BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 2. 119–128
Advance Access publication February 29, 2008
doi:10.1093/bib/bbn008

Machine learning methods for predictive proteomics

Annalisa Barla, Giuseppe Jurman, Samantha Riccadonna, Stefano Merler, Marco Chierici and Cesare Furlanello

Submitted: 14th September 2007; Received (in revised form): 25th January 2008

Abstract

The search for predictive biomarkers of disease from high-throughput mass spectrometry (MS) data requires a complex analysis path. Preprocessing and machine-learning modules are pipelined, starting from raw spectra, to set up a predictive classifier based on a shortlist of candidate features. As a machine-learning problem, proteomic profiling on MS data needs caution like the microarray case. The risk of overfitting and of selection bias effects is pervasive: not only potential features easily outnumber samples by 10^3 times, but it is easy to neglect information-leakage effects during preprocessing from spectra to peaks. The aim of this review is to explain how to build a general purpose design analysis protocol (DAP) for predictive proteomic profiling: we show how to limit leakage due to parameter tuning and how to organize classification and ranking on large numbers of replicate versions of the original data to avoid selection bias. The DAP can be used with alternative components, i.e. with different preprocessing methods (peak clustering or wavelet based), classifiers (e.g. Support Vector Machine) or feature ranking methods (recursive feature elimination or I-Relief). A procedure for assessing stability and predictive value of the resulting biomarkers' list is also provided. The approach is exemplified with experiments on synthetic datasets (from the Cromwell MS simulator) and with publicly available datasets from cancer studies.

Keywords: *proteomics; selection bias; feature selection; functional profiling*

INTRODUCTION

Proteomic profiling is applied to identify biomarkers relevant for molecular diagnostic or prognostic purposes. Technologies for high-throughput mass spectrometry (MS) are experiencing a substantial growth and diffusion, giving researchers the opportunity not only to face new biological questions but also to improve the techniques used at different steps of the profiling process. First, a growing number of both matrix-assisted laser desorption/ionization (MALDI) and surface-enhanced laser desorption/ionization (SELDI) datasets has become publicly accessible. Secondly, availability of signal processing and machine-learning methods has also recently spread, with the final aim of selecting a list of biomarkers that stand the best chance of predicting

Corresponding author. Cesare Furlanello, FBK, via Sommarive 18, I-38100 Povo (Trento), Italy. Tel: +39-0461-314580; Fax: +39-0461-314591; E-mail: furlan@fbk.eu

Annalisa Barla (PhD, Computer Science, University of Genua, 2005) is a postdoctoral fellow at DISI, Genua. She worked on statistical learning theory with applications of kernel engineering for image classification. Her most recent research deals with bioinformatics methods for variable selection in genomics and proteomics.

Giuseppe Jurman (PhD, Mathematics, University of Trento, 1998) is a fellow researcher in the Predictive Models for Biological and Environmental Data Analysis (MPBA) unit at Fondazione Bruno Kessler (FBK), dealing with the mathematical structure of machine-learning and bioinformatics algorithms.

Samantha Riccadonna (MSc, Mathematics, University of Trento, 2004). She is currently enrolled in the PhD program at DISI, University of Trento with an internship in the FBK-MPBA lab focusing on statistical data mining and integrative bioinformatics.

Stefano Merler (MSc, Mathematics, University of Trento, 2004). He is a member of FBK-MPBA, with main interests in the development and application of machine-learning techniques for modeling of infectious diseases, bioinformatics and environmental epidemiology.

Marco Chierici (PhD, Bioengineering, University of Padua, 2007) is currently a postdoctoral fellow at the MPBA lab at FBK, with proteomics data analysis and population genetics as main areas of interests.

Cesare Furlanello (MSc, Mathematics, University of Padua, 1986) is Senior Researcher and responsible for the MPBA unit at FBK, Trento, Italy. He has developed predictive models and software infrastructures based on machine learning with applications to bioinformatics and geoinformatics data.

a class label (classification) or a response indicator (regression). In this review, we focus on the design of predictive classifiers and of feature selection on MS data; we treat the different preprocessing and biomarker selection methods that have been proposed as given building blocks of a chain (or pipeline) of spectra analysis and profiling methods. We consider the problem of either studying a new element of the preprocessing and classification chain or just of setting up one chain to address a clinical or prognostic question. In practice, we aim at defining how data preparation and profiling methods have to be organized in a complete schema of procedures, given a set of labeled spectra and a classification task with classes defined by labels.

The first aim is just to achieve reproducible results. Indeed, similar to the evolution of microarray technology, a critical revision of proteomic profiling has arisen, pointing out the need for the most careful handling of the preprocessing and modeling tools [1–4]. In order to ensure reproducibility of experiments, a main critical issue is the experimental design methodology, which may pose the results at risk of selection bias [1, 5–10]. For microarrays, a set of comprehensive guidelines for the development and validation of predictive models or classifiers, likely to give biomarkers and classifiers with reproducible outcomes on novel data, is the expected results of MAQC-II, a multicentric project promoted (by the U.S. Food and Drug Administration (FDA) [11]). Here we try to introduce similar concepts and a general data analysis workflow that deals also with the preprocessing issues typical of MS data.

