for each run). The results are enclosed in Table 3. In average, we obtain 98.4% of good classification.

To pursue the reliability study, despite the small size of the data, an external leave-one-out cross-validation was performed on this data set. A complete GA was applied 37 times leading to 37 solutions. In each run, internal cross-validation was performed with the 36 remaining spectra. In order to be able to compare with more classical method, the leave-one-out was also applied with Linear Discriminant Analysis (LDA), *k*-nearest neighbours (KNN), classification trees (CT) and random forests (RF) (Hastie et al., 2001). They have been implemented on R© (R Development Core Team, 2004) using the following functions: (*lda*(MASS), *knn*(class), *tree*(tree), *randomForest*(randomForest)). The default arguments were used in each function excepted KNN were the number of neighbours have been chosen between 1 and 10 so that the number of good classifications using leave-one-out was the best one (one neighbour was selected). The results are enclosed in Table 4.

The results without cross-validation are given as indication but cross-validation is the more interesting as it will allow a comparison with our method. In a general point of view, the global results are quite poor especially for the second and third groups. This is likely to be due to the data structure (is there a real difference between groups 2 and 3 ?...) and to the small size of the two last groups. Our method obtains the best global results (by CV) followed by KNN and RF which tie. It has to be noticed that the best number of neighbours for KNN is only one which seems a bit weird but quite efficient (especially without CV). The worst one is LDA. It is certainly linked with the non linearity of the separations between groups and also with the fact that the data are not discretized leading to overfit them. Concerning the number of selected variables, our method chooses the smallest amount showing the efficiency of our algorithm to select very relevant peaks.

Finally, it can be noticed that the peaks found here were not identified using the univariate statistical tools provided by the Ciphergen software. It is mainly due to the fact that certain peaks can be inefficient to discriminate groups by themselves but combined with other ones, they become very useful. None of the single peaks could have provided

a discrimination which would have been as good as the one obtained here. This may also be due to the step of binarization of the data which become more coarse but more robust too.

6. Conclusion

SELDI-TOF mass spectrometry is an interesting technology but it mainly suffers from its lack of accuracy, especially in intensity. Then, the binarization of the data seems to be a good way of taking this problem into account. Then, two questions arise: the choice of the peaks to be used to discriminate and, simultaneously, the choice of their intensity thresholds. As there are a huge number of possibilities, the GA are a good method of stochastic optimization, indeed, they make a whole population of potential solutions evolve. Thanks to a classical data set, the relevance of those choices have been shown to lead to satisfying results.

However, it must be noticed here that the power of GA has got a time cost. It has been shown that it is possible to extend our method to more than two groups using pairwise coupling. The implementation of this method in GA has been performed and gives promising results. There is theoretically no limit to the number of groups but as $k(k-1)/2$ populations have to be built and run for k groups, the limitation will come from the computation time. This computation time is reasonable for 3 groups, indeed the GA has been implemented in Matlab© and run on a basic PC, it took eight minutes for 500 generations and a population size of 200 individuals.

The obtained results are satisfying for the two data sets concerning both the relevance of selected peaks and their small number: a very good discrimination is obtained thanks to a few peaks. The first data set has been often used and the peaks selected by our method had already been found by other methods (but never together). Moreover, we obtained very precise predictions. So this application allowed showing the consistency of our method. Concerning the second data set, it allows seeing that the generalization to more than two groups is efficient too but the results will have to be verified on the bigger samples we will have at our disposal in several months.

Finally, having implemented a method which is completely independent on the Ciphergen software, it not only compensates the lack in their alignment method and in the statistical analysis but it can be applied on other spectrometry technologies, allowing comparisons.

The GAs (for two and more than two groups) have been implemented in Matlab© (2004) code and are available on demand to the corresponding author.

Acknowledgements

The authors want to thank the associate editor and the two reviewers for their relevant comments. They involved a real improvement of the manuscript.

References

Alexe, G., Alexe, S., Liotta, L.A., Petricoin, E., Reiss, M., Hammer, P.L., 2004. Ovarian cancer detection by logical analysis of proteomic data. Proteomics 4, 766–783.

Ambrogi, F., Lama, M., Boracchi, P., Biganzoli, E., 2007. Selection of artificial neural network models for survival analysis with Genetic Algorithms. Computational Statistics and Data Analysis 52, 30–42.

