Distinguishing between Spectral Clustering and Cluster Analysis of Mass Spectra

Hélène Borges, Romain Guibert, Olga Permiakova, Thomas Burger

Distinguishing between spectral clustering and cluster analysis of mass spectra

Hélène Borges[#,1], Romain Guibert[#,1,2], Olga Permiakova[#,1], Thomas Burger*[,1,2]
[#]equal contribution, listed in alphabetical order
*thomas.burger@cea.fr

[1] Univ. Grenoble Alpes, CEA, INSERM, BIG-BGE, 38000 Grenoble, France
[2] CNRS, BIG-BGE, F-38000 Grenoble, France

***Abstract**: The term "spectral clustering" is sometimes used to refer to the clustering of mass spectrometry data. However, it also classically refers to a family of popular clustering algorithms. To avoid confusion, a more specific term could advantageously be coined.*

***Keywords:** Cluster analysis; Spectral clustering; Mass spectrometry; Proteomics*

Introduction: In proteomics literature, "spectral clustering" refers to performing a cluster analysis on a dataset resulting from mass spectrometry (MS) acquisitions, with the objective to answer a wide variety of analytical questions (which have recently been surveyed by Perez-Riverol et al.).[1] However, this term also names a widespread family of algorithms for cluster analysis. This ambiguity deteriorates the keyword indexing quality of any work focusing on cluster analysis of MS spectra, and thus complicates the inevitable state-of-the-art review of new research in computational proteomics.[2] We believe there are advantages to adjusting the naming convention. Therefore, we propose to refer to the cluster analysis of MS data as "spectrum clustering" (its original name, see below); or to avoid spelling similarities and improve indexing, as "mass spectrum clustering".

Cluster analysis refers to a wide family of unsupervised statistical learning and multivariate analysis techniques. Roughly speaking, its goal is to aggregate similar observations into clusters, so that the resulting clusters are as dissimilar as possible. Cluster analysis has numerous applications in a variety of scientific domains, including omics biology. Thus, its use on large-scale proteomics data is bound to develop.[1]

To the best of the authors' knowledge, the idea of applying cluster analysis to MS based proteomics data[3,4,5,6] goes back to the mid-2000s. Interestingly, the first proposal (in 2004)[3] referred to such clustering task as "spectrum clustering", before the term "spectral clustering" was coined in 2005 by Tabb et al.[4] The latter term was then used from time to time (for instance Bonanza algorithm[6] in 2008, or PRIDE Cluster[7] in 2013), before witnessing a recent regain of interest,[8,1] notably through a scholarly discussion on the subject relayed by Journal of Proteome Research.[9,8]

The term "spectral clustering" also designates a specific family of clustering algorithms, with theoretical foundations that are nearly twenty years old.[10,11] Its name roots in algebra vocabulary, where the set of eigenvalues of a matrix is commonly referred to as its "spectrum". From an analytical chemistry viewpoint, this naming convention is surprising, while remaining compliant with the general meaning of "spectrum", i.e. a decomposition into elementary constituents (light spectrum, mass spectrum, etc.): Conceptually, eigenvalues amount to the atomic elements of a matrix.

Spectral clustering first developed in the machine learning community, and for several years, it has been almost exclusively applied to computer vision problems. This largely explains why, *(i)* the term was independently coined to refer to the cluster analysis of mass spectra; *(ii)* no cluster analysis of spectra reported in the proteomics literature has been conducted with it so far (on the contrary, other cluster analysis techniques, e.g. hierarchical clustering, are regularly applied to MS-based proteomics

data).[7,12]. However, since its original development, spectral clustering techniques have stepped out of computer vision applications, and have been demonstrated to be extremely powerful on various application domains, so that their popularity is now unparalleled. To date, Shi and Malik's seminal article[10] has gathered more than 15000 citations according to Google Scholar, and other articles explicitly referring to the term "spectral clustering" in their title gather as many of them (see Table 1).

Table 1: List of the five first articles proposed by Google Scholar when searching « Spectral Clustering », accompanied with the number of citations of these articles (on October 23, 2018).

Authors & year	Reference	# citations
Ng, Jordan & Weiss (2002)	13	6967
Von Luxburg (2007)	14	6087
Zelnik-Manor & Perona (2005)	15	1793
Dhillon, Guan & Kulis (2004)	16	999
Bengio et al. (2004)	17	1050