The workflow for a proteomic profiling process is typically seen as a pipeline concatenating two engines: a preprocessor engine with a classification-ranking one. Spectra enter the pipeline as raw data, undergo a sequence of elementary normalization and feature extraction steps so to get finally encoded as a feature matrix. One row, i.e. a vector of real values, and one label become one input–output pair to be learned by a classifier. Defining the classifier (between different alternatives) and tuning it for best performance is the goal of the second engine, which also provides a sublist of features with top relevance for discrimination. To ensure reproducibility, we need to specify how the available data material is fed to the system. The key to reproducibility is vastly in the system of specifications that are given for selecting training, and test datasets (extracted by partitioning or subsampling from the data at hand) to simulate the accuracy of the classifier on future data and the interest of using the selected features as prognostic markers. It is worth laying out a design analysis protocol (DAP) in which data preparation, preprocessing, model selection, and performance evaluation are all detailed. The diagram in Figure 1 outlines a generic DAP workflow for proteomic profiling. We suggest focusing first on the organization of the system than on its specific elements. As an example, the preprocessing encoding from spectra to features (upper left of Figure 1) can be organized by combination of elementary preprocessing methods. When dealing with MS profiling, crucial is such initial upstream preprocessing phase obtained through a concatenation of smoothing, normalization and feature extraction methods (in some order). A wide variety of techniques have been introduced to tackle the preprocessing steps from raw data to the peak extraction and quantification [5, 12–14], as well as to deal with the downstream phase, from peaks to classification models and biomarker identification [4, 15].

UPSTREAM ANALYSIS

Extensive low-level processing can be combined with machine learning also in the first phase: clustering approaches have been successfully used to match peaks across spectra (see [13] with updates in [16]). Computer vision approaches to signal handling are often applied to proteomics spectra as in [14, 17]: the authors deal with the multiple peak alignment task by using a scale-space method in the first case and a robust point matching algorithm in the second. Noy and Fasulo [18] introduced a model-based approach to feature extraction in which spectra are decomposed into a mixture of distributions derived from peptide models. Klann and coworkers [19] perform deconvolution of signals with peak-like structures and propose wavelet shrinkage, i.e. a regularization procedure, for model selection. A hybrid ant colony optimization (ACO) and support vector machines (SVM) approach is used by Ressom and coworkers [20] to select a parsimonious set of biomarkers from MALDI time-of-flight data (MALDI-TOF), after a preprocessing phase that involves outlier screening, binning, baseline removal and spectra normalization by total ion current (TIC). Note that normalization by area under the curve (AUC) usually requires multiplication for a scale factor such as the mean or median AUC to ease

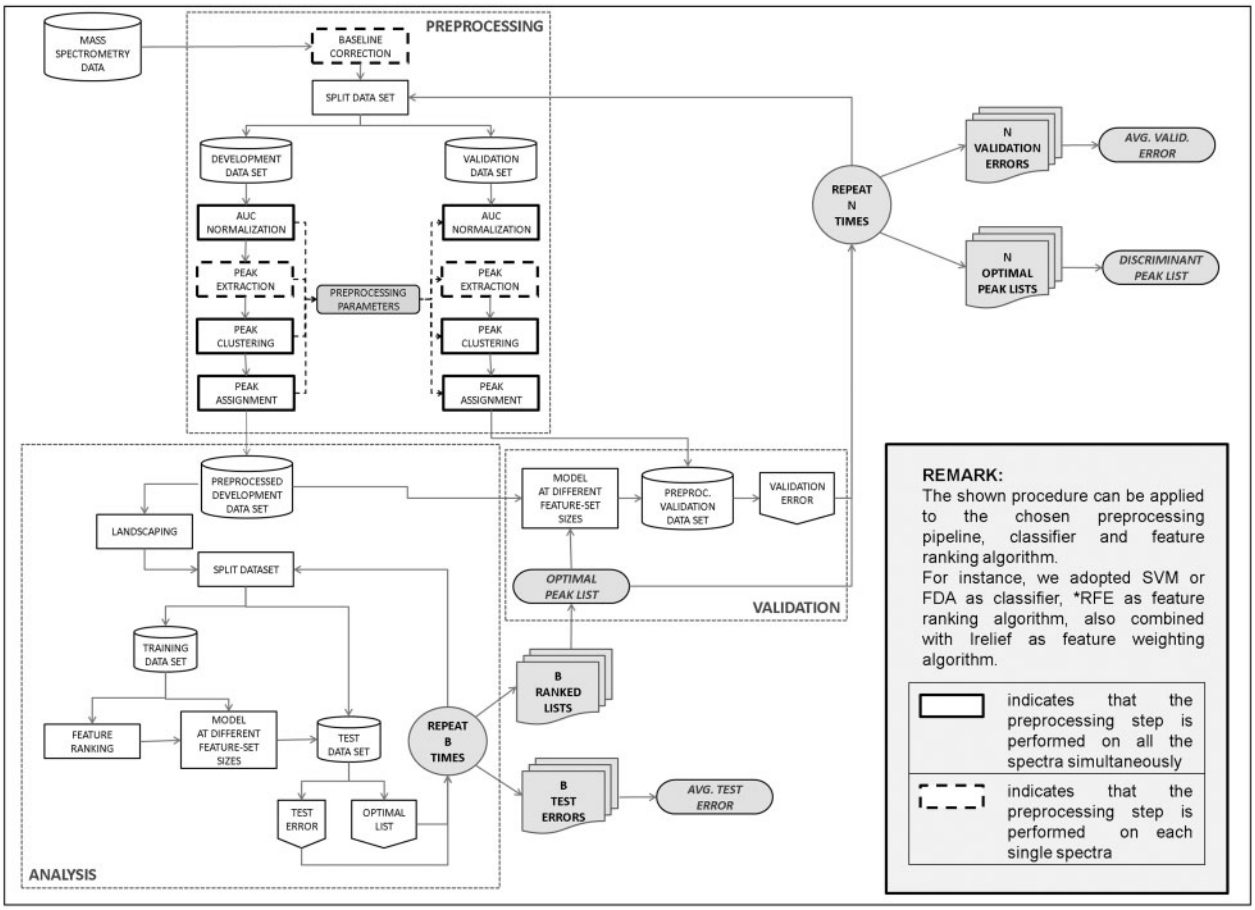

Figure 1: A workflow for proteomic profiling.

visualization [21, 22]. In [23], a low-level processing algorithm for MS spectra is presented, based on a denoising phase with the undecimated discrete wavelet transform (UDWT). To complete the preprocessing phase two main parameters have to be set in this case: the type of wavelets to be used and the threshold level. The authors employ Daubechies wavelets [24] and set the wavelet degree to eight for all analysis. By transforming the signal to the wavelet domain thresholding, and back-transforming they achieve an effective noise reduction, while detecting the interesting peaks.

