Peak Aggregation as an Innovative Strategy for Improving the Predictive Power of LC–MS Metabolomic Profiles

ARTICLE *in* ANALYTICAL CHEMISTRY · JANUARY 2014

Impact Factor: 5.64 · DOI: 10.1021/ac403702p · Source: PubMed

4 AUTHORS:

Francesc Fernández
Polytechnic University of Catalonia
4 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Rafael Llorach
University of Barcelona
81 PUBLICATIONS **2,613** CITATIONS
SEE PROFILE

Cristina Andres-Lacueva
University of Barcelona
177 PUBLICATIONS **5,313** CITATIONS
SEE PROFILE

Alexandre Perera
Polytechnic University of Catalonia
79 PUBLICATIONS **455** CITATIONS
SEE PROFILE

analytical chemistry

Technical Note

pubs.acs.org/ac

Peak Aggregation as an Innovative Strategy for Improving the Predictive Power of LC-MS Metabolomic Profiles

Francesc Fernández-Albert,*[,†,‡,§] Rafael Llorach,*[,‡] Cristina Andres-Lacueva,[‡] and Alexandre Perera-Lluna[†,§]

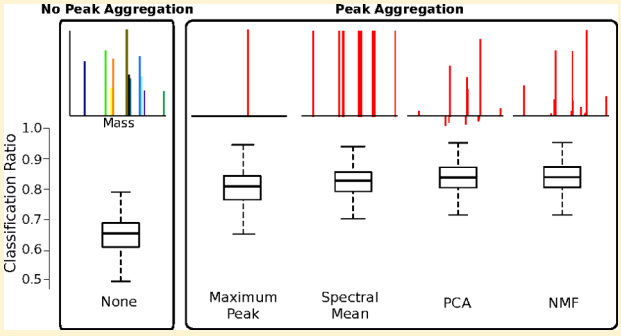
[†]B2SLab, ESAII Department, Polytechnic University of Catalonia, Barcelona, Barcelona, Spain

[‡]Biomarkers & Nutrimetabolomic Lab., Nutrition and Food Science Department, XaRTA INSA, INGENIO−CONSOLIDER Program, FUN-C-Food CSD2007-063, Campus Torribera, Pharmacy School, University of Barcelona, Barcelona, Spain

[§]CIBER in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain

Supporting Information

ABSTRACT: Liquid chromatography−mass spectrometry (LC-MS)-based metabolomic datasets consist of different features including (de)protonated molecules, fragments, adducts, and isotopes that may show high correlation values related to a high level of collinearity. There have been described several sources of these high correlation patterns regarding metabolomic datasets. Among these sources, it should be highlighted the high level of correlation computed between features coming from the same metabolite. It is well-known that soft ionization methods (such as electrospray) produce several mass features from a particular compound (i.e., metabolite spectrum). Typically, the statistical methods used in metabolomics consider spectral peaks as variables. However, it has been reported that a high collinearity between variables might be the responsible for high uncertainty values in the predictors of a regression. In this context, this technical note proposes a new strategy based on the application of the so-called peak aggregation methods (NMF Reduction, PCA Decomposition, Maximum Peak, and Spectrum Mean) to take advantage of the variable collinearity and solve the issue of high variable collinearity. A set of real samples obtained after human nutritional intervention with placebo or polyphenol-rich beverages was used to test this methodology. The results showed that applying any peak aggregation method (especially NMF and PCA) improves the statistical prediction power of class pertinence independently of the nature of the classifier (linear PLS-DA or nonlinear SVM). Overall, the introduction of this new approach resulted in a reduction of the dimensionality of the data and, in addition, in a significant increase in the overall predictive power of the data.



Liquid chromatography−mass spectrometry (LC-MS)-based metabolomic analyses produce datasets with a high level of complexity. High feature (i.e., peak) collinearity values are an important characteristic of this complexity. The origin and meaning of this collinearity (e.g., correlations) has already been revised.[1,2] Among the different types of correlations found in the LC-MS metabolomics datasets, the statistical relevance of the correlations between features coming from the same metabolite should be highlighted. In this context, it is assumed that an LC-MS metabolomics dataset consists of a mixture where, with the (de)protonated molecule, it is possible to find different features corresponding to the formation of adducts,[3] isotopes, and fragments ions coming from the ionization source[4] that show high levels of correlation. In particular, Moco et al.[5] showed that, after analysis of the correlations in an LC-MS dataset containing more than 3.000 mass signals, the highest positive correlations were found for mass signals belonging to the same metabolite.