Briefly, the principle of spectral clustering is the following: First, the dataset is endowed with a graph structure. Then, one performs a dimensionality reduction guided by the eigenvalues of a specific matrix, referred to as the graph Laplacian (in the algorithm name, "spectral" thus refers to the graph Laplacian eigenvalues). Lastly, k-means clustering is performed via a Lloyd-type algorithm.[18] Concretely, the graph Laplacian encodes the connectivity levels between the vertices of the data graph (a kind of "diffusion capability" for each data item towards its neighborhood). Therefore, working on this matrix makes sense, as good clusters supposedly correspond to sets of highly connected vertices with few inter-cluster connections. This explains why an accurate clustering is expectable, even for datasets with a complex structure that cannot be captured by ball-shape clusters (as with classical k-means). For instance, Figure 1 represents a famous toy dataset with two intermingled spiraling clusters (classically referred to as Swiss-rolls), on which spectral clustering achieves good performance. Nowadays, applying spectral clustering algorithms to data of various types is rather straightforward thanks to very detailed and pedagogical tutorials[14] as well as efficient toolboxes (e.g. Kernlab R package).[19]

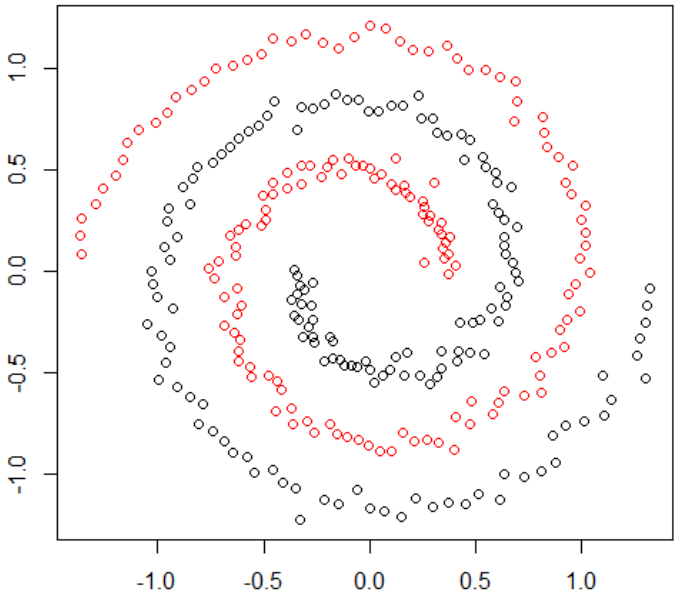


Figure 1: A typical toy dataset with a complex non-linear structure (two intermingled Swiss-rolls) accurately clustered thanks to the default spectral clustering algorithm available in Kernlab.[19]

Finally, one nowadays uses "spectral clustering" to name two different notions: First, it has been regularly used since 2000 to refer to a family of clustering algorithms. Second, it has been sometimes used since 2005 to refer to cluster analysis of spectral data. Among the large number of co-existing scientific domains, it is common to have the same names used to refer to different concepts. However, as the computational aspects of proteomics grow up, and as big data tools become pervasive in the processing of MS data, some confusion may appear. This is likely for the following reason: Spectral clustering popularity mainly comes from the underlying data embedding into a graph structure, which makes it particularly efficient for network-based data (ranging from social network[20] to interaction networks in biology).[21, 22, 23] As this type of data becomes customary in interactomics studies,[24] spectral clustering techniques are likely to become essential tools for proteomics data analysis. As a result, the proteomics community has much to gain in avoiding vocabulary confusion.

In the past, similar vocabulary confusions due to proteomics getting closer to data science were already witnessed. Notably, the concept of *False Discovery Rate*[25] was confused with what are respectively called in the biostatistics literature, the *False Positive Rate* and the *False Discovery Proportion*, as opportunely pointed by Käll et al[26] (for the former) and Serang & Käll[27] (for the latter). These two confusions did not help MS experts to get involved with the increasing use of statistics in proteomics. To avoid similar misunderstanding, it would make sense to return to the original naming convention[3] (i.e. "spectrum clustering") or to coin a more precise one, specific enough to be well visible and well indexed, such as "mass spectrum clustering".

Author Information

Corresponding author

*Email : thomas.burger@cea.fr . Tel: +33 4 38 78 22 72

ORCID

orcid.org/0000-0003-3539-3564

Notes

The authors declare no competing financial interest.

Acknowledgements

This work was supported by grants from the French National Research Agency: ProFI project (ANR-10-INBS-08), GRAL project (ANR-10-LABX-49-01) and LIFE project (ANR-15-IDEX-02).

References

1 Perez-Riverol, Y., Vizcaíno, J. A., & Griss, J. (2018). Future prospects of spectral clustering approaches in proteomics. *Proteomics*, 1700454.

2 An anonymous reviewer kindly remarked this question has been debated during the 2015 Mid-Winter Proteomics Informatics meeting at Semmering, (https://coreforlife.eu/events/2014/midwinter-proteomics-bioinformatics-seminar) where the consensus was that of the status quo. The goal of this Letter is to extend and enlarge this discussion.

3 Beer, I., Barnea, E., Ziv, T., & Admon, A. (2004). Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, *4*(4), 950-960.