DATA ANALYSIS PROTOCOL

The first step in a proteomic profiling task is the design of a protocol organizing the data life cycle, from a collection of raw MS spectra to a list of candidate biomarkers. We need to specify elementary steps for the preprocessing, how to partition data in development-validation splits (and development sets in training-test parts), classification and feature ranking procedures and validation of the resulting models.

The structure of a DAP workflow for proteomic profiling is displayed in Figure 1. The MS spectra are first passed to the preprocessing engine. As soon as possible, we set apart a portion of the data for validation, and then apply a pipeline of preprocessing modules to the development data. Ideally, each step should be applied to spectra separately (dashed boxes in Figure 1), i.e. preprocessing parameters should be used unmodified during upstream analysis of validation (e.g. with AUC normalization).

The workflow expands an experimental schema for gene expression profiling [25]. Predictive models and feature ranking are based on SVM classifiers coupled with an Entropy Based-Recursive Feature Elimination (E-RFE) variable selection step. The classification and ranking analysis (lower left box in Figure 1) is formed by a preparation phase and an internal validation loop of B runs derived from the BioDCV profiling method [25]. For each run, data

are split in two datasets by random partitioning at fixed ratio (e.g. 75% for training and 25% test: 'Split data set' block in Figure 1). Splits include stratification for class labels and any potential sub grouping available from clinical data and a priori knowledge. The training portion is used for training classifiers and for model selection. Within the BioDCV internal training steps, different kernel methods (e.g. linear and non-linear SVM, data-driven kernels [26], linear SVM) can be trained for classification. We can also apply alternative feature ranking algorithms (mostly from the wrapper family: recursive feature elimination (RFE) and variants, I-Relief [27]). The pair (classifier, ranking method) will be considered as elementary model and evaluated on the test dataset at increasing feature subset sizes. As a first result, all the B accuracy estimates (test error) will be used to obtain average test error (ATE) measures, with confidence intervals, at each feature set size. The set of B ranked lists will be used to create an optimal list of the features most present in the top positions or having best mean positions [28]. One set of ranked peak lists and predictive accuracy estimates (total and per class) will be derived, given the experimental condition and the current development dataset. In [28] we show by algebraic methods that the entire set of ranked lists obtained in the B internal validation runs can be used to evaluate the distributions of the mutual distances between partial top-k sublists, obtaining a '*stability*' measure of the resulting biomarker list. Stability can be used together with accuracy to select the optimal models. Classifiers based on optimal models and optimal feature lists are applied on the validation dataset. Preprocessing, analysis and validation are then repeated N times, thus obtaining an average validation error that can be used as an estimate of performances on novel data, as shown in the rightmost upper section of the DAP graph.

The same DAP diagram can be used to compare different classifiers or alternative preprocessing methods. In particular, to ensure that the procedure is not affected by any systematic bias, a suitable number of replicated experiments are run on the dataset after having randomly permuted the sample labels: to be evaluated as unbiased, the procedure should produce an ATE curve lying close to the dataset no information error rate (i.e. the samples ratio between the smallest class and the whole dataset sizes, corresponding to the error reached by classifying all samples as belonging to the most populous class). The preprocessed dataset (from development data) is often analyzed by a battery of different classifiers: different classification models are evaluated at a grid of parameters so to get a first landscape of predicted measures of accuracy as function of classifier type and parameters' choice. Alternative models will be evaluated by k-fold cross-validation (k-CV), with k chosen according to dataset dimensions (typically $k \leq 10$), or by resampling. The list of tested models will include SVM (different kernels), kNN, classification trees and ensemble methods (bagging and boosting) amongst the aforementioned. Tables with expected accuracy (total and per class) will be analyzed and optimal parameters will be chosen for a reduced set of best classifiers. Typically one or two types of classifiers and a limited number of alternative parameters are selected for internal validation. Finally, outlier and subtype detection can also be obtained with the same schema: spectra can be evaluated for possible removal (shaving) [29] and evidence for potential unreported subtype structure investigated [30] and considered for stratification in the subsequent analyses.

We emphasize that correctly characterizing the overall schema so to avoid the effects of selection bias in the analysis is the core message of the paper. Working examples with our solutions and alternative approaches (both for preprocessing and machine-learning components) will be provided to exemplify its use in practice.

PEAK CLUSTERING

Within the proteomic profiling DAP, we can compare alternative preprocessing engines. The example used by the DAP in Figure 1 implements a pipeline of existing modules for a preprocessing based on peakclustering (PC): the steps are baseline correction, AUC normalization, peak extraction, peak clustering and peak assignment [31,13]. Each MS spectrum is described by a set of pairs of ordered mass-over-charge ratio (m/z) values and corresponding intensity: the preprocessing phase is generally meant to reduce dimensionality (number of m/z values) and to obtain a set of potential markers at the local maxima of the signal. In this example, baseline is estimated and removed by loess (local linear regression) [22] over the local minima with the *PROcess* package of the R statistical environment [32]. Each signal is then normalized 'across spectra' by using a function of the AUC [31] (the formula

