Journal of
Proteomics & Bioinformatics

Nadarajah et al., J Proteomics Bioinform 2012, 5:4
DOI: 10.4172/jpb.1000219

Research Article Open Access

Serum Peptide Profiles of Duchenne Muscular Dystrophy (DMD) Patients Evaluated by Data Handling Strategies for High Resolution Content

Vishna D Nadarajah[1], Bart JA Mertens[2], Hans Dalebout[3], Marco R Bladergroen[3], Sharmini Alagaratnam[1], Penny Garrood[4], Kate Bushby[4], Volker Straub[4], André M Deelder[3], Johan T den Dunnen[1], Gert-Jan B van Ommen[1], Peter AC 't Hoen[1] and Yuri EM van der Burgt[3]*

[1]*Leiden University Medical Center (LUMC), Department of Human Genetics, PO Box 9600, 2300 RC, Leiden, the Netherlands*
[2]*Leiden University Medical Center (LUMC), Department of Medical Statistics, PO Box 9600, 2300 RC, Leiden, the Netherlands*
[3]*Leiden University Medical Center (LUMC), Department of Parasitology, Biomolecular Mass Spectrometry Unit, PO Box 9600, 2300 RC, Leiden, the Netherlands*
[4]*Institute of Human Genetics, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, United Kingdom*

Abstract

The accuracy of Mass Spectrometry (MS)-based analysis of peptides in complex biological mixtures improves upon using high resolution instrumentation. However, high resolution content poses challenges to data processing and statistical analysis. Here, three different data handling strategies were evaluated with respect to classification performance using a well-defined cohort of serum samples from Duchenne Muscular Dystrophy (DMD) patients and controls. For this purpose, serum samples were purified using a solid-phase extraction (SPE) protocol based on Reversed-Phase (RP) C18 magnetic beads. Isotopically-resolved peptide profiles were acquired on a Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometer and examined by either using the full mass spectrum or after selecting peaks between 1000<*m/z*<4000 followed by *data filtering* or *data integration*. To identify discriminative peptides, linear Logistic Regression Analysis (LRA) with double-cross validation was applied for each method. The data integration strategy resulted in the lowest classification error rate while use of the filtered or full profile data gave higher error rates. From this it was concluded that peak selection methods may increase the discriminative power, however with the potential downside of loss of potentially interesting peptides. Seven peptides were found by all three methods when considering the top 15 discriminating peptides. Correlation analysis of discriminative peptides showed strong associations between peptides of different *m/z*-values, suggesting that the list of discriminative peptides reflected a smaller group of proteins. Validation studies using larger patient cohorts are required for further statistical evaluation of these results.

Keywords: Mass spectrometry; Data handling; High resolution; Serum peptide profiles; Duchenne Muscular Dystrophy

Introduction

Serum peptide and protein profiling studies are widely employed in biomarker discovery studies. In MS-based clinical proteomics, peptide- or protein levels in serum of healthy and diseased individuals are mapped in a single spectrum, aiming for identifying differences [1-4]. The signature of biomarker candidates that is found through proteomics studies holds great promise for personalized medicine [5]. In this respect, it should be stressed that for implementation in a diagnostic setting, automated and standardized high-throughput workup procedures are required and rigorous analysis methods need to be developed. An interesting approach involves the application of Solid-Phase Extraction (SPE) using functionalized magnetic beads for sample workup in combination with matrix-assisted laser desorption/ionization (MALDI) MS analysis [6]. Previously, the critical parameters in this type of clinical proteomics experiments have been evaluated in detail and the necessity of standardization and Quality Control (QC) in sample collection, the workup procedure and subsequent mass analysis have been emphasized [7]. With regard to the latter aspect, recent developments and improvements in MS instrumentation have shown great benefit for profiling studies [8]. Systems have become robust, more precise and sensitive, and cover a wider *m/z*-range. Moreover, high resolution profiles, where different isotope peaks can be resolved, allow internal quality control and the determination of overlapping peptides within each spectrum [9].

Multiple data handling strategies have been reported for the processing and statistical analysis of peptide- and protein profiles, either model-based or applying different feature selection strategies. Initially, feature selection was based on simple binning procedures or finding local maxima [10-12]. Later more sophisticated methods were used like continuous wavelet transformation or Smoothed Nonlinear Energy Operator (SNEO) [13,14]. Unfortunately, in many studies features or wavelet functions did not correspond to the underlying peptide peaks, *i.e. m/z*-signals. Classification algorithms integrate feature selection and statistical evaluation and avoid the need for prior "peak picking" [15]. The comparison of full mass spectra is unbiased and inherently allows the finding of unique peaks (*i.e.* species) in a single sample. However, it also carries the burden of comparing large amounts of noisy signals present in each sample [16]. Data handling for (ultra) high resolution profiles comes with additional challenges, including the high number of data points (up to several millions in Fourier Transform Ion Cyclotron Resonance (FTICR) data [8]) per spectrum, which can make the comparison of hundreds of spectra computationally challenging. Moreover, (ultra) high resolution spectra contain multiple *m/z*-signals for each peptide. The value of incorporating information about peak shapes and isotopic distribution for feature selection of high resolution data has been acknowledged, but a thorough evaluation of possible strategies is currently lacking [17]. In this study, we will evaluate three

***Corresponding author:** Yuri EM van der Burgt, Leiden University Medical Center (LUMC), Department of Parasitology, Biomolecular Mass Spectrometry Unit, PO Box 9600, 2300 RC, Leiden, The Netherlands, E-mail: y.e.m.van_der_burgt@lumc.nl