Typically, after the LC-MS acquisition stage,[6−8] the data take the form of a matrix having mass features (treated as variables) as columns and samples as rows. This matrix is used as the basic dataset in the algorithms that process the LC-MS data. In this context, multivariable methods such as principal component analysis (PCA), partial least squares (PLS), and partial least squares discriminant analysis (PLS-DA) are widely used to extract the significant features from such datasets.[9] Some more-specific procedures—for example, the application of an orthogonal filter before the PLS (OPLS) or PLS-DA (OPLS-DA) to separate the between-class from the within-class variance[9−11]—are also commonly used. Moreover, univariate statistical tests such as *t*-tests or analysis of variance (ANOVA)[12] are also applied, obtaining a *p*-value for each mass feature. It has been reported that, when collinearity among variables is found in performing regression or discriminant analysis, this may lead to biased regressor

Received: November 14, 2013
Accepted: January 28, 2014
Published: January 28, 2014

   dx.doi.org/10.1021/ac403702p | *Anal. Chem.* 2014, 86, 2320−2325

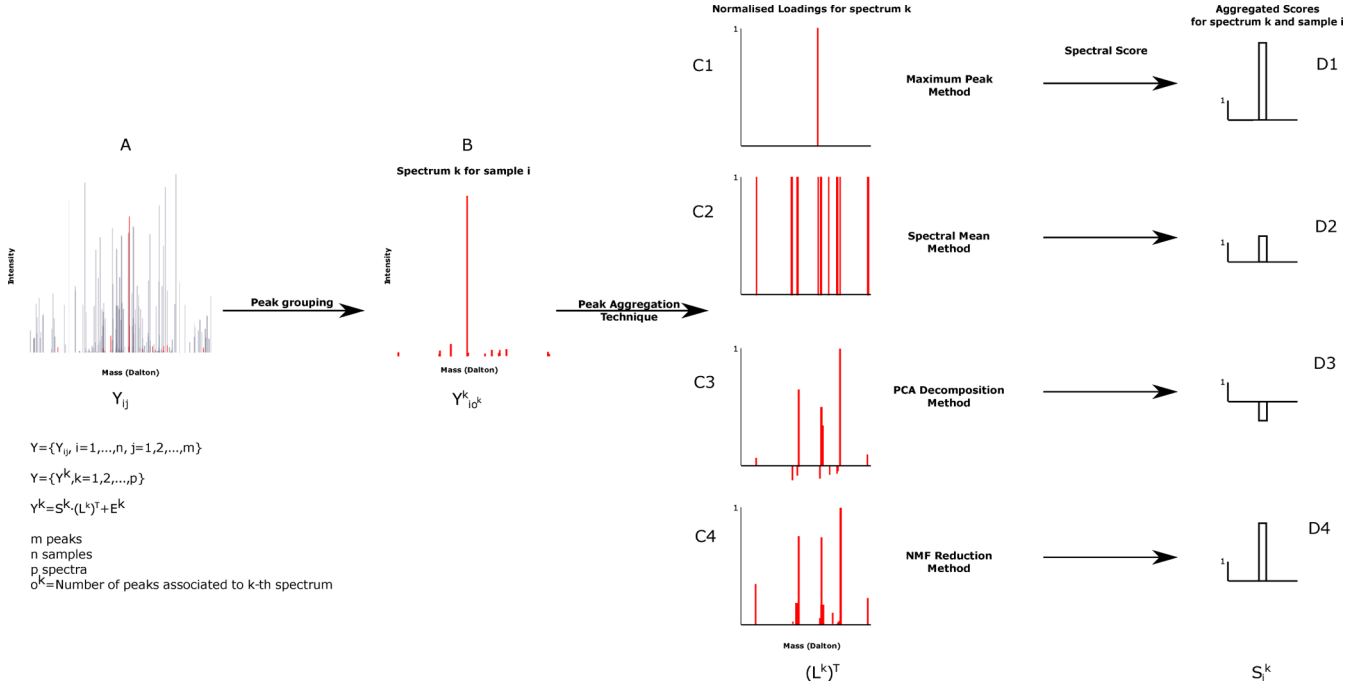


Figure 1. Sequence of plots depicting an example of the differences in using peak aggregation methods on a single spectrum. Plot A shows, in red, a spectrum for one of the samples among other peaks. Plot B depicts just this spectrum. Plots C1−C4 show the spectrum under the chosen peak aggregation technique (loadings matrix $(L^k)^T$ at eq 2), whereas Plots D1−D4 show how this spectrum is expressed across samples (scores matrix S^k at eq 2).

estimators.[13] In the case of an exact linear relationship between the variables, it is not even possible to find a unique predictor.[13] As a consequence, collinearity among variables should be controlled in order to arrive at better-fitting regressions.

The main aim of this paper is to explore the potential of shifting from single-mass features toward a mass-peak spectrum-oriented analysis. For each mass spectrum, the feature measure is obtained through peak aggregation techniques. The effect of considering this new variable measures instead of single features is evaluated by the selection of significant features using a standard ANOVA test and by obtaining the classification ratio in a classification stage by using two different classifiers: support vector machines (SVM) and PLS-DA.[14]

MATERIALS AND METHODS