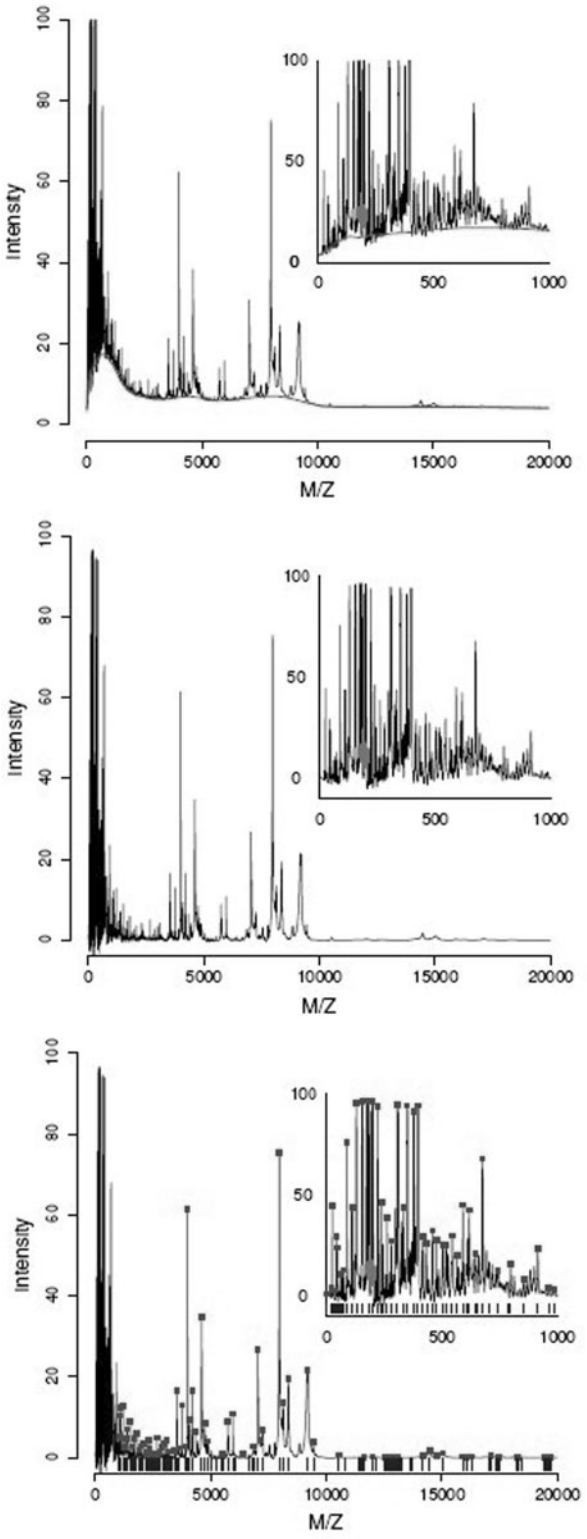

Figure 2: Preprocessing of one cancer patient spectrum (dataset OVI, see Supplementary Material). Top: raw spectrum with baseline identification. Middle: normalization after baseline subtraction. Bottom: peak extraction (squares) and centroid identification (lines). Insets show leftmost range ($m/z < 1000$Da). Modified from [15].

includes the median AUC). As in [13], a peak candidate is detected given three conditions: (i) it must have the highest intensity among its nearest neighbors, (ii) it must be higher than a chosen threshold and (iii) higher than a function of the signal-to-noise ratio (SNR). Furthermore, two candidate peaks overlap if they both lie in a window of a chosen width (the 'peak width', t). The peaks from all spectra are grouped through complete hierarchical clustering on the logarithmic scale of their m/z [13]. Another parameter (the 'peak gap', l) defines the cut level for the clustering and thus a partition whose cluster centroids define the 'common peaks', i.e. those common to all spectra. For each spectrum, a peak is selected if its distance from a common peak is $< s = \log(l)$. Peaks are much larger and less overlapping for higher m/z ranges, thus the threshold s needs to be focused or adapted to the sub range of interest: this is also sensible if the instrument is known to be calibrated in a fairly narrow range. Figure 2 displays an example of the whole process on a real spectrum.

Expected peak width and peak gap are good examples of parameters that are critical for DAP's automation: they are related to the physical characteristics of the spectrometer and to the m/z range of interest. The mean number of detected peaks decreases significantly as a function of the width parameter. Usually, these parameters are manually fine-tuned after a visual inspection or by using a priori knowledge.

One should mark which preprocessing steps are applied to each spectrum separately (i.e. baseline correction and peak extraction), and which (normalization, peak clustering and peak assignment) have to be performed on all spectra at the same time and thus need a careful handling. For the latter steps, we need to decouple preprocessing of development and validation datasets because there is a potential opportunity for overfitting effects whenever development and validation datasets are simultaneously preprocessed. In our case, to avoid information leakage, we use the median AUC from development to normalize spectra in validation—and not a median AUC computed on all data.

ALTERNATIVE PREPROCESSING

As an alternative example for the preprocessing engine within the DAP, we employ a wavelet denoising procedure proposed by Coombes and

coworkers in [33] (as an evolution of [23]) for common peaks identification and intensity quantification (see Supplementary Material). To the classification engine we pass the feature matrix from either one of the two approaches (the readers may substitute here their favorite approach or introduce a new component).

Preprocessing methods often require tuning of window sizes or of thresholds. Machine-learning principles can be used to automate this practice: for the peak-clustering method, we build a classifier whose features are picked up from all those detected by preprocessing at different peak width and peak gap choices. We call 'multiresolution' (MR) this procedure of ranking over the aggregated features, leaving to the feature selection algorithm in BioDCV the responsibility of choosing the most discriminating features, possibly defined by different peak width and peak gap pairs in distant intervals of the spectrometer range.

EXAMPLES

We can illustrate the general strategy for machine-learning profiling DAP in concrete with examples on classification of two synthetic and two real MS datasets (described in Supplementary Material). Given workflows as in Figure 1, we analyzed different preprocessing procedures, classifiers and feature ranking algorithms. The MR algorithm was also tested.