Received January 23, 2012; **Accepted** March 12, 2012; **Published** April 30, 2012

Citation: Nadarajah VD, Mertens BJA, Dalebout H, Bladergroen MR, Alagaratnam S, et al. (2012) Serum Peptide Profiles of Duchenne Muscular Dystrophy (DMD) Patients Evaluated by Data Handling Strategies for High Resolution Content. J Proteomics Bioinform 5: 096-103. doi:10.4172/jpb.1000219

different data analysis strategies for high resolution MS-data: a model-based approach in which the full mass spectrum is considered [6,15], and two data reduction strategies. For data reduction, only parts of the spectrum containing *m/z*-signals (*e.g.* peaks) with theoretically expected isotope patterns are used for evaluation and analysis ("peak picking"), using either the raw data in these regions or the area under the curve of the peaks. In this way only the signals that correspond to peptides present in human serum profiles are considered [9,18].

The aim of this study is to develop and compare strategies for data handling of high resolution MALDI-TOF serum peptide profiles using either full mass spectral- or peak selected data. Three different data handling strategies will be evaluated with respect to classification performance using a well-defined cohort of serum samples from Duchenne Muscular Dystrophy (DMD) patients and controls. DMD is an inherited X linked recessive disease and has an incidence of one in 3500 new born boys. Current diagnostic and prognostic methods for DMD remain challenging as muscle strength assessments, biopsies and genetic tests are limited by patient variability, comfort and costs [19]. Based on the hypothesis that biomarkers leak from dystrophic muscle into the circulation blood samples have been used for discovery purposes in DMD patients [20]. The collection of blood is less invasive (and painful) compared to muscle biopsies. Serum peptide profiling in combination with high resolution MS has great potential in finding biomarkers to diagnose and monitor DMD disease progression and to determine efficacy of new therapies currently evaluated in trials [5,7,21].

Materials and Methods

Study participants and serum collection

Ten boys with a molecular diagnosis of DMD (age range, 6.6 to 9.9 years; mean 8.2 years) were recruited for a longitudinal study involving MR imaging and blood sampling at 9-month intervals for a total of 18 months [22]. As part of the study, 14 healthy male children (age range, 6.3 to 12 years; mean 8.2 years) who were attending the Royal Victoria Infirmary, Newcastle-Upon-Tyne for venipuncture for renal isotope imaging after a urinary tract infection were recruited as controls. The DMD boys' families were initially approached at their routine clinic visit and informed consent was obtained from the parents after a further explanation of the study with both parents and child. The control children were recruited on the day of their isotope imaging after discussion with both parent(s) and child. The control children did not suffer from any significant medical disorders and were on no systemic medication. Any child with an abnormal renal isotope investigation was subsequently excluded from the study. All elements of the human study were approved by the local Research Ethics Committee prior to commencement.

Blood was drawn from both groups and collected in plain glass tubes (BD-Vacutainer 367694, 13×100mm). The blood was allowed to clot at room temperature for 10 min and then placed in a refrigerator at 4°C. Then, the samples were spun at 2800g at room temperature for 10 minutes. Finally, the serum supernatant was carefully removed and transferred into 2 ml Sarstedt polypropylene tubes, and stored at -80°C until further use. At the end of the 18 months, 24 longitudinal samples were obtained from 10 DMD boys and 14 samples from a single time point from 14 control boys, *i.e.* in total 38 samples. Before performing a profile analysis, each sample was thawed and aliquoted in eight 40µL Matrix 2D barcoded storage tubes (Thermo Fisher Scientific Inc., USA) using an 8-channel Hamilton pipetting robot. These barcoded tubes were kept in 96 samples Latch Racks and frozen and stored at −80°C.

Sample processing

The workflow of the experimental setup is depicted in Figure 1. Serum sample processing consisted of a solid-phase extraction (SPE) protocol using Reversed-Phase C18-Functionalized Magnetic Beads (RPC18-MB, Invitrogen). The magnetic beads were checked before use with a standard light microscope (Dialux EB-20, Leitz, Germany) in order to evaluate dispersion and possible aggregation within the suspension. All 38 serum samples were stored in one Latch rack that was taken out of the freezer 1h before the extraction procedure on the 96-channel Hamilton STARplus® liquid handling robot was started (Hamilton, Bonaduz, Switzerland). The activation, wash and desorption steps of the beads were based on the manufacturers instructions and optimized for implementation on the 96-channel pipetting robot [9]. Briefly, 10µL of RPC18 Dynabeads (Invitrogen, Carlsbad, CA, USA) were washed four times with 50 µL of a 0.1% trifluoroacetic acid solution before mixing with 5µL of serum sample. After 5min incubation time the beads were washed three times with 25µL of a 0.1% trifluoroacetic acid solution. Then, the peptides were released from the beads with 15µL elution solution (50% acetonitrile in water) and the eluates were transferred into a 96-well plate. MALDI spotting of RPC18 eluates was randomized and performed in quadruplicates using the same 96-channel pipetting robot according to a previously reported protocol [9], thus yielding 152 MALDI-spots (4x38 samples).

