Supporting information for: Pinpointing Biomarkers In Proteomic LC-MS Data By Moving Window Discriminant Analysis

Tom G. Bloemberg,[†] Hans J.C.T. Wessels,[‡,¶] Maurice van Dael,[‡] Jolein Gloerich,[‡] Lambert P. van den Heuvel,[‡,§] Lutgarde M.C. Buydens,[*,†] and Ron Wehrens[*,†,‖]

Radboud University Nijmegen, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands, Nijmegen Proteomics Facility, Department of Laboratory Medicine, Laboratory of Genetic, Endocrine and Metabolic diseases, Radboud University Nijmegen Medical Centre, Nijmegen Centre for Mitochondrial Disorders, Department of Laboratory Medicine, Laboratory of Genetic, Endocrine and Metabolic Diseases, Radboud University Nijmegen Medical Centre, and Department of Pediatrics, Radboud University Nijmegen Medical Centre

E-mail: l.buydens@science.ru.nl; ron.wehrens@iasma.it

Phone: +31-24-3653180; +39-0461-615563. Fax: +31-24-3652653; +39-0461-650872

[*] To whom correspondence should be addressed

[†] Radboud University Nijmegen - IMM

[‡] Radboud University Nijmegen Medical Center - NPF

[¶] Radboud University Nijmegen Medical Center - NCMD

[§] Radboud University Nijmegen Medical Center - Dept. Pediatrics

[‖] Current address: Centro Ricerca e Innovazione, Fondazione Edmund Mach, Via E. Mach, 1, 38010, San Michele all'Adige (TN), Italy.

LDA and PCA-LDA

LDA[1–4] is a classic multivariate classification technique whose performance still rates amongst the best[4–6]. More importantly, its results are readily interpretable, a feature that is lacking in some modern techniques like Support Vector Machines (SVMs) for instance, although recent advances have been made[7]). Two main types of LDA can be distinguished: Fisher LDA and maximum likelihood LDA. For a two-class problem, such as in the current paper, both formulations give exactly the same result. Here, we will use the formulation of Fisher[1].

Fisher LDA makes a linear combination of the variables (time points in chromatograms or *m/z*-values in mass spectra) in such a way as to maximize the ratio of the between-class variance and the within-class variance:

$$F = \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}, \tag{1}$$

with **B** and **W** the between-class and within-class covariance matrices, respectively and **a** the so-called discriminant coordinate[4] which contains the weights of the individual variables in the linear combination. Practically, performing LDA means finding the **a** (and thus the variable weights in it) that maximizes the ratio F. The value of the maximized ratio itself (the ‘Fisher quotient’, a scalar number) is a measure for the classifiability of the data: a high value of F means that the between-class variance is large compared to the within-class variance, so the two different groups of samples can easily be distinguished on the basis of the data. At the same time, the discriminant coordinate **a** is comparable to the loadings in principal component analysis (PCA) and contains the weights of the original variables in the linear combination. A variable or group of variables that has a high weight in the discriminant coordinate of a well-classifiable data matrix (i.e. with a high value of F) will be very likely to contain the specific information that leads to the good classification result, i.e. a peak caused by a potential biomarker.

Commonly, in chemical data, the number of samples n is much smaller than the number of measured variables p. This poses a problem, since the within-class covariance matrix $\mathbf{W}$ becomes singular, and LDA cannot be applied directly. Many solutions to this problem have been proposed[8,9], but a particularly popular one is the combination of PCA and LDA[5,10–13], which we will refer to as PCA-LDA here. Briefly, PCA is used to reduce the number of variables before LDA is applied, thus preventing $\mathbf{W}$ from becoming singular.

Datasets

Simulated Sets

Most relevant details about the simulated datasets are available from the main paper. Some additional details are provided here.

The elution density of the compounds in the datasets is not constant. Instead, the probability of finding a compound varies as a block function with a value of 0.7 in the range $t = 200$–600 and 0.2 outside that area, so as to qualitatively emulate the behaviour observed in real LC-MS measurements.

Peak heights were sampled from the set $\{1,2,...,500\}$ with probabilities according to a normal distribution (positive half only), adapted by adding 0.3 to all probabilities.