Two synthetic datasets (MR1 and MR2) were created with the Cromwell proteomic MALDI-TOF simulation engine [3]. The spectra in MR1 were set so to have 81 peaks of which 14 discriminant with different intensities (10 000 and 5000) assigned for two classes. Four instances of MR1 were generated for increasing levels of noise over intensity (see Supplementary Material for details). The synthetic dataset MR2 mimics a clinical diagnostic experiment setup in which potential discriminating peaks belong to two different and distant regions within the *m/z* operation range of the spectrometer. This two-band structure is sketched in Figure 3.

The two real datasets regard discrimination of ovarian cancer from controls on two different platforms (OV1: 253 SELDI-TOF spectra; OV2: 170 MALDI-TOF spectra, described in Supplementary Material).

A profiling experiment, following the DAP workflow outlined in Figure 1, was run on MR1, OV1 and OV2 datasets, choosing PC as preprocessing engine, SVM as classifier, and RFE as ranking algorithm.

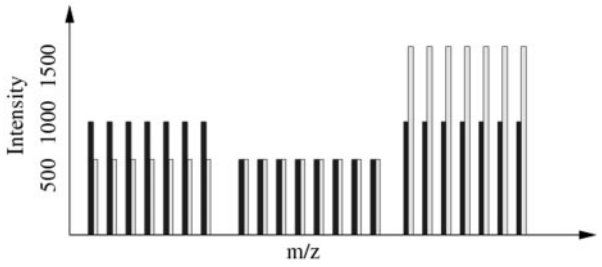

Figure 3: A two-band distribution of peaks (synthetic dataset MR2): bars indicate average intensities for class 1 (in light) and -1 (in dark); peaks discriminate classes in two distant regions of the simulated *m/z* range.

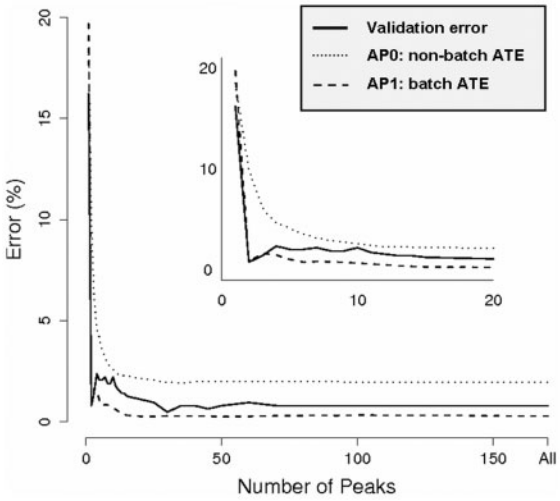

Figure 4: Comparing error estimates (cancer data OVI) on validation with ATE for batch preprocessed and non-batch data. Inset: detail of the error curves. Modified from [15].

For MR1, at all the noise levels, the 81 peaks were found. Higher noise levels introduced extra non-discriminant features but they were low ranked. In all runs, 13 out of 14 peaks were always detected as top discriminating (the leftmost was confused with non-discriminating peaks).

The effect of batch preprocessing of training and test data together is shown in Figure 4 for the OV1 data. A discrimination error below 3% (AVE: average error on validation) is obtained with <5 features by profiling on these controversial data [34]. The average test error (ATE) is more optimistic than the validation error when training and test data in the development are processed all together. The ATE is instead upward biased when test data are preprocessed with parameters computed on training.

Consistency with previous work has been shown on the OV2 dataset and $N = 5$ development/validation splits and $B = 400$ training/test runs [15]: in this configuration we use about 128 samples in development and get AVE = 24.5% with all features (best result on validation), to compare with a 5-CV error on 136 samples ranging from 21 to 27% in [35]. The most discriminating peaks for [35] are also confirmed: mean spectra (AVG) for the two classes are shown in Figure 5 for the first (3960 Da) and fifth (4060 Da) most relevant peaks. In the same DAP configuration, we also checked for potential selection bias by randomly permuting the class labels obtaining AVE = 49.1%, i.e. above the no-information rate (45.3%) on all data.

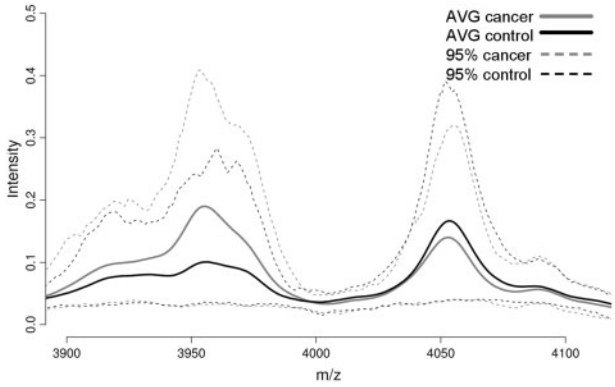

Figure 5: Discriminating peaks (cancer data OV2) in the [3900, 4100] Da interval, mean spectra (AVG) with 95% CI. Modified from [15].

The BioDCV profiler can be used with different preprocessing engines. We applied a DAP as in Figure 1 with $N = 10$ splits to compare the PC and the wavelet denoising (WD) preprocessing pipelines (described above) on the OV2 dataset. Although the final classification scores are fairly comparable, the methods found smaller sets of candidate features with WD preprocessing (PC set sizes: min = 76, med = 79.5, max = 83; WD: min = 53, med = 54.5, max = 60). Locations of candidate peaks can change for different splits, as shown in Figure 6. The diagram also displays the agreement between methods. Note that here for PC we choose peak width and peak gap to favor detection of peaks in the rightmost region of the spectrum: a different resolution would give a different alignment diagram (but not necessarily different final biomarker lists). Effects on AVE curves are discussed in Supplementary Material (Figure 2S).