Mass spectrometry

After sample workup the serum samples were mass analyzed within 12h after spotting using an UltraFlex II MALDI-TOF/TOF MS instrument (Bruker Daltonics) employing automatic acquisition of mass spectra in the positive reflectron mode (Figure 1). The spectra were acquired using FlexControl software version 3.0 (Bruker Daltonics) with identical data acquisition parameters for each sample. A SmartBeam™ 200 Hz solid-state laser, operating at a frequency of 100 Hz, was used for ionization. A profile, or summed spectrum, was obtained for each MALDI-spot by adding 20 spectra of 60 laser shots, each at different rasters. FlexControl software decided on-the-fly whether or not a scan was used for the summed spectrum. To this end, a resolution higher than 2000 was required. Peaks were detected using the SNAP centroid peak detection algorithm with signal-to-noise threshold of 1 and a "TopHat" baseline subtraction. All mass scans not fitting these criteria were excluded. The measurement of a MALDI spot was finished when 1200 laser shots had been summed in one profile. The MALDI-TOF spectra were measured from *m/z* 600 to *m/z* 4,600 and externally calibrated using a commercially available peptide mix (Bruker Daltonics). FlexAnalysis Software 3.0 (Bruker Daltonics) was used for visualization and initial data processing. All spectra were aligned using an internal calibration method with five monoisotopic peaks at *m/z* 1206.6, 1465.8, 2553.2, 2931.2, and 3261.6.

Data processing: data handling strategies for high resolution content

After MS-acquisition further data processing was performed as depicted in the second part of the workflow (Figure 1). To this end, three different data handling strategies for high resolution content were followed. In Method 1 the full peptide profile was used without any data reduction, whereas in the other two methods the number of datapoints was reduced through user-defined peak selection. The peptide peaks from the profiles were selected based on a manual inspection of all 152 mass spectra from *m/z* 1000 up to 4600. Thus

Citation: Nadarajah VD, Mertens BJA, Dalebout H, Bladergroen MR, Alagaratnam S, et al. (2012) Serum Peptide Profiles of Duchenne Muscular Dystrophy (DMD) Patients Evaluated by Data Handling Strategies for High Resolution Content. J Proteomics Bioinform 5: 096-103. doi:10.4172/jpb.1000219

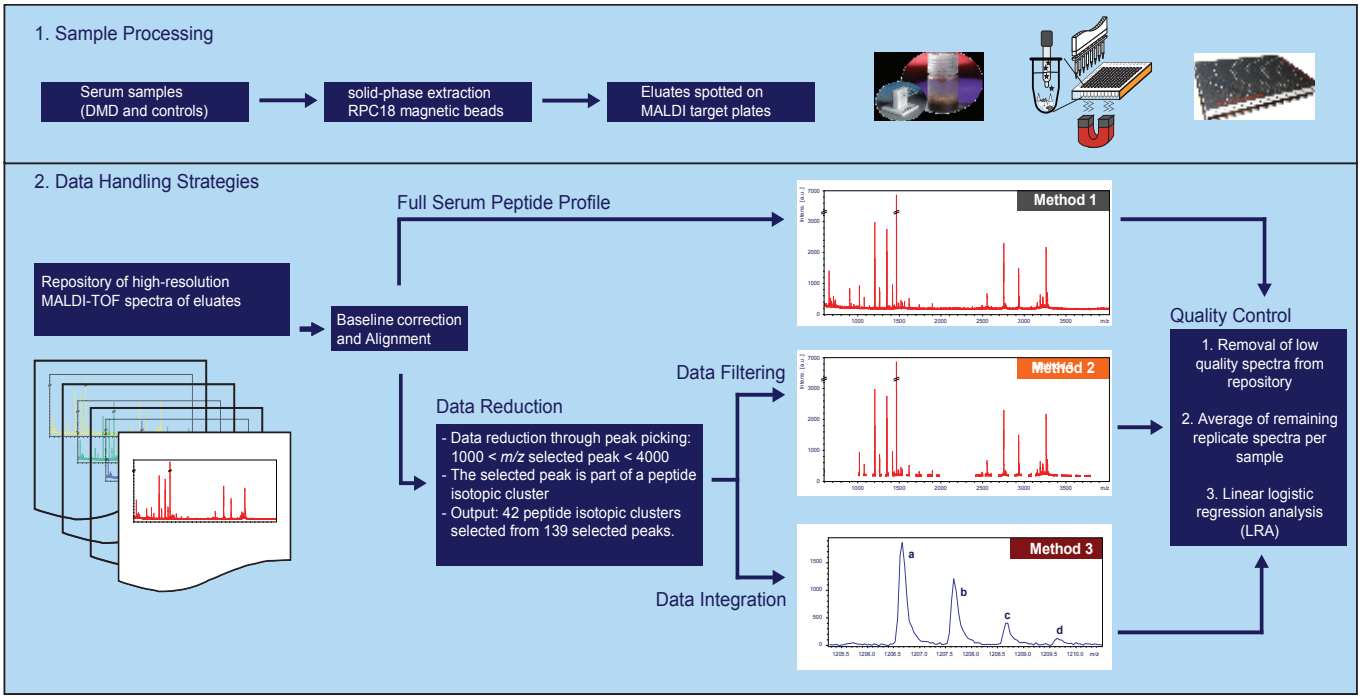

Figure 1: Data handling workflow using the three different strategies. Data handling for high resolution serum peptide profiles started with sample processing, followed by profile processing and data processing. During profile processing, the three methods for data handling were used to analyze the same high resolution serum peptide profiles. The methods: Method 1, the full serum peptide profile is analyzed; Method 2, R script to remove all datapoints except those within the 42 selected peptide isotope clusters at +0.5< m/z selected monoisotopic peak < 3.5; Method 3, all datapoints are removed except for 139 selected peaks of which the intensity of each peak is integrated and calculated using Xtractor.

selected peaks were part of an isotopic cluster that contained at least two *m/z*-signals per peptide species. Spectral signals between *m/z* 600 and 1000 were excluded for further analysis since these are not specific (*e.g.* MALDI matrix, chemical background). In summary, 42 peptide isotopic clusters were selected that contained 139 *m/z*-signals in total. Each cluster was numbered, the letter in the index referring to the isotopic peak (*e.g.* 21a is the monoisotopic peak of peptide cluster 21). Method 2 and Method 3 differ from one another in that, for Method 2, data filtering involved removing all raw datapoints except for those in the selected peptide isotope clusters at +0.5< *m/z* monoisotopic peak of each cluster < 3.5. In Method 3, the area under the 139 peaks was integrated. Data reduction was performed using an in-house developed R script for Method 2 and the previously reported tool Xtractor for Method 3 (www.ms-utils.org/Xtractor).