Spectrum Definition. In the context of data obtained in full scan mode, the concept of spectrum is defined as the set of features/peaks including the (de)protonated molecule, isotopes, adducts, and fragments originated in the ionization source that eluted at the same retention time. Therefore, all peaks produced by the same metabolite will show large correlation properties.

Peak Aggregation Techniques. Let Y be the peak dataset matrix (with dimensions of $Y = n \times m$, where the elements of this matrix are the intensity of the peaks in the samples, n is the number of samples in the peak dataset, and m is the number of peaks in the peak dataset). Let p be the total number of spectra present in the peak dataset with $p \leq m$ (assuming that p matches the actual number of metabolites in the data). In the process of changing to a metabolite-based scheme, each peak is assigned to a metabolite as defined in the previous section. Several criteria are available to perform this assignment. In this paper, we use the correlation-based approach proposed by Kuhl et al.[15] This method relates each spectrum to a submatrix of the peak dataset known as a spectral submatrix. In general, since each spectrum may have any number of peaks, the dimensions of these spectral submatrices are different. The complete set of peaks Y from now on is considered to consist of a set of p spectral submatrices Y^k $(n \times o^k)$, each one having o^k peaks:

$$Y = \{Y^k,\ k = 1, 2, ..., p\} \tag{1}$$

Applying either multivariate techniques or statistical tests on LC-MS samples using spectra as variables is not straightforward due to the different dimensions of the spectral submatrices. Different peak aggregation techniques are studied in this technical note to reduce the dimensionality of all spectral submatrices. Each peak aggregation method is based on a multivariate process that is applied independently over each spectrum, resulting in a one-dimensional (1D)-spectral submatrix per spectrum.

For each method and for a spectrum k, the effect of applying the method over the spectral data Y^k is mathematically expressed as shown in eq 2:

$$Y^k = S^k \cdot (L^k)^T + E^k \tag{2}$$

Matrices S^k $(n \times 1)$ and L^k $(o^k \times 1)$ are method-dependent and correspond to the scores and loadings matrices for the kth spectrum, respectively. Matrix E^k $(n \times o^k)$ is the error matrix of the model. The loadings matrix can be thought of as the spectral representation obtained by the applied peak aggregation technique, whereas the scores matrix is the expression of L^k across the samples. Graphically, the interpretation of both loadings and scores matrices for each peak aggregation method used and for a certain spectrum k is depicted in Figure 1.