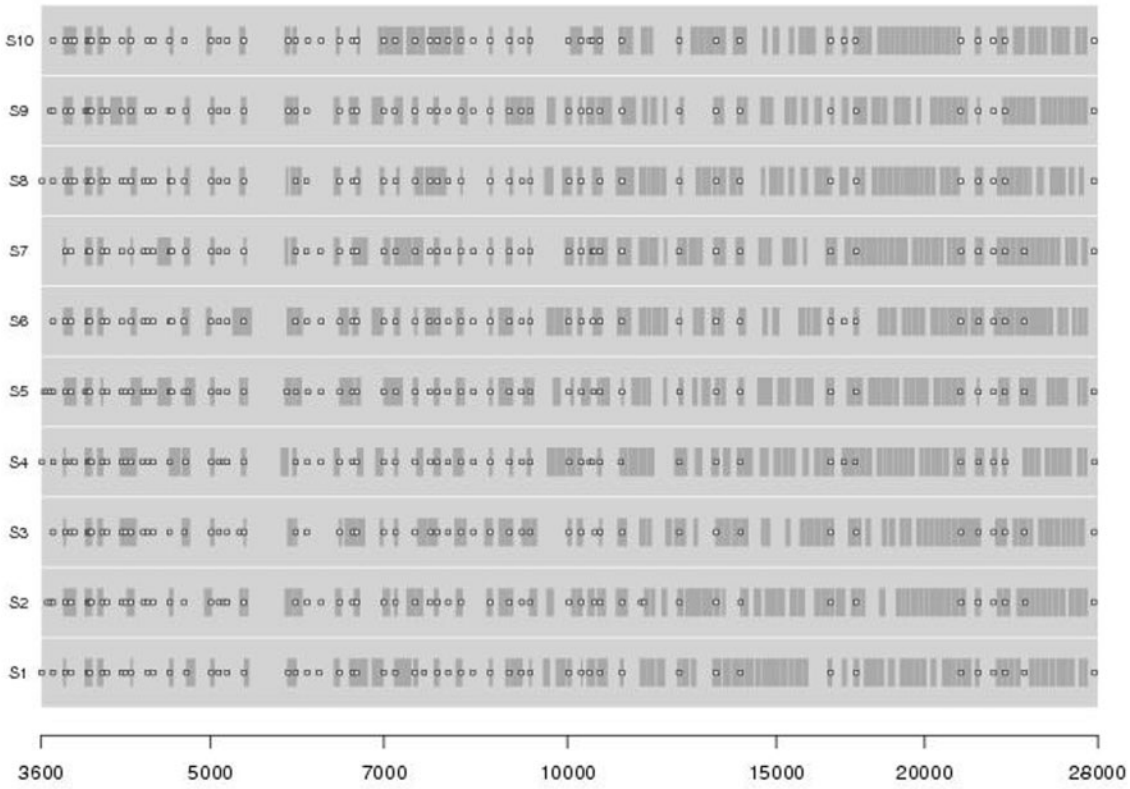

Figure 6: OV2 MS data: peak alignment diagram for two preprocessing methods (PC: darker rectangles for peak cluster ranges; WD: white squares for peak locations; *x*-axis: *m*/*z*) and 10 dev/valid splits (S1–S10).

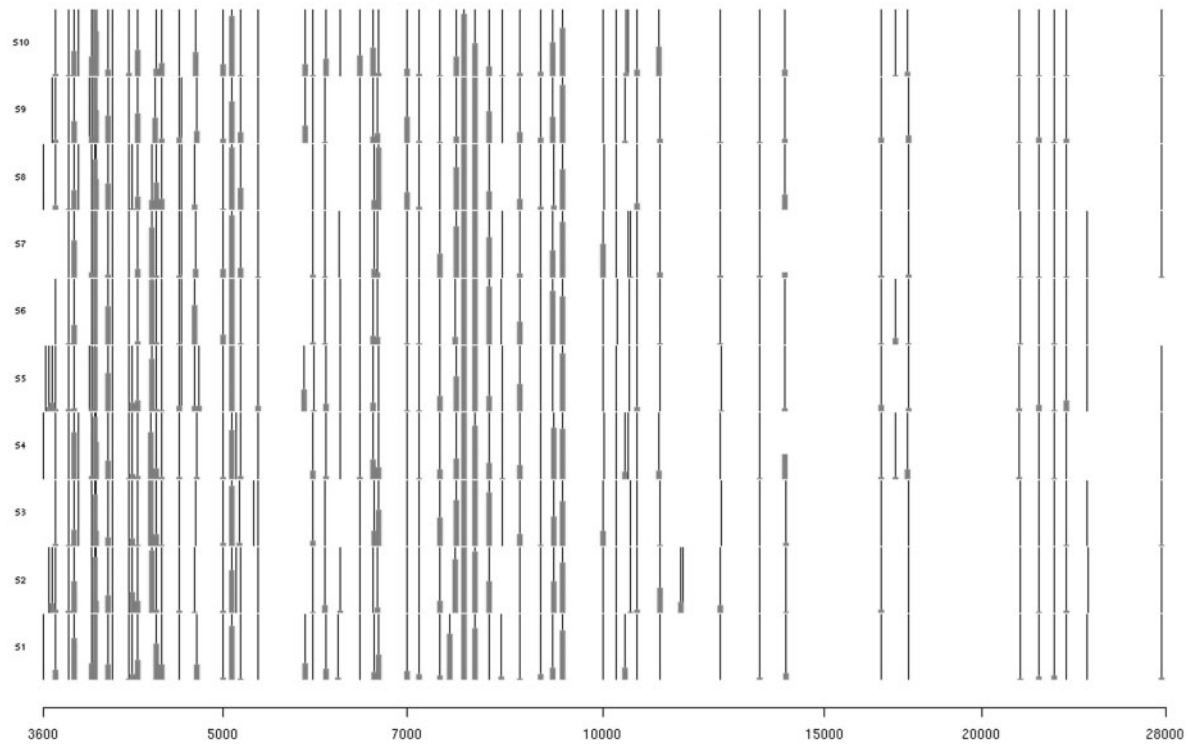

Figure 7: Peak relevance diagram (WD preprocessing, SVM/RFE, $N = 10$ splits S1–S10, *x*-axis: *m*/*z*) on OV2 cancer dataset. Candidate peaks are marked by thin vertical segments: height of grey bars indicates relevance as fraction of times in top-10 lists after classification and ranking.