Double-cross validatory penalized linear logistic regression analysis (LRA)

The logistic regression model as described by Alagaratnam *et al.* [23] was applied on the MALDI-TOF data obtained using Method 1, 2 and 3 [15]. For each sample eluate, the MALDI-TOF data from the quadruplicates were averaged prior to LRA analysis. The familiar problems of overfitting and the identification associated with diagnostic model fitting with such high dimensional data (as previously described for linear discrimination) was addressed by applying a penalty to the vector of regression coefficients by assuming that these are normally distributed with mean zero (the expected value of any regression coefficient is zero) with some small variance σ^2, such that most regression parameter values will be shrunken towards zero, unless there is considerable evidence of differential spectral expression for the associated spectral bin between both groups. The reverse value $\lambda=1/\sigma^2$ is sometimes referred to as a so-called penalization constant, as greater values imply ever greater constraint on the fitted coefficients β which will eventually all be forced to zero as λ approaches infinity. Choice of σ^2 is again performed via leave-one-out cross-validation, with a secondary double-cross-validation layer to account for potential bias in the reporting of error.

Results and Discussion

Statistical data analysis and classification of peptide profiles

The peptides were obtained from human serum samples using SPE with RPC18-magnetic beads in an automated procedure and were profiled in mass spectra with peaks in the *m/z* range 1000–4000 Da. From the high resolution MALDI-TOF mass spectrometry, five peptides were selected for alignment. It was found that the four technical replicates for each of the 38 samples were highly similar and it was thus decided to average these spectra. The three methods for data handling of the serum peptide profile (Figure 1) were compared using linear logistic regression analysis (LRA) as a statistical classifier to distinguish between DMD and control samples and as a tool to generate discriminant weightings coefficients. These coefficients can be used in a subsequent exploratory analysis to identify peaks or peptide isotopic clusters that significantly differ between sample groups. As explained in the methods section, the isotopically-resolved peptide profiles were examined using either the full mass spectrum (Method 1) or after selecting peaks (Method 2, filtered data and Method 3, integrated data). The results are summarized in Table 1. Twenty-three DMD and ten control samples were correctly classified, resulting in an error rate of 13% using peak lists with individual isotopes obtained through integration. Application of filtered data or full profile data, respectively,

	METHOD 1		METHOD 2		METHOD 3	
	1	2	1	2	1	2
1=DMD	19	5	23	1	23	1
2=Control	9	5	7	7	4	10
Total	28	10	30	8	21	11
Error Rate	0.208	0.642	0.041	0.5	0.041	0.286
Total Error Rate	0.368		0.211		0.132	

Error rates: number of misclassified samples per group, DMD or control.

Total error rates: number of misclassified samples per method

Table 1: Classification table and error rates of applied strategies. The table is generated from double cross validation LRA and compares error rates and total error rates of Methods 1, 2 and 3.

yielded higher error rates of 21% and 36%. In Figures 2a, 2b and 2c the corresponding discriminant weightings plots are depicted. Here, peaks with higher weighting coefficient values contribute more to class separation and are therefore referred to as discriminant peptides. In general, the peak weightings obtained by the three different methods were in good agreement. Interestingly, the discriminant weightings plot obtained from Method 1 (see Figure 2a) reveals an identical peptide isotopic cluster as is observed in the original mass spectra (Figure 2a, inset A). From this it follows that as expected peptide isotopes contribute similarly to the class discrimination, which can be used for internal quality control of the applied data handling strategy [9]. Similar isotope clusters are also observed in the results of Methods 2 and 3. Evidently, only a subset of the selected 42 peptide isotopic clusters can be observed.