Apart from spikes (that were removed, see below), noise in the *E. coli* data is only present on the peaks, due to thresholding by the spectrometer software. The simulated noise was made to qualitatively resemble this by adding half of the absolute value of random standard normally distributed values only to values in the data larger than 0.01.

Apart from the random noise on the peaks, the exact simulated data described above and in the main paper can be reproduced via installing the SimSpec-package (provided as part of this Supporting Information) in R and following the instructions in the `GenerateDataset.R` script.

E. coli Benchmark Set: Sample Preparation and Data Collection

The spike-in set consists of LC-MS measurements on fifteen samples that each consist of tryptic peptides originating from 5 μg *Escherichia coli* protein homogenate. Ten samples are spike-free and five were spiked with 0.5 μg carbonic anhydrase (*Bos taurus*), prior to tryptic digestion. One sample from each class (with and without spike) was measured in duplicate, the rest was measured once, leading to a total of seventeen LC-measurements, eleven of class one (without spike) and six of class two (with spike).

E. coli K12 strain was grown on glucose medium and cells were harvested by centrifugation (5 minutes at 2000 *g*). For each sample, cells were taken up in 500 μL 8M urea 10mM Tris-HCl buffer pH 8.0 and were sonicated for 5 minutes. Cell debris was removed by centrifugation for 5 minutes at 14,000 *g*. Protein concentration was determined using the 2D Quant kit (GE Healthcare), carbonic anhydrase spike was added as required, and proteins were digested (in-solution) using trypsin as described elsewhere[14]. For each analysis 5 μg of tryptic *E. coli* peptides were extracted using stop and go elution (STAGE) tips according to[15].

Measurements were performed using an Agilent 1100 nanoflow liquid chromatograph coupled online via a nano electrospray ionization source to a 7 T linear ion trap Fourier Transform ion cyclotron resonance mass spectrometer (LTQ FT, Thermo Scientific). The acquired *m/z* range was 350–2000 Th. Samples were analyzed using multiple in-house packed columns over a period of 2 months at different days and in random order to include real life chromatographic and mass spectrometric variations.

Chromatographic separations were performed using fused silica emitters (New Objective, PicoTip® Emitter, Tip: $8 \pm 1\mu$m, ID: 100 μm, FS360-100-8-N-5-C15) that were packed in-house with reversed phase ReproSil-Pur C18AQ 3 μm resin (Dr. Maisch GmbH)[16]. Peptides were loaded directly onto the analytical column at a flow of 600 nL/min buffer A (0.5 % acetic acid). Next, a 60 minutes linear gradient was applied of 10-40 % buffer B (80 % acetonitrile, 0.5 % acetic acid) at a flow of 300 nL/min for peptide separations. All measurements were performed with intermediate blank runs to avoid carry-over effects. MS scans of *m/z* 350–2000 Th were acquired by the ICR cell at a selected resolution of $R = 1 \cdot 10^5$ using $1 \cdot 10^6$ ions and allowed for a maximum injection time of 500 ms.

E. coli Benchmark Set: Preprocessing

To make the LC-MS data amenable to analysis, the raw data were converted from the proprietary Thermo format (.RAW) to mzXML, using ReAdW version 1.1[17]. The mzXML files were subsequently imported in the statistical computation environment R[18] using the `read.mzXML` function from the **caMassClass** package[19]. Spikes in the data (having, in general, a low intensity) were removed using in-house written functions. The data were converted from a list of *m/z*-time-intensity triplets that is typical for mass spectrometric data, to matrix format. This was done by binning along the *m/z* axis, using 1801 bins of width 0.5 Th spanning the range 350–1250 Th. The relevant part of the time axis between 2000 s and 5500 s typically consisted of 1500–1600 time points and was converted to 2000 timepoints for all samples by linear interpolation. The individual 1801×2000 matrices were then combined in a $1801 \times 2000 \times 17$ array.

