Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

# High-Throughput Metabolomics by 1D NMR

*Alessia Vignoli[+], Veronica Ghini[+], Gaia Meoni[+], Cristina Licari, Panteleimon G. Takis, Leonardo Tenori, Paola Turano, and Claudio Luchinat**

anie_201804736_sm_miscellaneous_information.pdf

*Table S1*. Summary of normalization methods for metabolomics.[a]

| Category | Name | Main Features |
|---|---|---|
| Physiological | Urine output (UO) | Concentrations are multiplied by the volume of urine excreted per hour per kilogram of mass |
| | Osmolality (OSM) | Concentrations are divided by the osmolality of each sample. It reflects physiological mechanisms but requires intensive lab work |
| | Internal Standard (IS) | Concentrations are divided by the concentration of physiologically motivated internal standard (i.e. creatinine for urines) |
| Analytical | External standard[1] (ES) | Intensities are divided by the signal intensity of a proper external standard of known concentration |
| Numerical (developed or mainly used for metabolomics) | Total area (TA) | A binned spectrum (or a list of concentrations) is divided by the total sum of all spectral bins (or of all concentrations). Different variations exist depending on whether specific spectral regions (or specific metabolites) are excluded in the total sum calculation (e.g. urea region, or lactate concentration) |
| | Total Vector Length[2] (TVL) | The procedure is the same of TA, but instead of dividing by total sum (1-norm), the total vector length (2-norm) is used |
| | Probabilistic Quotient Normalization (PQN) | A reference spectrum is calculated (e.g. the median spectrum). For each spectrum, each bin is divided by the corresponding bin in the reference spectrum, and the median of all these quotients is taken as the normalization factor. It outperforms other normalization methods in different comparisons |
| | EigenMS (EMS) | Scaling factors are estimated via singular value decomposition of the residuals matrix calculated by an ANOVA model. Developed for MS, never tested for NMR |
| | Multiplicative Scatter Corection[3] (MSC) | Developed and mainly used for near and mid infrared spectra. Applicable also to NMR |
| | Normalization to Noise (Amix Manual, Bruker) (NN) | The mean intensity of the noise is taken as a normalization parameter in NMR spectra. It corrects only for scale factors due to technical reasons |
| Numerical (originally developed for transcriptomics) | Cyclic loess normalization (CLN) | A normalization curve is fitted using non-linear local regression (loess) |
| | Contrast normalization (CN) | Data are firstly mapped in a contrast space, then normalizing curves are fitted, using a robust distance measure based on the Euclidean norm |
| | Quantile normalization (QN) | Data (bins or metabolites) are converted in quantiles, in a way that the set of intensities become the same set of values in all the samples, however, distributed differently |
| | Linear baseline normalization (LBN) | The scaling factor is computed for each spectrum as the ratio of the mean intensity of the baseline to the mean intensity of the spectrum |
| | Non-linear baseline normalization (N-LBN) | The scaling factor is computed by fitting a non-linear normalization curve that map a spectrum to the baseline |

| | | |
|---|---|---|
| | Cubic-spline normalization (CSN) | Normalization curves are computed using robust cubic splines built on quantiles. |
| | Shapiro–Wilk (SW) | Metabolites showing high variability in concentration are iteratively removed, and only low-variability metabolites are used as references for data normalization |
| | Linear mixed (LM) | A mixed model is fitted to metabolite concentration to estimate the correlation matrix and scaling factors |
| | Variance stabilization normalization (VSN) | VSN methods are set of non-linear methods that keep the variance constant over the entire data range, leading to roughly equal variable variance. Found to work well with NMR data |
| Compositional data analysis | Centered log-ratio (CLR) | The set of data (concentrations or bins) is divided by the geometric mean and the logarithms of the ratios are taken |
| | Additive log-ratio (ALR) | The set of data (concentrations or bins) is divided by an arbitrary reference and the logarithms of the ratios are taken |
| | Isometric log-ratio (ILR) | The set of data (concentrations or bins) is isometrically transformed using an appropriate orthonormal basis |
| | Pairwise log-ratios (PLR) | All pairwise log-ratios among all variables (bins or concentrations) are calculated and used in place of the original variables. Reported as the best performing log-ratio method for metabolomic data, comparable with PQN |

[a] The table is mainly adapted from,[4] and it includes information taken from [5] and [6]additional methods from.[7] See[4–6] for references.