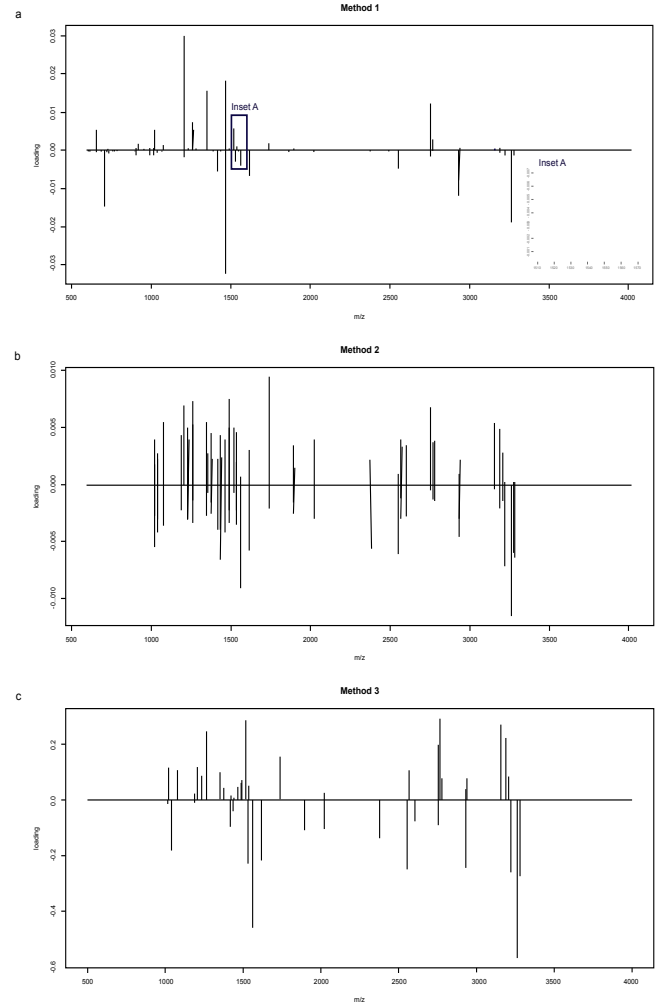

Figure 2: Discriminant Weighting Plot of Methods 1, 2 and 3. The plot shows discriminating peaks between DMD and control samples based on weighting coefficients (labeled as loading on y axis). The positive and negative weightings coefficient values observed in Figure 2a, 2b and 2c indicate whether these peptides are increased or decreased in the DMD samples compared to the controls. Peaks with negative weightings coefficients indicate higher intensities for DMD samples compared to controls, whereas peaks with positive weightings coefficients indicate higher intensities in controls compared to DMD samples

Overlapping peptides between the three strategies

In order to evaluate and compare the classification results obtained with the three strategies, the overlap between 15 peptides with the highest discriminant weightings coefficient was determined. In Table 2 the *m/z*-value are listed for the 15 discriminant peptides from each strategy. Moreover, based on peptide identification data from previous serum profiling reports the predicted protein identities are given [24-27]. Seven peptides were found with all three strategies, *i.e.* at the nominal *m/z*-values 1260, 1518, 1561, 1616, 2753, 2931 and 3261. This group likely constitutes the most robust group of discriminating peptides. The predicted protein identities show that the 15 discriminative peptides are made up of a small group of proteins. For example, with Method 1, five peptides are fragments of fibrinopeptide A, four peptides are fragments of fibrinogen and two peptides were from complement 3f. Other than the previously mentioned proteins, Inter-α-trypsin inhibitor fragments were identified with Method 2 and 3, whilst C4a was identified with Method 2. The integrated intensities of these discriminant peaks were compared between disease and control samples, as a verification step for the statistical classification. A typical example of such an analysis is shown in the box plots in Figure 3a. It is clear that the monoisotopic peaks 40a, 41a and 42a, showed significant differences in intensities between the DMD and control groups. In addition, these peaks showed strong and positive correlations in the correlation map of Figure 4, as well as with peptides 38 and 39. This could be explained by the fact that these peptides result from the same "source", *i.e.* protein. Indeed, three of these peptides originate from fibrinogen alpha-chain (Table 2). On the other hand, Figure 3b shows the box plots comparing the intensities of peptide clusters 21, 23 and 24. As expected, the marked differences in intensities between DMD and control groups for each cluster are demonstrated in all isotopes within a cluster. The same is true for other peptides in the correlation map, such as 11 and 23, or 14, 15 and 16. Moreover, as mentioned earlier for the discriminant weighting plots, strong positive correlations were observed between peaks within one specific peptide isotopic cluster (blue framed boxes in Figure 4). Additionally, negative correlations were mainly observed between peptide clusters of low *m/z*- and high *m/z*-values.

The presence of fibrinopeptide A fragments at lower *m/z*-

Discriminant peptide* peak m/z value	Monoisotopic Peak Number	Preliminary protein identification #	Method 1, 2 or 3		
3277.7	42a	Fibrinogen α [d]		2	3
3261.6	41a	Fibrinogen α [a]	1	2	3
3222.5	40a			2	3
3190.6	38a	Fibrinogen α [a]			3
3156.8	37a	Inter-α-trypsin inhibitor fragments [b]		2	3
2931.5	35a	Fibrinogen α [a]	1	2	3
2768.4	33a	Fibrinogen α [a]	1		3
2753.6	32a	Albumin precursor (25-28) [d]	1	2	3
2553.3	29a	Fibrinogen α [a]	1		3
1740.1	25a	C4a [a]		2	
1616.8	24a	Fibrinopeptide A [b]	1	2	3
1561.9	23a	C3f [a]	1	2	3
1533.2	21a		1		3
1518.8	20a	Fibrinopeptide A-H_20 [d]	1	2	3
1487.7	18a			2	
1465.8	17a	Fibrinopeptide A[a]	1		
1418.7	14a		1		
1347.6	11a	C3f [a]	1	2	
1260.6	9a		1	2	3
1206.6	6a	Fibrinopeptide A [a,c]	1	2	
1077.5	4a	Fibrinopeptide A [a]		2	
1039.6	3a				3
1020.5	2a	Fibrinopeptide A [a,c]	1		

*determined from weighting coefficients generated from LRA, for each method top 15 weighting coefficients were selected

#as determined from the publications of (a) Villanueva et al. [26] (b) Hortin et al. [24], (c) Gianazza et al. [23] and (d) Tiss et al. [25].