All the resulting p-spectral submatrices S^k can be combined to build a new dataset called a spectral dataset S $(n \times p)$.

dx.doi.org/10.1021/ac403702p | Anal. Chem. 2014, 86, 2320−2325

Statistical tests or multivariate techniques can then be applied to the spectral dataset to extract its significant features. We report results for the following peak aggregation methods: *Maximum Peak*, *Spectral Mean*, *PCA Decomposition*, and *Non-negative Matrix Factorisation Reduction* (*NMF Reduction*).

No Peak Aggregation Method. When peaks are used as variables, no peak aggregation is performed and the spectral dataset $S_{N\,one}$ equals the peak dataset Y:

$$S_{N\,one} = Y \tag{3}$$

Maximum Peak. The *Maximum Peak* method consists in taking the peak having maximum mean values across samples within the spectrum. Mathematically, it can be expressed as shown in eq 4:

$$S^k_{\max} = \left(Y^k_{lq} \middle| q = \max_j \left(\sum_{i=1}^{n} \frac{Y^k_{ij}}{n} \right), l = 1, 2, \ldots, n \right) \tag{4}$$

where $S^k_{\max}$ is an $(n \times 1)$ dimensional matrix. In some metabolomic processing algorithms, the *Maximum Peak* approach is used after building the spectra, to select the peak to be sent as a query to the database.[16] In this type of algorithm, the spectral maximum peak is chosen as the representative of the spectra and sent to the database to identify the entire metabolite.

Spectral Mean. The *Spectral Mean* is a peak aggregation method that applies a mean to the peaks of the spectrum, expressed as eq 5:

$$S^k_{\text{mean}} = \left(\sum_{j=1}^{o^k} \frac{Y^k_{ij}}{o^k} \middle| i = 1, 2, \ldots, n \right) \tag{5}$$

This method considers all peaks in a spectrum with the same weight, disregarding their statistical properties.

Principal Component Analysis Decomposition. A natural evolution from the *Spectral Mean* method is the *PCA Decomposition* method. In this peak aggregation method, a PCA is performed on every spectrum k, as shown in eq 6. This method builds an aggregated factor through a maximum variance criteria through a PCA decomposition. A data-centered PCA model is constructed for each Y^k matrix and the set of scores corresponding to the first principal component is employed as the aggregated index. In eq 6, T^k $(o^k \times 1)$ is the first principal component and P^k $(n \times 1)$ the first score vector, whereas E^k_{PCA} $(n \times o^k)$ is the error matrix.

$$Y^k = P^k \cdot (T^k)^{\mathrm{T}} + E^k_{\text{PCA}} \tag{6}$$

The spectral dataset when *PCA Decomposition* method is used is shown in eq 7:

$$S^k_{\text{PCA}} = P^k \tag{7}$$

Provided that the loadings change for each of the peaks of the spectrum, this peak aggregation method can take into account complex peak collinearity patterns between peaks of the same spectrum. However, interpretability of the output of the PCA method is not possible as, in that case, negative values are allowed in both the loadings and the scores.

Non-Negative Matrix Factorisation Reduction. An alternative decomposition is to impose non-negativity on the computation of the peak loadings corresponding to each spectrum. This can be achieved through a Non-Negative Matrix Factorisation of Y^k into the product of two matrices H^k $(n \times 1)$ and W^k $(o^k \times 1)$ plus an error matrix E^k_{NMF} $(n \times o^k)$ shown in eq 8.[17] NMF is a versatile technique that has been used in some other pattern discovery fields[18] and which may be obtained by several different mathematical criteria.[19] In the NMF method, all the matrix components of the H-matrix and the W-matrix are non-negative.[17] The matrix components were obtained by performing an optimization procedure, which consisted of a minimization of the Kullback–Leibler divergence[20] between the spectral dataset Y^k and the product $H^k \cdot (W^k)^{\mathrm{T}}$:

$$Y^k = H^k \cdot (W^k)^{\mathrm{T}} + E^k_{\text{NMF}} \tag{8}$$

When NMF is applied to a spectrum, the component capturing maximum variance of the Y^k matrix is chosen to be the 1D-spectral matrix, as shown in eq 9:

$$S^k_{\text{NMF}} = H^k \tag{9}$$

This method has the advantage of better interpretability. Since the initial data values are intensities, all components of the peak dataset X are positive. The resulting spectral dataset under the NMF method, S_{NMF}, is a matrix all of whose components are positive. The components of the spectral dataset $S_{ij;\text{NMF}}$ can be understood as the expression of a certain spectrum j in the sample i. As the W-matrix represents a certain averaged spectrum, the components $S_{ij;\text{NMF}}$ can be understood as a type of spectral intensity. The R source code of the methods will be available through the Bioconductor repository.