*Table S2*. Summary of the main multivariate statistical techniques used in metabolomics.

| Category | | Name | Main Features | Remarks | References |
|---|---|---|---|---|---|
| Unsupervised Methods | Projection | Principal Component Analysis (PCA) | Builds new variables (principal components) from linear combinations of the original ones that are orthogonal each other, and maximize the variations in the samples, in a way that few PCs are the most accurate representation of the original data. | For metabolomic data PCA is mostly used as an exploratory technique for visualization, outlier detection, and data reduction. | [8–11] |
| | | Independent Component Analysis (ICA) | Decomposes a multivariate signal into original independent components. | For metabolomic spectral data ICA attempts to extracts from the spectra the signals of the individual metabolites. | [12,13] |
| | | Multilevel Component Analysis (MLCA) | Component analysis of multilevel data | It is useful for the exploratory analysis of metabolomic data obtained by repeated sampling of the same individuals. | [14] |
| | | Simultaneous Component Analysis (SCA) | Extension of PCA for simultaneous analysis of variables observed in several populations or in different occasions. Several extensions of this procedure, including a multilevel version (MSCA), are available. | Useful for the analysis of metabolomics data collected in different cohorts or with different experimental conditions. | [15,16] |
| | | Group-Wise Principal Component Analysis (GPCA) | A sparse version of PCA that use clusters of variables. For each calculated component, only the variables in the same cluster have non-zero loadings. | At variance with PCA, in GPCA each loading ideally contains only signals of biologically correlated metabolites. | [17] |
| | Clustering | *K*-Means (KM) | This method groups objects on the basis of their distances. | Can be used to group individuals based on the similarity of their metabolic profiles. | [18–20] |
| | | Partition Around Medoids (PAM) | A clustering algorithm related to K-means with improved robustness to noise and outliers. | | [21] |
| | | Spectral Clustering (SC) | A clustering technique based on the graph theory that exploit the eigen decomposition of the graph Laplacian. | | [22,23] |
| | | Hierarchical Clustering (HC) | A family of algorithms that groups data by creating a hierarchy of clusters organized in a tree (dendrogram). | Useful to obtain a visual representation of how different metabolic profiles are related. | [24,25] |

| | | | | | |
|---|---|---|---|---|---|
| | | Knowledge Discovery by Accuracy Maximization (KODAMA) | An unsupervised and semi supervised learning algorithm for feature extraction from noisy and high-dimensional data, driven, driven by an integrated procedure of cross-validation of the results. | Particularly effective with metabolic data. | [26,27] |
| | Neural Networks | Self-Organizing Map (SOM) | A neural network that produces a distribution of input data using a regular grid such that topological relations are preserved. | SOM transform metabolomics data into a visually interpretable map that captures inherent relationships among metabolites. | [28,29] |
| | | Autoencoder (AE) | An artificial neural network that learn a representation (encoding) of the input data for the purpose of dimensionality reduction. | Can be used like PCA but it can learn both linear and nonlinear transformations. | [30,31] |
| Supervised Methods | Projection | Linear Discriminant Analysis (LDA) | Finds projections that simultaneously maximize the between groups variance and minimize the within groups variance. | Fails when the number of variables exceed the number of samples, that is a common feature of metabolomics datasets. However, can be applied after a data reduction (e.g. PCA) step. | [32,33] |
| | | Partial Least Squares (PLS) | PLS is similar to PCA, but instead of maximizing the variance of the data, it maximizes the covariance between the data and the response variable. Many different variants of the original procedure exist. | A fundamental and ubiquitous method in metabolomics. Used both for regression and classification. | [34–37] |
| | | Orthogonal Partial Least Squares (OPLS) | The extracted components are separated into “predictive”, i.e. related to the target variable, and “orthogonal”, i.e. uncorrelated with the target variable. | The most common variant of PLS used in metabolomics due to its improved interpretability. | [38] |
| | | Analysis of Variance Simultaneous Component Analysis (ASCA) | A direct generalization of the univariate analysis of variance for multivariate case. | It was designed explicitly with metabolomics in mind. It can model different experimental designs. | [39] |
| | | Multilevel Partial Least Squares (MPLS) | A modified PLS to extend the univariate paired t-test to multivariate data. | MPLS find systematic variations in metabolic profiles in paired experiments, for example after a drug or nutritional challenge. | [40,41] |
| | | Group-Wise Partial Least Squares (GPLS) | The PLS adaptation of GPCA. It is an efficient sparse version of PLS | GPLS improve the interpretability of the model helping to find meaningful | [42] |