Table 2: Discriminant peptides* of found in different data handling strategies. The table shows the top 15 discriminant peptides* for each of the three methods, its related monoisotopic peak number, preliminary protein identification# and the overlap between the 3 methods.

values and fibrinogen α-chain fragments at the higher *m/z*-values can be rationalized as follows. Fibrinogen is partly degraded into fibrinopeptides during the clotting process. Furthermore, both fibrinopeptide A and B are trimmed by exopeptidases suggesting that the amounts and *m/z-value*s of these short peptides depend on the extent and duration of clotting of an individual specimen [25]. Other fragments of fibrinogen can also accumulate in serum depending on physiologic and serum collecting variables. Previously, it has been shown that fibrinopeptide and fibrinogen fragments can discriminate between disease and control samples for diabetes nephropathy [24], liver disease [28] and classify types and stages of cancer [27]. The entirely different set-ups of those studies make it unlikely that differences in fibrinopeptide and fibrinogen are solely due to artifacts in the sample collection procedure. It is well possible that alterations in blood coagulation are common to many diseases. In early stages of DMD, coagulation and fibrinolysis disorders are suggested to play a role in muscle degeneration through microcirculation abnormalities in muscle tissues [29]. In addition, fibrinogen has been recently reported as promoting inflammation and muscle fibrosis in the *mdx* mouse model (animal model for muscular dystrophy) through activation of the Transforming growth factor (TGF)-beta pathway [30]. Other than specific diseases, there are genetic factors that can affect blood fibrinogen levels. Single Nucleotide Polymorphisms (SNPs) have been identified in genes encoding the alpha fibrinogen chain and associated with variability in plasma fibrinogen levels [31]. Since in the current study peptide fragments of this protein have been found as discriminating features it may be necessary in future research to check for such mutations in our DMD cohort. This means that validation studies on whether fragments of fibrinogen have a role as DMD biomarkers will require both immunoassays with larger and independent cohorts of DMD serum as well as genotype data on fibrinogen SNPs.

Applicability for ultrahigh resolution profiles

Recent advances in separation methods, MS and bioinformatics, have pushed proteomics to the forefront of biomarker discovery. Still, several challenges remain, such as the detection of low abundance proteins, preparing (processing) data for the use of advanced statistics, and translating the results generated from MS-based proteomics into useful clinical knowledge [23,32]. The aim in this study was to contribute to the latter part and smoothen the pipeline for further statistical analysis. To this end, high resolution MALDI-TOF serum peptide profiles of a relatively small DMD cohort were used. Three different data handling methods were applied and the results were compared with respect to the classification performance. The key difference in the data handling was either the use of data reduction methods (Method 2, data filtering and Method 3, data integration) or analysis of full peptide profile data (Method 1). It was found that in both data reduction methods the information on the peaks within a peptide isotopic cluster was kept intact. User-defined peak selections were carried out between *m/z* 1000 and 4000 after manual evaluation