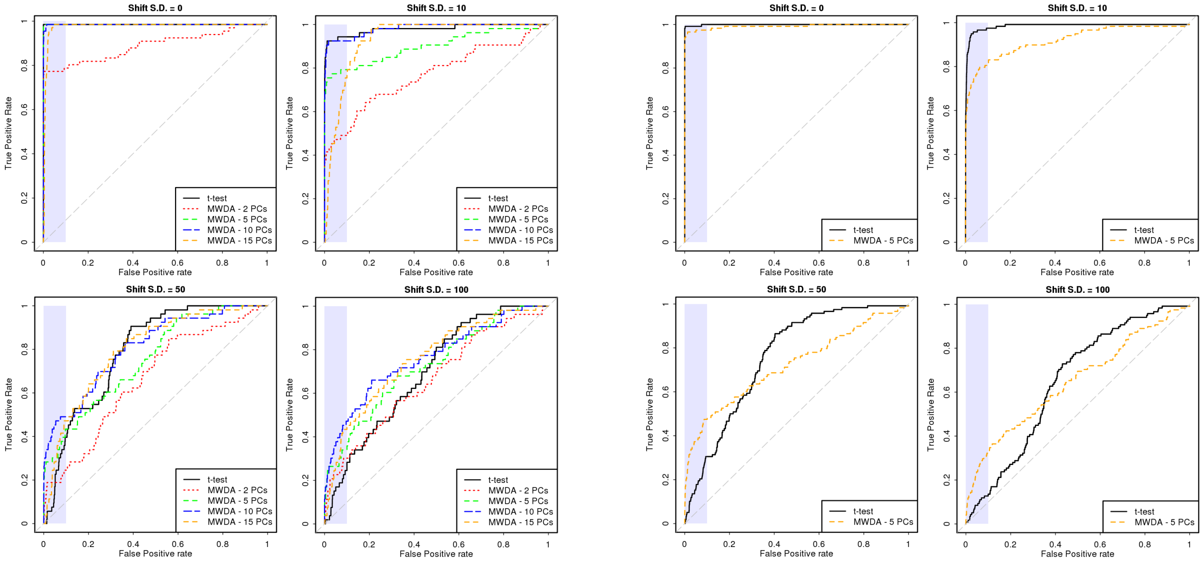
Carbonic Anhydrase: Sample Preparation and Data Collection

Carbonic anhydrase was digested (in-solution) using trypsin[14] and peptides were extracted using stop and go elution tips[15]. The tryptic carbonic anhydrase digest was first measured by NSI MS using a TriVersa NanoMate robot (Advion) coupled to a 7 T linear ion trap Fourier-Transform ion cyclotron resonance mass spectrometer (LTQ FT Ultra, Thermo Scientific). Full MS spectra were acquired by the ICR cell using 5 microscans at a resolution of $R = 1 \cdot 10^5$ using $1 \cdot 10^6$ ions. Both collision induced dissociation (CID) and electron capture dissociation (ECD) fragmentation spectra were acquired. The CID fragmentation spectra were acquired using the linear ion trap (3 microscans, $3 \cdot 10^3$ ions, 3 Th isolation width, 30% normalized collision energy, 30 ms activation time, activation $Q = 0.25$) and ECD spectra were acquired by the ICR cell (3 microscans, $1 \cdot 10^6$ ions, 3 Th isolation width, resolution $R = 1 \cdot 10^5$, 5% normalized ECD energy, 70 ms activation time, 0 ms delay). The CID spectra were only acquired for ions with chargestates 1+, 2+, 3+