| | | | | | |
|---|---|---|---|---|---|
| | | | were loading vectors are defined by using only separated groups of correlated variables. | biologically connected clusters of metabolites. | |
| | Machine Learning | *K*-Nearest Neighbours (*K*-NN) | One of the simplest classification algorithms. It classifies an unknow instance depending on the class of the majority of its nearest neighbours. | Classification techniques that can be used to classify a unknown sample, given a train set of metabolomics data. | [25,43] |
| | | Support Vector Machines (SVMs) | An SVM model maps the input data into a high- or infinite-dimensional space so that the different groups are separated by a gap that is as wide as possible. Unknown data are then mapped into the same space and are predicted to belong to a group based on which side of the gap they fall. SVMs can efficiently perform a non-linear classification because items that are not linearly separable in the actual space became separable in the transformed space. | | [44,45] |
| | | Boosting (BO) | A general family of very successful algorithms that use an ensemble of weak classifier to build a final strong classifier. | | [46,47] |
| | | Random Forest (RF) | Uses data from the training set to build an ensemble of uncorrelated decision trees. Each tree is build using only a random subset of both the data items and of the variables. The final classifier is obtained by pooling the decisions of each tree in the forest. | Although RF is not commonly used in metabolomics, it has several benefits: i) it compares in accuracy to SVM, ii) it generates an estimate of the error of the model, iii) it computes proximities between pairs of samples that can be used for visualization and clustering. | [48,49] |
| | Neural Networks | Multilayer Perceptron (MLP) | A kind of feedforward neural network made by layers of nodes (neurons) that can learn bot linear and non-linear classification problems. | Seldom used in metabolomics but potentially very effective to classify, integrate and model different kind of data (e.g. metabolomics data, clinical data, demographical data). Usually these approaches require large amount of experimental data. | [50,51] |
| | | Deep Learning (DL) | A class of neural networks algorithms that use a cascade of multiple layers of nonlinear processing units for feature extraction and classification. | | [31,52] |

*Table S3.* Summary of the main univariate statistical tests for comparing different groups used in metabolomics studies*.

| Number of Groups | Type of Groups | Category | Name of the test | Main Features | References |
|---|---|---|---|---|---|
| Population | Independent | Parametric | One-sample *t*-test | Tests to assess whether the mean (median) of a normally distributed population (given a sample) has the hypothesized value. | [53] |
| | | Non-Parametric | One sample median test | | [54] |
| 2 | Independent | Parametric | Student's *t*-test | Used to determine whether two independent samples were selected from populations having the same distribution. The normality assumption of the *t*-test is not required for Wilcoxon-Mann-Whitney | [55] |
| | | Non-Parametric | Wilcoxon-Mann-Whitney test | | [56] |
| | Paired | Parametric | Paired *t*-test | Tests to be used for dependent groups in the paired design, e.g. same individuals before and after treatment. | [55] |
| | | Non-Parametric | Sign test or Wilcoxon signed ranks test | | [57] |
| >2 | Independent | Parametric | One-way analysis of variance (ANOVA) | Collection of procedures to analyse different experimental designs. In its simpler form can be used to test the differences among group means. Kruskal-Wallis test is its non-parametric counterpart. | [58] |
| | | Non-Parametric | Kruskal-Wallis test | | [59] |
| | Paired | Parametric | Repeated measure ANOVA | Extension of ANOVA for paired samples, when the measures are repeated at multiple times, e.g. same individuals at different time points. | [60] |
| | | Non-Parametric | Friedman test or Quade test | The multigroup extensions of sign and signed ranks tests. According to Conover, Friedman test is typically more powerful when the number of groups is ≥ 5, and vice versa. | [54] |