Baggerly, K.A., Morris, J.S., Wang, J., Gold, D., Xiao, L.-C., Coombes, K.R., 2003. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics 3 (9), 1667–1672.

Bhandari, D., Murthy, C.A., Pal, S.K., 1996. Genetic algorithm with elitist model and its convergence. International Journal of Pattern Recognition and Artificial Intelligence 10 (6), 731–747.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth/Brooks/Cole, Monterey.

Chatterjee, S., Laudato, M., Lynch, L.A., 1996. Genetic algorithms and their statistical applications: An introduction. Computational Statistics and Data Analysis 22, 633–651.

Chen, S, Hong, D, Shyr, Y., 2007. Wavelet-based procedures for proteomic mass spectrometry data processing. Computational Statistics and Data Analysis 52, 211–220.

Coombes, K.R., Fritsche, H.A., Clarke, C., Chen, J.-N., Baggerly, K.A., Morris, J.S., Xiao, L.-C., Hung, M.-C., Kuerer, H.M., 2003. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clinical Chemistry 49 (10), 1615–1623.

DiMatteo, I., Genovese, C.R., Kass, R.E., 2001. Bayesian curve-fitting with free-knot splines. Biometrika Trust 88, 1055–1071.

Duda, R., Hart, P., Strok, D., 2001. Pattern Classification and Scene Analysis. John Wiley and Sons, New York.

Goldberg, D.E., 1989. Genetic Algorithms in Search, optimisation and Machine Learning. Addison-Wesley, Reading, Massachussetts.

Green, P.J., 1995. Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. Biometrika 82, 711–732.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning, Data Mining, Inference and Prediction. Springer-Verlag, New York, 533p..

Jeffries, N.O., 2004. Performance of a genetic algorithm for mass spectrometry proteomics. BMC Bioinformatics 5 (180).

Jong, K., Marchiori, E., van der Vaart, A., 2004. Analysis of proteomic pattern data for cancer detection. Lecture Notes in Computer Science 3005, 41–51.

Kapetanios, G., 2007. Variable selection in regression models using nonstandard optimisation of information criteria. Computational Statistics and Data Analysis 52, 4–15.

Liu, H., Li, J., Wong, L., 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics 13, 51–60.

Matlab©, version 7.0.0.19920. 2004. The MathWorks, Inc.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A., 2002. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359, 572–577.

Prados, J., Kalousis, A., Sanchez, J.-C., Allard, L., Carrette, O., Hilario, M., 2004. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. Proteomics 4 (8), 2320–2332.

Qu, Y., Adam, B.-L., Yasui, Y., Ward, M.D., Xazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J., Wright, G.L., 2002. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clinical Chemistry 48 (10), 1835–1843.

Reeves, C.R., Rowe, J.E., 2003. Genetic Algorithms — Principles and Perspectives, A Guide to GA Theory. Kluwer Academic Publishers, London.

R Development Core Team. 2004. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

Reynès, C., Roche, S., Tiers, L., Sabatier, R., Jouin, P., Molinari, M., Lehmann, S., 2007. Comparison between surface and bead based MALDI profiling technologies using a single bioinformatics algorithm. Clinical Proteomics (in press).

Salomon, R., 1996. The influence of different coding schemes on the computational complexity of genetic algorithms in function optimisation. In: Proceedings of the Fourth International Conference on Parallel Problem Solving from Nature. Springer-verlag, Berlin, pp. 227–235.

Sorace, J.M., Zhan, M., 2003. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 4 (24).

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le, Q.-T., 2004. Sample classification from protein mass spectrometry, by peak probability contrasts. Bioinformatics 20 (17), 3034–3044.

Wagner, M., Naik, D., Pothen, A., 2003. Protocols for disease classification from mass spectrometry data. Proteomics 3, 1692–1698.

Wu, T.-F., Lin, C.-J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005.

Yasui, Y., Pepe, M.S., Thompson, M.L., Adam, B.-L., Wright, G.L., Qu, Y., Potter, J.D., Winget, M., Thornquist, M., Feng, Z., 2003. A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. Biostatistics 4 (3), 449–463.

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., Kovach, J.S., 2003. Detection of cancer-specific markers amid massive mass spectral data. Proceedings of the National Academy of Sciences 100 (25), 14666–14671.