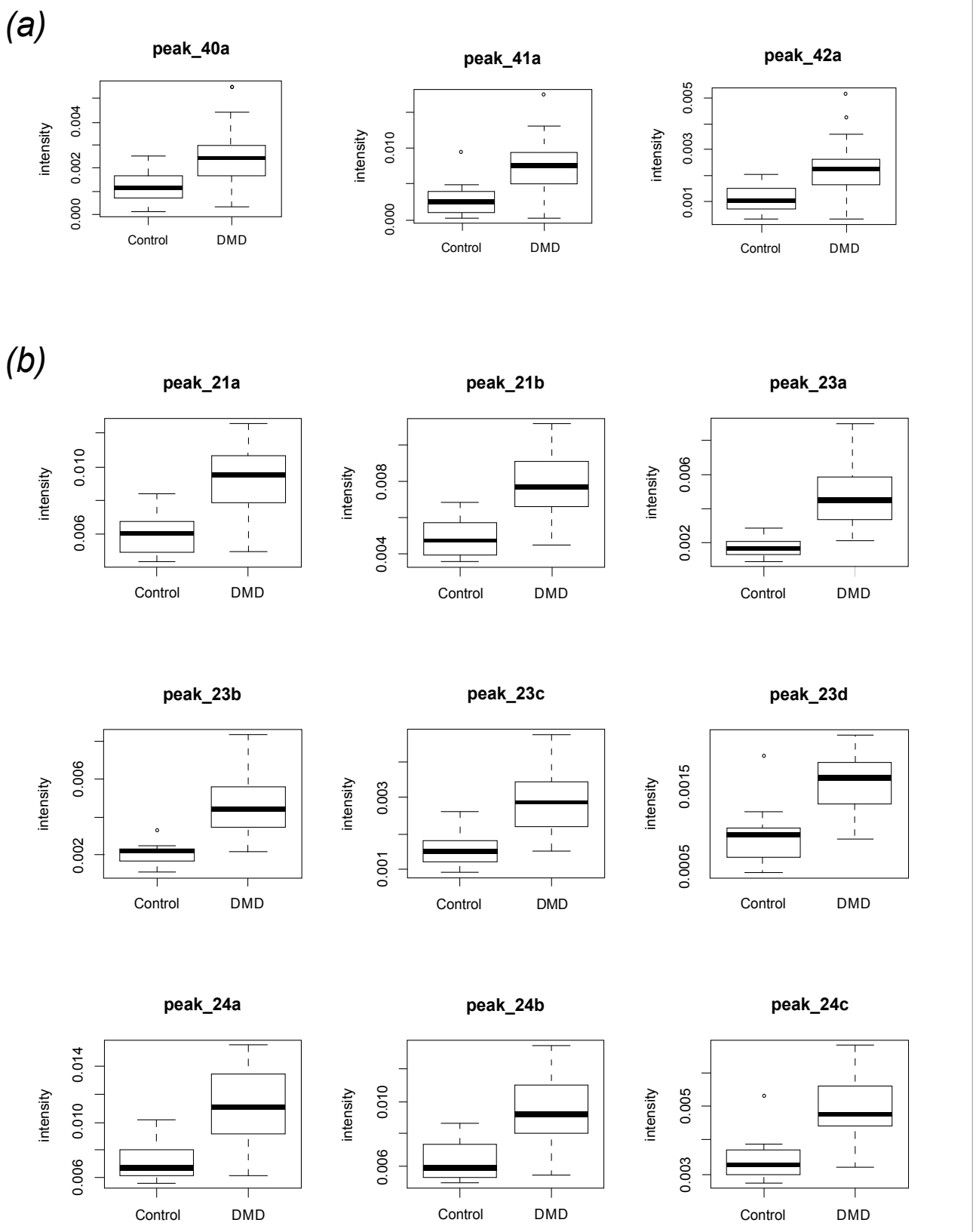

Figure 3: Comparisons of intensities between control and DMD samples. Box plots comparing the distribution of intensities for discriminant peaks. The peaks 40a, 41a and 42a were selected as examples to compare intensities in DMD and controls as they were shown to be discriminant monoisotopic peaks in Table 2a. Peaks 21a,b; 23a,b,c,d; 24a,b,c were selected as examples to show the distribution of integrated intensities of isotopes within a peptide isotopic cluster (b).

Figure 4

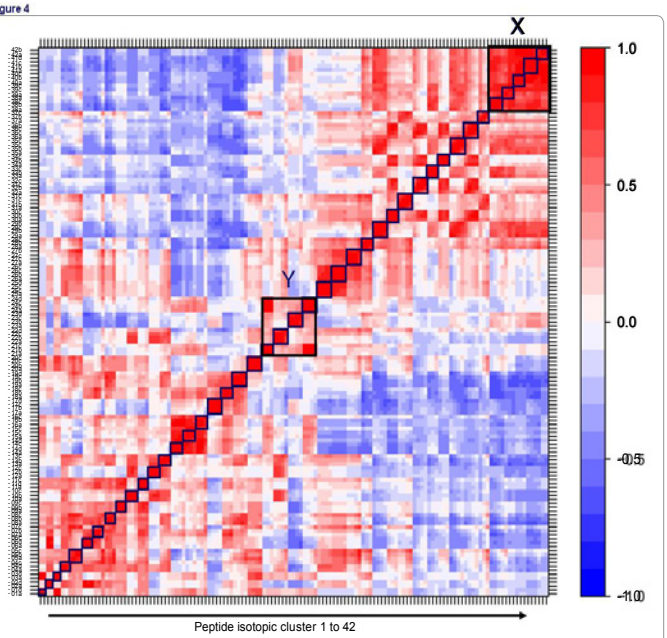

Figure 4: Correlation map of the 139 selected peaks representing 42 peptide isotopic clusters. The correlation map shows the relationship between the integrated intensities of the 42 peptide isotopic clusters across all serum profiles based on Pearson correlation. The color scale is shown on the right side of the figure, whereby positive correlation is indicated with an increased intensity for red, while negative correlation is shown with an increase for blue. We observe (i) that the peaks within a peptide isotopic cluster correlate strongly and positively, and (ii) that some discriminant peptides correlate strongly with others (examples, shown in Box X (3277.7, 3261.6, 3222.5, which are peptide clusters 40, 41 and 42) and Box Y (1533. 2 and 1616.8, peptide clusters 21 and 24).

of all *m/z*-signals in the mass spectra. Note that each selected peak was part of a peptide isotopic cluster, *i.e.* two, three or four isotopes per peptide were considered. It is these above-mentioned strategies that differentiate our study from those previously described as it applies data reduction concepts without the loss of valuable high resolution content in serum peptide profiles. Furthermore we suggest that the data handling strategies for high resolution content that we apply in this study are also applicable for current studies using ultrahigh resolution mass spectrometers [8].

In conclusion, it was shown that despite differences in analysis methods and classification errors, the lists of discriminant peptides in serum profiles from DMD patients and age-matched controls derived from the classification analysis using the three methods are similar. The differences between the three methods in resulting classification originate from different shrinkage levels between the respective calibrations across the three datasets. This also follows from the scaling of the weightings axes in Figures 2. Simply put, the regression for "Method 1" must be more heavily penalized to cope with the much higher dimensionality of the data and the absence of prior selection, for which a price is paid in prediction. Nevertheless, provided robust statistical methods are applied, profiling data can be handled with and without data reduction since the same peptides (and proteins) are found as an end result. Obviously, performing a peak selection is beneficial for data reduction purposes, however additional larger studies are required to investigate whether data reduction may either bias the process of selecting discriminant peptides or peaks or lead to sub-optimal prediction rules.

Acknowledgement

This work was supported by the Dutch Organisation for Scientific Research (Medical council ZON-MW; principal funding recipient, Annemieke Aartsma-Rus), the BIO-NMD Grant (EC,7th FP, proposal#241665; www.bio-nmd.eu; principal funding recipient, Peter-Bram 't Hoen) and the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO).