* An exhaustive compendium of statistical tests can be found in [61]

## References


[1] A. M. De Livera, D. A. Dias, D. De Souza, T. Rupasinghe, J. Pyke, D. Tull, U. Roessner, M. McConville, T. P. Speed, *Anal. Chem.* **2012**, *84*, 10768–10776.
[2] L. G. Rasmussen, F. Savorani, T. M. Larsen, L. O. Dragsted, A. Astrup, S. B. Engelsen, *Metabolomics* **2011**, *7*, 71–83.
[3] H. Martens, J. P. Nielsen, S. B. Engelsen, *Anal. Chem.* **2003**, *75*, 394–404.
[4] A.-H. Emwas, E. Saccenti, X. Gao, R. T. McKay, V. A. P. M. dos Santos, R. Roy, D. S. Wishart, *Metabolomics* **2018**, *14*, 31.
[5] P. Filzmoser, B. Walczak, *J. Chromatogr. A* **2014**, *1362*, 194–205.
[6] S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, W. Gronwald, *Metabolomics* **2012**, *8*, 146–160.
[7] E. Saccenti, *J. Proteome Res.* **2017**, *16*, 619–634.
[8] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
[9] H. Abdi, L. J. Williams, *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459.
[10] R. Bujak, M. J. Markuszewski, in *Identif. Data Process. Methods Metabolomics*, Future Science Ltd, **2015**, pp. 82–95.
[11] M. Ringnér, *Nat. Biotechnol.* **2008**, *26*, 303–304.
[12] A. Hyvärinen, E. Oja, *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2000**, *13*, 411–430.
[13] M. V. Baptiste Féraud Réjane Rousseau, Pascal de Tullio, B. Govaerts*, *J. Biom. Biostat.* **2017**, *8*.
[14] M. E. Timmerman, *Br. J. Math. Stat. Psychol.* **2006**, *59*, 301–320.
[15] H. A. L. Kiers, J. M. F. ten Berge, *Psychometrika* **1989**, *54*, 467–473.
[16] D. L. S. Ferreira, S. Kittiwachana, L. A. Fido, D. R. Thompson, R. E. A. Escott, R. G. Brereton, *Analyst* **2009**, *134*, 1571–1585.
[17] J. Camacho, R. A. Rodríguez-Gómez, E. Saccenti, *J. Comput. Graph. Stat.* **2017**, *26*, 501–512.
[18] J. A. Hartigan, M. A. Wong, *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108.
[19] A. Srinivasan, C. J. Galbán, T. D. Johnson, T. L. Chenevert, B. D. Ross, S. K. Mukherji, *Am. J. Neuroradiol.* **2010**, *31*, 736–740.
[20] M. Cuperlović-Culf, N. Belacel, A. S. Culf, I. C. Chute, R. J. Ouellette, I. W. Burton, T. K. Karakach, J. A. Walter, *Magn. Reson. Chem. MRC* **2009**, *47 Suppl 1*, S96-104.
[21] L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley. Com, **2009**.
[22] J. Shi, J. Malik, *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
[23] P. Tiwari, M. Rosen, A. Madabhushi, *Med. Phys.* **2009**, *36*, 3927–3939.
[24] C. C. Bridges, *Psychol. Rep.* **1966**, *18*, 851–854.
[25] O. Beckonert, E. Bollard, T. M. D. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon, J. K. Nicholson, *Anal. Chim. Acta* **2003**, *490*, 3–15.
[26] S. Cacciatore, C. Luchinat, L. Tenori, *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 5117–5122.
[27] S. Cacciatore, L. Tenori, C. Luchinat, P. R. Bennett, D. A. MacIntyre, *Bioinformatics* **2017**, *33*, 621–623.
[28] T. Kohonen, *Neural Netw.* **2013**, *37*, 52–65.
[29] H. Zheng, J. Ji, L. Zhao, M. Chen, A. Shi, L. Pan, Y. Huang, H. Zhang, B. Dong, H. Gao, *Oncotarget* **2016**, *7*, 59189–59198.
[30] P. Baldi, in *PMLR*, **2012**, pp. 37–49.
[31] F. M. Alakwaa, K. Chaudhary, L. X. Garmire, *J. Proteome Res.* **2018**, *17*, 337–347.
[32] R. A. Fisher, *Ann. Eugen.* **1936**, *7*, 179–188.
[33] D. Yuan, Y. Liang, L. Yi, Q. Xu, O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* **2008**, *93*, 70–79.
[34] P. Geladi, *Chemom. Intell. Lab. Syst.* **1992**, *15*, vii–viii.
[35] S. Wold, L. Eriksson, *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
[36] H. Abdi, *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 97–106.
[37] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, R. Goodacre, *Anal. Chim. Acta* **2015**, *879*, 10–23.
[38] J. Trygg, S. Wold, *J. Chemom.* **2002**, *16*, 119–128.