It is interesting to study how peak importance (according to machine learning) is reproducible for different validation splits. We applied profiling and quantified the best ranked features (WD preprocessing, SVM/RFE) for OV2 on $N = 10$ splits (S1–S10). The diagram in Figure 7 (see colour version in Supplementary Material) highlights peaks' relevance by counting for each peak candidate its frequency in the resulting top-10 positions after ranking. Several locations are consistently top ranked in all splits, with most discriminant peaks concentrated below 10 000 Da. The peaks shown in Supplementary Figure 5S are also consistently top ranked.

The main use of this approach is possibly the fair comparison of alternative preprocessing or classification modules. We preprocessed the MS cancer dataset OV2 by both PC and WD methods, and then applied a suite of classifiers (kNN, trees, Linear, Gaussian, data-driven SVM: see Table 3S in Supplementary Material) without feature selection and different parameters. In this 'landscape' of experimental conditions we found that kNN and tree stumps have poor accuracy even on the development sets. The other methods (i.e. SVM with different kernels and trees) reach zero classification error on the development data, but SVM perform better in validation. The linear kernel is the best classifier for both preprocessing pipelines, but all classifiers have better performances with WD.

Finally, we consider the testing of hybrid methods that integrate preprocessing and machine-learning components. The new MR profiling procedure was applied with PC preprocessing, linear SVM-RFE to the synthetic MR2 and cancer OV2 datasets, with $N = 3$ splits. On the MR2 dataset, we applied the machine learning engine on the union set of peaks preprocessed at two different resolutions R_1 and R_2, one tuned for <4400 Da and another for >17 500 Da. At MR, peaks from both the resolutions are selected, with AVE(MR) = 4.6% which improves over AVE(R1) = 5.6% and AVE(R2) = 30.6% (see Supplementary Material for details). For the OV2 dataset we considered 10 resolutions R_1–R_{10} (see Supplementary Material). The peaks detected at resolution R_2 and R_8 in two different regions of the spectra are shown in Supplementary Figure 3S. Increasing values of peak width and gap allow identifying peaks toward the rightmost part of the spectra, as shown in Supplementary Figure 4S. The average *m*/*z* of the detected peaks does not depend

on the resolution, but their average position weighted by relative importance increases with peak width and gap while the discriminant peak positions are concentrated in the leftmost part of the spectra. As shown in Supplementary Table 5S, with MR, the minimum validation error AVE = 23% is obtained with only three peaks. The accuracy is comparable to AVE = 26% obtained with resolution R_5, but can be obtained without predefining resolution or region of interest.

Grid computing resources were provided by the infrastructures of the European project Enabling Grids for E-sciencE (EGEE) within its Biomed virtual organization. The BioDCV system was run simultaneously on up to 120 CPUs distributed in several computing centers in France, Spain, UK, the Netherlands, Greece, FYROM and Italy.

DISCUSSION

A proteomic profiling study is a complex task. A DAP can be seen as a roadmap leading to best reproducibility of results on novel data through reliable choices of the preprocessing and machine-learning modules. The risk of information leakage is potentially high, thus the need of implementing robust preprocessing methods with limited use of intra-spectra information. To this purpose, automating of parameter tuning and organizing classification and ranking based on large numbers of replicate versions of the original data are essential. We tested the MR approach to include part of parameter tuning within the feature ranking process. We remark that the method identifies peak locations and interval of interest in the MS spectra, i.e. reaching one step before the biomarker identification. Also, in our experiments we did not include information on the calibration range of the instrument, but the method can be easily adapted to focus on specific regions of interest.

The results obtained on synthetic datasets demonstrate the method, and those on the cancer datasets match with literature. Although we are aware of the limitations of the OV1 dataset [34], we demonstrated our methodology by obtaining results similar to those in [1, 36, 37]. While the authors reach near perfect classification with almost any attempted method (Wilcoxon test, kNN, SVM), our slightly worse accuracy is probably due to the caution of separately preprocessing training and test data. On the other hand, with batch preprocessed training/test splits, a close to perfect classification error is achieved. The OV2 data set is harder to classify but includes more reliable data. The obtained results are compliant with previously published results [35].

Moreover, the DAP allowed unbiased estimates of accuracy for different preprocessing engines (the wavelet denoising method performed better than the PC one) and of alternative classification modules. In the MR experiments, the results show that one resolution will be always optimal for one dataset and one instrument. We believe that leaving the task of choosing the best features to the system is especially convenient in order to limit manual tuning, and thus potential dependence effects affecting predictive use of the models and of the biomarkers. Finally, the multiresolution method could be clearly used to combine together features coming from heterogeneous platforms, thus supporting integrative genomic studies.

Key Points

- Experimental design to avoid selection bias.
- DAP workflow: a roadmap for profiling.
- Different approaches inside the same DAP.
- Averaging over replicated experiments.
- Biomarkers stability as a performance measure.
- Multiresolution automates parameter tuning.
- Cromwell Simulation Engine for MS synthetic data.
- Batch preprocessing leads to overfitting.
- Randomize labels to detect flaws.
- Grid computing for proteomic profiling.