EXPERIMENTAL DATA

Experimental Design. A randomized, crossover, placebo-controlled, double-blind intervention study was performed with 30 volunteers, who consumed 187 mL of a control placebo (class A) or a functional beverage (FB) (class B) in an acute study, and twice a day during 15 days for a chronic consumption study (15 days placebo samples were labeled as class C and 15 days FB samples as class D). Twenty-four-hour (24 h) urine samples were collected the day before the acute intervention study and on the last day of the chronic study. For the acute study, the urine samples were collected in the first 4 h. The study protocol was approved by the Ethics Committee of the University of Barcelona.

Urine analysis. The samples were analyzed by liquid chromatography coupled with a hybrid quadrupole time-of-flight (LC-q-TOF, AB/MDS Sciex) in positive mode using the protocol proposed by Tulipani et al.[10] LC was performed in HPLC Agilent using a RP 18 Luna column (50 × 2.0 mm, 5 μm), with a sample injection volume of 15 μL. A linear gradient elution was performed consisting of [A] Milli-Q water 0.1% HCOOH (v/v) and [B] acetonitrile 0.1% HCOOH (v/v). The gradient elution (v/v) of [B] was: (time, min; B, %): (0, 1), (4, 20), (6, 95), (7.5, 95), (8, 1), (12, 1). Q-TOF spray parameters were set as previously described[21,22] and full data acquisition was performed scanning from 70 to 700 m/z. The TOF was calibrated with reserpine (1 pmol/μL). LC-MS data were acquired in random order to avoid possible bias. Finally, the files were translated to the netCDF format.

Statistical Validation. In order to evaluate the quality of the significant features found by all five methods, a cross-validation classification step was performed over the spectral dataset S, using only the significant features (see section S2 in the Supporting Information) and five training samples for each class in a repeated random subsampling cross-validation stage. For each of the five methods (the standard method and the

dx.doi.org/10.1021/ac403702p | *Anal. Chem.* 2014, 86, 2320−2325

four peak aggregated methods), a total of 300 repetitions was performed. In each repetition, the training set of five samples/class was chosen randomly between all the samples. Two different classifiers were used: PLS-DA, which is a linear classifier, and SVM which is a nonlinear one.[14] (Further details are provided as section S2 in the Supporting Information.) In order to selectively check the effect of the peak aggregation strategy, only spectra having more than one peak are considered valid in this analysis.

RESULTS

Effect of the Peak Aggregation. Figure 2 shows a typical pattern for correlation values between peaks in LC-MS datasets (see section S1 in the Supporting Information). This figure shows that there exist peak blocks having high correlation values. The peaks of these highly correlated peak blocks showing similar retention times are likely to come from the same metabolite and are considered to constitute a spectrum.