[39] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, *Bioinforma. Oxf. Engl.* **2005**, *21*, 3043–3048.
[40] van V. E. al et, “Multilevel data analysis of a crossover designed human nutritional intervention study. - PubMed - NCBI,” can be found under https://www.ncbi.nlm.nih.gov/pubmed/18754629/, **n.d.**
[41] J. A. Westerhuis, E. J. van Velzen, H. C. Hoefsloot, A. K. Smilde, *Metabolomics* **2010**, *6*, 119–128.
[42] J. Camacho, E. Saccenti, *J. Chemom.* **n.d.**, n/a-n/a.
[43] T. Cover, P. Hart, *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
[44] C. Cortes, V. Vapnik, *J Mach Learn Rres* **1995**, *20*, 273–297.
[45] S. Mahadevan, S. L. Shah, T. J. Marrie, C. M. Slupsky, *Anal. Chem.* **2008**, *80*, 7562–7570.
[46] Y. Freund, R. E. Schapire, *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
[47] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, P. M. Thompson, *IEEE Trans. Med. Imaging* **2010**, *29*, 30–43.
[48] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[49] T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, A. Zhao, T. Chen, Y. Cao, et al., *Evid.-Based Complement. Altern. Med. Evid.-Based Complement. Altern. Med.* **2013**, *2013*, *2013*, e298183.
[50] C. V. D. Malsburg, in *Brain Theory*, Springer, Berlin, Heidelberg, **1986**, pp. 245–248.
[51] J. S. Sonawane, D. R. Patil, in *Int. Conf. Inf. Commun. Embed. Syst. ICICES2014*, **2014**, pp. 1–6.
[52] J. Schmidhuber, *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2015**, *61*, 85–117.
[53] J. H. McDonald, K. W. Dunn, *J. Microsc.* **2013**, *252*, 295–302.
[54] W. J. Conover, *Practical Nonparametric Statistics, 3rd*, Wiley, New York, **1999**.
[55] D. FRALICK, J. Z. ZHENG, B. Wang, X. M. TU, C. FENG, *Shanghai Arch. Psychiatry* **n.d.**, *29*, 184–188.
[56] M. Neuhäuser, in *Int. Encycl. Stat. Sci.*, Springer, Berlin, Heidelberg, **2011**, pp. 1656–1658.
[57] G. W. Snedecor, W. G. Cochran, *Statistical Methods*, Iowa State University Press, Ames, Iowa, **1989**.
[58] J. Kaufmann, A. Schering, in *Wiley StatsRef Stat. Ref. Online*, John Wiley & Sons, Ltd, **2014**.
[59] W. H. Kruskal, W. A. Wallis, *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.
[60] R. Bakeman, *Behav. Res. Methods* **2005**, *37*, 379–384.
[61] G. K. Kanji, *100 Statistical Tests*, SAGE Publications Ltd, London; Thousand Oaks, Calif., **2006**.