whereas ECD spectra were acquired for ions detected with chargestates 2+ and higher. Additional LC-MS/MS analyses of carbonic anhydrase digest were performed using an Easy nano LC (Proxeon) coupled online with the 7 T linear ion trap FT-ICR mass spectrometer with settings for CID and ECD spectra as described above. Chromatographic conditions were identical to those mentioned previously in this manuscript. The instrument was set to run cycles that consisted of a survey Full MS scan (m/z = 350–1600 Th) which was followed by either 4 data dependent CID linear ion trap (in parallel to the survey scan) or 3 ECD FT-ICR fragmentation spectra for each separate analysis. Peaklists were generated from the raw mass spectrometer data using ExtractMSn (Thermo Scientific) and in-house developed Perl scripts. Database searches were performed using Mascot v2.2 (Matrix Science) against a curated NCBI Refseq Escherichia coli K12 database supplemented with known contaminant proteins (e.g. skin proteins, trypsin, LysC) and carbonic anhydrase II sequence. Search parameters for both CID and ECD spectra specified tryptic specificity (allowing for 1 missed cleavage) with a precursor mass tolerance of 20 ppm and "Error tolerant" search type. Specific settings for linear ion trap CID spectra included 0.8 Da mass tolerance and ESI-TRAP was set as instrument type (b & y ions). ECD spectra were searched using a mass tolerance of 0.05 Da and specified FT-ECD as instrument type (c & z ions). Peptides identified with a Mascot identification score of 20 or higher were manually validated when their chargestate, retention time (when applicable) and m/z value (tolerance 0.05Th) corresponded with selected ions (MWDA, *t*-tests, or randomized target list). Peptides identified for each selection list were used to assess the performance of both MWDA and *t*-test methodologies versus a random target list.

Results

In the main text, it is stated that the plots of the true positive rates versus the total number of LC-matrices designated positive for the simulated sets are near-identical with the complete ROC-curves. To substantiate that claim, the indicated ROC-curves are provided in Figure 1a and Figure 1b.



(a) Training sets

(b) Test sets

Figure 1: Averaged ROC-curves for the simulates training and test sets, respectively, using *t*-tests and MWDA. Averaging was performed via the merge-sorting approach discussed by Fawcett[20].

References

(1) Fisher, R. A. *Annals of Eugenics* **1936**, *7*, 179–188.

(2) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; Data Handling in Science and Technology; Elsevier, 1997; Vol. 20A.

(3) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; Data Handling in Science and Technology; Elsevier, 1998; Vol. 20B.

(4) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, First ed.; Springer, 2001.

(5) Mertens, B. J. A. *Journal of Proteomics* **2009**, *72*, 785–790.

(6) Hand, D. J. *Statistical Science* **2006**, *21*, 1–14.

(7) Krooshof, P. W. T.; Üstün, B.; Postma, G. J.; Buydens, L. M. C. *Analytical Chemistry* **2010**, *82*, 7000–7007.

(8) Jonathan, P.; McCarthy, W. V.; Roberts, A. M. I. *Journal of Chemometrics* **1996**, *10*, 189–213.

(9) Friedman, J. H. *Journal of the American Statistical Association* **1989**, *84*, 165–175.

(10) Ami, D.; Natalello, A.; Mereghetti, P.; Neri, T.; Zanoni, M.; Monti, M.; Doglia, S. M.; Redi, C. A. *Spectroscopy - An International Journal* **2010**, *24*, 89–97.

(11) Charlton, A.; Allnutt, T.; Holmes, S.; Chisholm, J.; Bean, S.; Ellis, N.; Mullineaux, P.; Oehlschlager, S. *Plant Biotechnology Journal* **2004**, *2*, 27–35.

(12) Cozzolino, D.; Smyth, H. E.; Cynkar, W.; Dambergs, R. G.; Gishen, M. *Talanta* **2005**, *68*, 382–387.

(13) Kher, A.; Mulholland, M.; Green, E.; Reedy, B. *Vibrational Spectroscopy* **2006**, *40*, 270–277.

(14) Wessels, H. J. C. T.; Gloerich, J.; der Biezen, E.; Jetten, M. S.; Kartal, B. *Methods in Enzymology* **2011**, *486*, 465–482.

(15) Rappsilber, J.; Ishihama, Y.; Mann, M. *Analytical Chemistry* **2003**, *75*, 663–670.

(16) Ishihama, Y.; Rappsilber, J.; Andersen, J. S.; Mann, M. *Journal of Chromatography* **2002**, *A979*, 233–239.

(17) http://sashimi.sourceforge.net/software_glossolalia.html#ReAdW.

(18) R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2009.

(19) Tuszynski, J. The caMassClass Package. 2007.

(20) Fawcett, T. *Pattern Recognition Letters* **2006**, *27*, 861–874.