References

1. Barbarini N, Magni P (2010) Accurate peak list extraction from proteomic mass spectra for identification and profiling studies. BMC Bioinformatics 11: 518.
2. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207.
3. Diamandis EP (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 3: 367-378.
4. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, et al. (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359: 572-577.
5. Surinova S, Schiess R, Huttenhain R, Cerciello F, Wollscheid B, et al. (2011) On the development of plasma protein biomarkers. J Proteome Res 10: 5-16.
6. Villanueva J, Philip J, Entenberg D, Chaparro CA, Tanwar MK, et al. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. Anal Chem 76: 1560-1570.
7. Callesen AK, Madsen JS, Vach W, Kruse TA, Mogensen O, et al. (2009) Serum protein profiling by solid phase extraction and mass spectrometry: a future diagnostics tool? Proteomics 9: 1428-1441.
8. Nicolardi S, Palmblad M, Hensbergen PJ, Tollenaar RA, Deelder AM, et al. (2011) Precision profiling and identification of human serum peptides using Fourier transform ion cyclotron resonance mass spectrometry. Rapid Commun Mass Spectrom 25: 3457-3463.
9. Nicolardi S, Palmblad M, Dalebout H, Bladergroen M, Tollenaar RA, et al. (2010) Quality control based on isotopic distributions for high-throughput MALDI-TOF and MALDI-FTICR serum peptide profiling. J Am Soc Mass Spectrom 21: 1515-1525.
10. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, et al. (2004) High-resolution serum proteomic features for ovarian cancer detection. Endocr Relat Cancer 11: 163-178.
11. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, et al. (2005) 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 5: 4107-4117.
12. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21: 1764-1775.
13. Du P, Kibbe WA, Lin SM (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 22: 2059-2065.
14. He S, Li X, Viant MR, Yao X (2009) Profiling MS proteomics data using smoothed non-linear energy operator and Bayesian additive regression trees. Proteomics 9: 4176-4191.
15. Mertens BJA, van der Burgt YEM, Velstra B, Mesker WE, Tollenaar RAEM, et al. (2011) On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction. Stat Probab Lett 81: 759-766.
16. Noy K, Fasulo D (2007) Improved model-based, platform-independent feature extraction for mass spectrometry. Bioinformatics 23: 2528-2535.
17. Rompp A, Dekker L, Taban I, Jenster G, Boogerd W, et al. (2007) Identification of leptomeningeal metastasis-related proteins in cerebrospinal fluid of patients with breast cancer by a combination of MALDI-TOF, MALDI-FTICR and nanoLC-FTICR MS. Proteomics 7: 474-481.
18. Hilario M, Kalousis A, Pellegrini C, Muller M (2006) Processing and classification of protein mass spectra. Mass Spectrom Rev 25: 409-449.
19. Bushby K, Lochmuller H, Lynn S, Straub V (2009) Interventions for muscular dystrophy: molecular medicines entering the clinic. Lancet 374: 1849-1856.

20. Nadarajah VD, van Putten M, Chaouch A, Garrood P, Straub V, et al. (2011) Serum matrix metalloproteinase-9 (MMP-9) as a biomarker for monitoring disease progression in Duchenne muscular dystrophy (DMD). Neuromuscul Disord 21: 569-578.

21. Pietrowska M, Marczak L, Polanska J, Behrendt K, Nowicka E, et al. (2009) Mass spectrometry-based serum proteome pattern analysis in molecular diagnostics of early stage breast cancer. J Transl Med 7: 60.

22. Garrood P, Hollingsworth KG, Eagle M, Aribisala BS, Birchall D, et al. (2009) MR imaging in Duchenne muscular dystrophy: quantification of T1-weighted signal, contrast uptake, and the effects of exercise. J Magn Reson Imaging 30: 1130-1138.

23. Alagaratnam S, Mertens BJ, Dalebout JC, Deelder AM, van Ommen GJ, et al. (2008) Serum protein profiling in mice: identification of Factor XIIIa as a potential biomarker for muscular dystrophy. Proteomics 8: 1552-1563.

24. Gianazza E, Mainini V, Castoldi G, Chinello C, Zerbini G, et al. (2010) Different expression of fibrinopeptide A and related fragments in serum of type 1 diabetic patients with nephropathy. J Proteomics 73: 593-601.

25. Hortin GL (2006) The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. Clin Chem 52: 1223-1237.

26. Tiss A, Smith C, Menon U, Jacobs I, Timms JF, et al. (2010) A well-characterised peak identification list of MALDI MS profile peaks for human blood serum. Proteomics 10: 3388-3392.

27. Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, et al. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest 116: 271-284.

28. Liangpunsakul S, Lai X, Ringham HN, Crabb DW, Witzmann FA (2009) Serum Proteomic Profiles In Subjects with Heavy Alcohol Abuse. J Proteomics Bioinform 2: 236-243.

29. Saito T, Yamamoto Y, Matsumura T, Fujimura H, Shinno S (2009) Serum levels of vascular endothelial growth factor elevated in patients with muscular dystrophy. Brain Dev 31: 612-617.

30. Vidal B, Serrano AL, Tjwa M, Suelves M, Ardite E, et al. (2008) Fibrinogen drives dystrophic muscle fibrosis via a TGFbeta/alternative macrophage activation pathway. Genes Dev 22: 1747-1752.

31. Jacquemin B, Antoniades C, Nyberg F, Plana E, Mueller M, et al. (2008) Common genetic polymorphisms and haplotypes of fibrinogen alpha, beta, and gamma chains affect fibrinogen levels and the response to proinflammatory stimulation in myocardial infarction survivors: the AIRGENE study. J Am Coll Cardiol 52: 941-952.

32. Righetti PG, Boschetti E (2008) The Proteo Miner and the Forty Niners: searching for gold nuggets in the proteomic arena. Mass Spectrom Rev 27: 596-608.