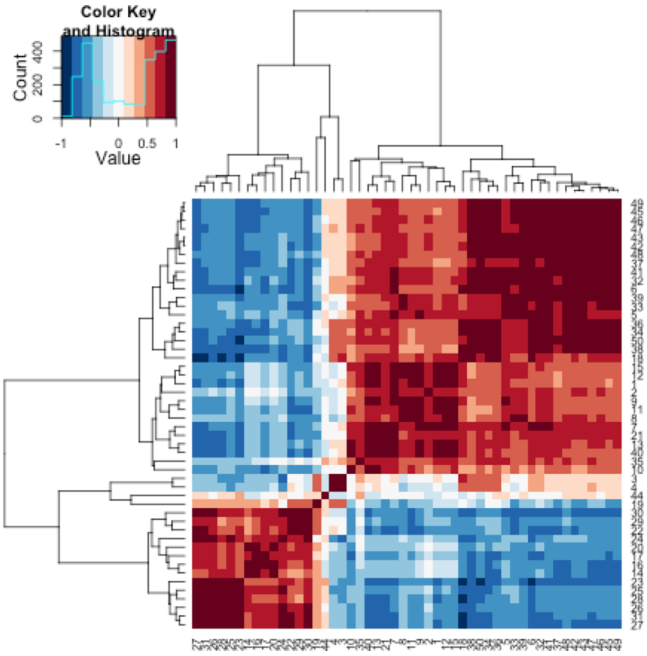


Figure 2. Heatmap showing the correlation values profile for a subset of 50 peaks in the metabolomic LC-MS sample set. Both dendrograms were performed using correlation distance. The figure depicts a typical pattern for correlation values between peaks in LC-MS datasets (see section S1 in the Supporting Information) showing that there exist peak blocks having high correlation values. It is possible to conclude that those peaks showing high correlation and similar retention time (because they were produced after chromatography) came from the same metabolite.

In order to illustrate in more detail the effect of using one or another peak aggregation method on the spectra, Figure 1 shows how a sample spectrum is processed, depending on which method is applied. Using different peak aggregation techniques implies having a different spectral pattern (matrix L^k at eq 2) and a different expression of this spectrum across samples (cf. matrix S^k at eq 2). In Figure 1, picture B corresponds to the original spectrum for only one sample, whereas the C plots of the figure show the result of applying peak aggregation methods to a single spectrum. The D plots correspond to the aggregated value for the sample spectrum and the analyzed sample.

From Figure 1, it can be seen that the *Maximum Peak* method (plots C1 and D1) is a restrictive one, because only the maximum peak of each spectrum is considered. In this case, the expression of the spectrum over samples is the expression of the maximum peak over samples. The *Spectral Mean* method (plots C2 and D2) enforces all the peaks of the spectrum to have equal weight. The main difference between using *PCA Decomposition* (plots C3 and D3) and *NMF Reduction* (plots C4 and D4) as opposed to using the *Maximum Peak* or the *Spectral Mean* is that the weight of each peak in the final spectrum value depends not only on the peaks, but also on the expression of those peaks over the samples. The C3 and D3 plots of Figure 1 show that the *PCA Decomposition* allows for negative values for both the spectrum pattern and for the expression of the spectrum over samples

Finally, plots C4 and D4 show the effect of applying the *NMF Reduction* on the sample spectrum. All the values in both the spectrum pattern $(L^k)^T$ and the expression of the spectrum over samples S^k are positive. This gives better interpretability of the spectral pattern, which can now be understood as weighted intensity measures over samples of the spectrum. In consequence, the C4 plot would be the spectrum where the weights have taken into account the different peak patterns across samples and the D4 plot would be the expression of this spectrum on sample *i*.

Class Prediction Results. After applying a peak aggregation method, the variables are now the metabolite spectra L^k rather than the peaks themselves. In consequence, the peak correlation patterns due to the peaks coming from the same metabolite are no longer present in the data and hidden

Table 1. Table Showing the Output Parameters of the Fisher's LSD Test for the PLS-DA and SVM Classifiers after 300 Classification Steps[a]

	PLS-DA				SVM			
method	mean value	std error	LCL	HCL	mean value	std error	LCL	HCL
mean[b]	0.803	0.003	0.797	0.809	0.804	0.003	0.798	0.809
NMF[c]	0.817	0.003	0.812	0.824	0.820	0.003	0.814	0.825
none	0.636	0.004	0.629	0.643	0.652	0.003	0.646	0.659
PCA[c]	0.821	0.003	0.814	0.827	0.817	0.003	0.812	0.822
maximum[b]	0.795	0.003	0.789	0.801	0.790	0.003	0.784	0.797

[a]The method labelled "none" corresponds to the standard method in which no peak aggregation is performed. The least significant difference was found to be 0.013 for PLS-DA and 0.012 for SVM. Mean differences lower than this value cannot be discerned. The labels LCL and HCL in the columns correspond to lower confidence limit and higher confidence limit, respectively. [b]Fisher's LSD test only found them to be equal in the PLS-DA classifier. [c]Fisher's LSD test found them to be equal, regardless of the classifier used.

dx.doi.org/10.1021/ac403702p | Anal. Chem. 2014, 86, 2320−2325

correlation patterns then emerge. To test if the spectral data set *S* improves the predictive power of the data, the metabolomic LC-MS sample files were processed following the steps described in section S2 in the Supporting Information, after which the classification procedure explained in the section "Statistical Validation" was applied.