SUPPLEMENTARY MATERIALS

Supplementary Materials are available at *Briefings in Bioinformatics* Online.

Acknowledgements

Research partially funded by AIRC, Associazione Italiana per la Ricerca sul Cancro. The authors thank D. Albanese, R. Flor and S. Paoli for computing support.

References

1. Sorace J, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;**4**:24.
2. Baggerly K, Coombes K, Morris J. Bias, randomization and ovarian proteomics data: a reply to "Producers and Consumers". *Cancer Inform* 2005;**1**:9–14.
3. Coombes K, Koomen J, Baggerly K, *et al*. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform* 2005;**1**:41–52.
4. Marchiori E, Jimenez C, West-Nielsen M, *et al*. Robust SVM-based biomarker selection with noisy mass

spectrometric proteomic data. In: *Applications of Evolutionary Computing. EvoBIO: Evolutionary Computation and Machine Learning in Bioinformatics*, LNCS vol. 3907. Berlin: Springer, 2006, 79–90.

5. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 2002;**99**(10):6562–6.
6. Yu W, Wu B, Liu J, *et al*. MALDI-MS data analysis for disease biomarker discovery. *Methods Mol Biol* 2006;**328**(14): 199–216.
7. Baggerly K, Morris J, Coombes K. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004;**20**(5): 777–85.
8. Morris J, Coombes K, Koomen J, *et al*. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 2005; **21**(9):1764–75.
9. Petricoin E, Ardekani A, Hitt B, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002; **359**:572–7.
10. Sun Y, Goodison S, Li J, *et al*. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 2007;**23**(1):30–7.
11. Shi L, Perkins RG, Fang H, *et al*. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol* 2007 Dec 21 [Epub ahead of print] PMID: 18155896.
12. Yasui Y, McLerran D, Adam BL, *et al*. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J Biomed Biotechnol* 2003;(4);242–8.
13. Tibshirani R, Hastie T, Narasimhan B, *et al*. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* 2004;**20**(17): 3034–44.
14. Yu W, Li X, Liu J, *et al*. Multiple peak alignment in sequential data analysis: a scale-space-based approach. *IEEE/ ACM Trans Comput Biol Bioinform* 2006;**3**(3):208–19.
15. Barla A, Irler B, Merler S, *et al*. Proteome profiling without selection bias. *Proc CBMS 2006*; IEEE 2006:941–6.
16. Codrea MC, Jimenez C, Piersma S, *et al*. Robust peak detection and alignment of nanoLC-FT mass spectrometry data. In: *The Fifth European Conference on Evolutionary Computation, Machine Learning and Datamining in Bioinformatics*, LNCS vol. 4447. Berlin: Springer, 2007, 35–46.
17. Yu W, Wu B, Lin N, *et al*. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comput Biol Chem* 2006;**30**(1):27–38.
18. Noy K, Fasulo D. Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics* 2007;**23**(19):2528–35.
19. Klann E, Kuhn M, Lorenz DA, *et al*. Shrinkage versus deconvolution. *Inverse Problems* 2007;**23**:2231–48.
20. Ressom H W, Varghese R S, Drake S K, *et al*. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 2007;**23**(5):619–26.
21. Coombes K, Morris J, Hu J, *et al*. Serum proteomics profiling–a young technology begins to mature. *Nat Biotechnol* 2005;**23**(3):291–2.
22. Chambers J, Cleveland W, Kleiner B, *et al*. *Graphical Methods for Data Analysis*. New York: Chapman and Hall, 1983.
23. Coombes K, Tsavachidis S, Morris J, *et al*. 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* 2005;**5**(16): 4107–17.
24. Daubechies I. *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.
25. Furlanello C, Serafini M, Merler S, *et al*. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics* 2003; **4**:54.
26. Merler S, Jurman G. Terminated ramp–support vector machines: a nonparametric data dependent kernel. *Neural Networks* 2006;**19**(10):1597–611.
27. Li J, Sun Y. Iterative relief for feature weighting: algorithms, theories and applications. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press, New York, NY, 2006, 913–20.
28. Jurman G, Merler S, Barla A, *et al*. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics* 2008;**24**:258–64.
29. Paoli S, Jurman G, Albanese D, *et al*. Integrating gene expression profiling and clinical data. *Int J Approx Reason* 2008;**47**(1):58–69.
30. Furlanello C, Serafini M, Merler S, *et al*. Semisupervised learning for molecular profiling. *IEEE/ACM Trans Comput Biol Bioinform* 2005;**2**(2):110–18.
31. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. *Proteomics* 2003; **3**(9):1692–8.
32. R Development Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria 2004.
33. Coombes KR, Baggerly KA, Morris JS. Pre-processing mass spectrometry data. In: Dubitzky M, Granzow M, Berrar D, (eds). *Fundamentals of Data Mining in Genomics and Proteomics*. Boston: Kluwer, 2007, 79–99.
34. Baggerly K, Morris J, Edmonson R, *et al*. Signal in noise: evaluating reported reproducibility of serum proteomics tests for ovarian cancer. *J Natl Cancer Inst* 2005;**97**(4):307–9.
35. Wu B, Abbot T, Fisherman D, *et al*. Ovarian cancer classification based on mass spectrometry analysis of sera. *Cancer Inform* 2006;**2**:123–32.
36. Zhu W, Wang X, Rao M, *et al*. Detection of cancer-specific markers amid massive mass spectral data. *PNAS* 2003; **100**(25):14666–71.
37. Jong K, Marchiori E, Sebag M, *et al*. Feature selection in proteomic pattern data with support vector machines. In: *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE, 2004, 41–8.