4 Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., VerBerkmoes, N. C., & McDonald, W. H. (2005). MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *Journal of the American Society for Mass Spectrometry*, *16*(8), 1250-1261.

5 Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., & Pevzner, P. A. (2007). Clustering millions of tandem mass spectra. *Journal of proteome research*, *7*(01), 113-122.

6 Falkner, J. A., Falkner, J. W., Yocum, A. K., & Andrews, P. C. (2008). A spectral clustering approach to MS/MS identification of post-translational modifications. *Journal of proteome research*, *7*(11), 4614-4622.

7 Griss, J., Foster, J. M., Hermjakob, H., & Vizcaíno, J. A. (2013). PRIDE Cluster: building a consensus of proteomics data. *Nature methods*, *10*(2), 95.

8 Griss, J., Perez-Riverol, Y., The, M., Käll, L., & Vizcaíno, J. A. (2018). Response to "Comparison and Evaluation of Clustering Algorithms for Tandem Mass Spectra". *Journal of proteome research*, *17*(5), 1993-1996.

9 Rieder, V., Schork, K. U., Kerschke, L., Blank-Landeshammer, B., Sickmann, A., & Rahnenführer, J. (2017). Comparison and evaluation of clustering algorithms for tandem mass spectra. *Journal of proteome research*, *16*(11), 4035-4044.

10 Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), 888-905.

11 Kannan, R., Vempala, S., & Veta, A. (2000). On clusterings-good, bad and spectral. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on* (pp. 367-377). IEEE.

12 The, M., & Käll, L. (2016). MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *Journal of proteome research*, *15*(3), 713-720.

13 Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems* (pp. 849-856).

14 Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416.

15 Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. In *Advances in neural information processing systems*(pp. 1601-1608).

16 Dhillon, I. S., Guan, Y., & Kulis, B. (2004, August). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 551-556). ACM.

17 Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Roux, N. L., & Ouimet, M. (2004). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Advances in neural information processing systems* (pp. 177-184).

18 Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

19 Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, 11(9), 1-20.

20 Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web* (pp. 695-704). ACM.

21 de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC bioinformatics*, *9*(1), 497.

22 Thurlow, J. K., Peña Murillo, C. L., Hunter, K. D., Buffa, F. M., Patiar, S., Betts, G., ... & Ozanne, B. W. (2010). Spectral clustering of microarray data elucidates the roles of microenvironment remodeling and immune responses in survival of head and neck squamous cell carcinoma. *Journal of clinical oncology*, *28*(17), 2881-2888.

23 Qin, G., & Gao, L. (2010). Spectral clustering for detecting protein complexes in protein–protein interaction (PPI) networks. *Mathematical and Computer Modelling*, *52*(11-12), 2066-2074.

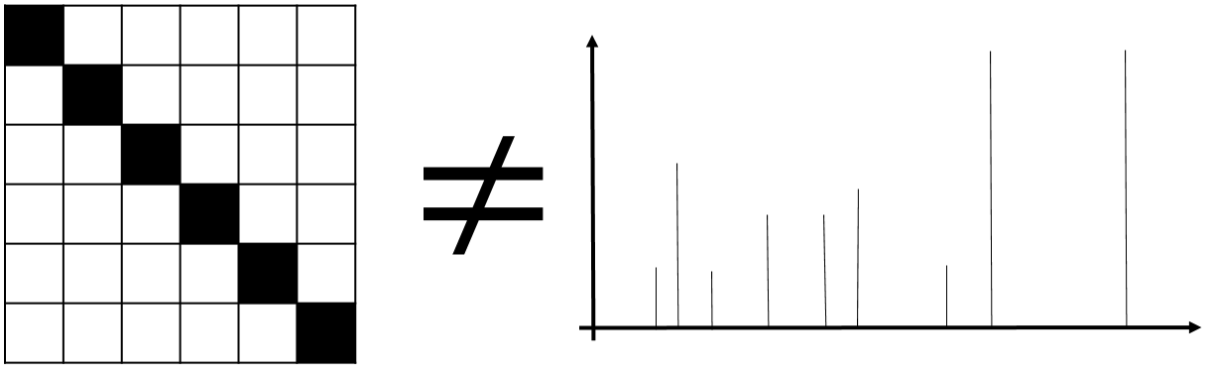
24 Bensimon, A., Heck, A. J., & Aebersold, R. (2012). Mass spectrometry–based proteomics and network biology. *Annual review of biochemistry*, *81*, 379-405.

25 Burger, T. (2017). Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *Journal of proteome research*, 17(1), 12-22.

26 Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2007). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, *7*(01), 29-34.

27 Serang, O., & Käll, L. (2015). Solution to statistical challenges in proteomics is more statistics, not less. *Journal of proteome research*, *14*(10), 4099-4103.

For TOC only



Spectrum (mathematics)

Spectrum (proteomics)