The results of performing the 300 classification steps and their associated Fisher's Least Significant Difference (LSD) parameters for each method and both classifiers are shown in Table 1 (also see Figures S-4 and S-5 in the Supporting Information). Irrespective of the classifier used, clear tendencies emerge when peak aggregation methods are used. The results, using both classifiers, show that applying peak aggregation methods where spectra are used as variables improves the predictive performance compared to the standard methods, where peaks are used. This improvement is at least equal to 14% and as much as 18% in the mean value of the classification ratio, depending on the method used (see Table 1). *PCA Decomposition* and *NMF Reduction* showed the best performance and were indistinguishable in both classifiers according to the Fisher's LSD results. The main difference between using the PLS-DA and the SVM classifiers was that, when PLS-DA was used, there was no difference between applying the *Spectral Mean* or the *Maximum Peak* as peak aggregation methods, whereas when SVM is used, the *Spectral Mean* showed better performance than the *Maximum Peak* method.

To give a deeper insight to the improvement of the overall classification ratio, Figure 3 shows the classification ratios for each of the four classes, using either *NMF Reduction* as a peak aggregation technique or not using any method. The results show improvements in the classification ratios for all four classes when *NMF Reduction* is used. This suggests that using *NMF Reduction* as a peak aggregation method improves the overall statistical power of the data, regardless of the similarity between classes.

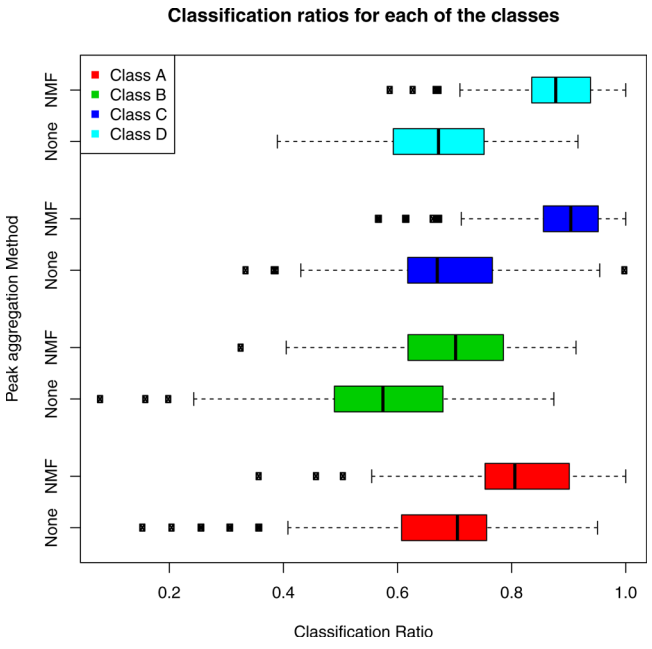


Figure 3. Comparative boxplot showing the different classification ratios for the four classes of samples, using either NMF as a peak aggregation technique or using no peak aggregation technique. The classifier used was SVM.

CONCLUSIONS

This paper has studied the effect of performing peak aggregated measures in the statistical analysis of metabolomic LC-MS urine samples. Using peak aggregation techniques implies a change from a single mass features to a metabolite spectrum oriented analysis. Different peak aggregation techniques, which imply different spectral datasets, have been compared to each other and also to the non-aggregated standard analysis. The results showed that using peak aggregation methods in the statistical analysis improves the statistical power of the LC-MS data, independent of whether the classifier is linear (PLS-DA) or nonlinear (SVM). Considering the classification ratio as the quality metrics for both classifiers, it was shown that using *NMF Reduction* or a *PCA Decomposition* methods over each spectrum are the processes that most improve the detection of significant features.

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: francesc.fernandez.albert@upc.edu (F.F.-A.).
*E-mail: rafallorach@ub.edu (R. Ll.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by Spanish national grants (Nos. AGL2009-13906-C02-01/ALI and AGL2010-10084-E), and the CONSOLIDER INGENIO 2010 Programme, FUN-C-FOOD (CSD2007-063) from the MICINN, as well as Merck Serono Research Grants 2010 (Fundación Salud 2000) and RyC-2010- 07334. R.Ll. and A.P.-L. thank MICINN and The European Social Funds for their financial contribution to the Ramón y Cajal contract (Ramon y Cajal Programme, MICINN-RYC). This work has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, through Grant Nos. TEC2010-20886-C02-02 and TEC2010-20886-C02-01. A.P.-L. is part of the 2009SGR-1395 consolidated research group of the Generalitat de Catalunya, Spain. CIBER-BBN is an initiative of the Spanish ISCIII. F.F-A. especially thanks EVALXARTA-UB and Agència de Gestió d'Ajuts Universitaris I de Recerca, AGAUR (Generalitat de Catalunya) for his contract.

REFERENCES

(1) Camacho, D.; Fuente, A.; Mendes, P. *Metabolomics* **2005**, *1*, 53−63.
(2) Steuer, R. *Briefings Bioinf.* **2006**, *7*, 151−158.
(3) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. *J. Chromatogr., B* **2008**, *871*, 143−163.
(4) Plumb, R. S.; Johnson, K. A.; Rainville, P.; Smith, B. W.; Wilson, I. D.; Castro-Perez, J. M.; Nicholson, J. K. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1989−1994.
(5) Moco, S.; Forshed, J.; Vos, R. C. H.; Bino, R. J.; Vervoort, J. *Metabolomics* **2008**, *4*, 202−215.
(6) Katajamaa, M.; Orešič, M. *BMC Bioinf.* **2005**, *6*, 179.
(7) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(8) Danielsson, R.; Bylund, D.; Markides, K. E. *Anal. Chim. Acta* **2002**, *454*, 167−184.
(9) Trygg, J.; Holmes, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469−479.
(10) Tulipani, S.; Llorach, R.; Jáuregui, O.; López-Uriarte, P.; Garcia-Aloy, M.; Bullo, M.; Salas-Salvadó, J.; Andrés-Lacueva, C. *J. Proteome Res.* **2011**, *10* (11), 5047−5058.
(11) Yamamoto, H.; Yamaji, H.; Abe, Y.; Harada, K.; Waluyo, D.; Fukusaki, E.; Kondo, A.; Ohno, H.; Fukuda, H. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 136−142.
(12) Hui-Ming, L.; Selley, J. E.; Nuala, A. H.; Ferguson, L. R.; Rowan, D. D. *J. Proteome Res.* **2009**, *8*, 2045−2057.
(13) Ns, T.; Mevik, B.-H. *J. Chemom.* **2001**, *15*, 413−426.
(14) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*, Corrected Edition; Springer: New York, 2003.
(15) Kuhl, C.; Tautenhahn, R.; Boettcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.
(16) Tikunov, Y. M.; Laptenok, S.; Hall, R. D.; Bovy, A.; Vos, R. C. H. *Metabolomics* **2011**, 1−5.
(17) Lee, D. D.; Seung, H. S. *Nature* **1999**, *401*, 788−791.
(18) Brunet, J.-P.; Tamayo, P.; Golub, T. R.; Mesirov, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4164−4169.
(19) Berry, M.; Browne, M.; Langville, A.; Pauca, V.; Plemmons, R. *Comput. Stat. Data Anal.* **2007**, *52*, 155−173.
(20) Kullback, S.; Leibler, R. A. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(21) Llorach, R.; Urpi-Sarda, M.; Jauregui, O.; Monagas, M.; Andres-Lacueva, C. *J. Proteome Res.* **2009**, *8*, 5060−5068.
(22) Llorach, R.; Garrido, I.; Monagas, M.; Urpi-Sarda, M.; Tulipani, S.; Bartolome, B.; Andres-Lacueva, C. *J. Proteome Res.* **2010**, *9*, 5859−5867.