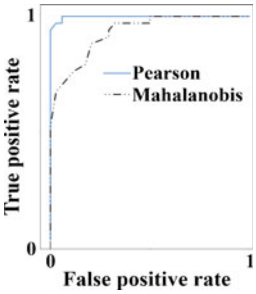
J. Am. Soc. Mass Spectrom. (2014) 25:293–296
DOI: 10.1007/s13361-013-0773-z

APPLICATION NOTE

# Comparison of the Mahalanobis Distance and Pearson's $\chi^2$ Statistic as Measures of Similarity of Isotope Patterns

Fatemeh Zamanzad Ghavidel,[1] Jürgen Claesen,[1] Tomasz Burzykowski,[1] Dirk Valkenborg[1,2,3]

[1]I-BioStat, Hasselt University, Hasselt, Belgium
[2]Applied Bio and Molecular Systems, Flemish Institute for Technological Research, VITO, Mol, Belgium
[3]Center for Proteomics, Antwerp, Belgium



**Abstract.** To extract a genuine peptide signal from a mass spectrum, an observed series of peaks at a particular mass can be compared with the isotope distribution expected for a peptide of that mass. To decide whether the observed series of peaks is similar to the isotope distribution, a similarity measure is needed. In this short communication, we investigate whether the Mahalanobis distance could be an alternative measure for the commonly employed Pearson's $\chi^2$ statistic. We evaluate the performance of the two measures by using a controlled MALDI-TOF experiment. The results indicate that Pearson's $\chi^2$ statistic has better discriminatory performance than the Mahalanobis distance and is a more robust measure.

**Key words:** Similarity statistics, Isotope distributions, Mass spectral data interpretation, Bioinformatics, Mahalanobis distance

Received: 30 April 2013/Revised: 7 October 2013/Accepted: 14 October 2013/Published online: 19 November 2013

## Introduction

In high-resolution mass spectrometry, proteins and peptides appear in a mass spectrum as a series of locally correlated peaks. This specific characteristic is related to the isotope distribution of a peptide. The isotope distribution is given by the probabilities of occurrence of all possible isotope variants of a peptide. It can be conveniently calculated when the atomic composition and the elemental isotope distribution are known [1, 2]. The information about the isotope distribution is used to interpret mass-spectrometry data in many applications. For example, it can be employed to discern genuine peptide peaks from noise [3, 4], to determine the monoisotopic peak [5, 6], or to study conformational dynamics of peptides and proteins using the hydrogen/deuterium exchange [7].

In a spectrum, peptide peaks can be scrutinized by assessing the degree of similarity between the observed pattern of peaks and the isotope distribution expected for a peptide with a similar mass [8–10]. The idea is illustrated in Figure 1. To this aim, a similarity measure is needed. Currently, the standard measure is Pearson's $\chi^2$ statistic and it has been rigorously investigated [7]. It is based on a weighted sum of the squared deviations between the expected and observed peaks [11, 12].

However, alternative similarity measures could be considered that can include information about possible correlation between the intensity peaks of a isotope distribution. In this short communication, we evaluate the use of Pearson's $\chi^2$ statistic and compare it to the Mahalanobis distance [13]. The latter similarity metric calculates the generalized distance and was described in a seminal paper by J. C. Mahalonobis. In mass spectrometry, the Mahalanobis distance is employed as a metric for outlier detection in the context of data quality assement and it operates on a particular feature set [14–16]. Additionally, the metric is often included in the object function of machine learning methods as a global distance measure [17] to classify spectral data. Nevertheless, the Mahalonobis distance has never been proposed for the interpretation of the isotope patterns observed in mass spectra. For this purpose, a controlled MALDI-TOF experiment on bovine cytochrome *c* was conducted to evaluate its performance on resolved isotope peaks.

## Experimental

A peptide mixture of tryptic-digested bovine cytochrome *c* was purchased from Sunnyvale, CA, USA and mixed with five internal standards from Sophia-Antipolis, France used for

*Correspondence to:* Dirk Valkenborg; *e-mail:* dirk.valkenborg@vito.be

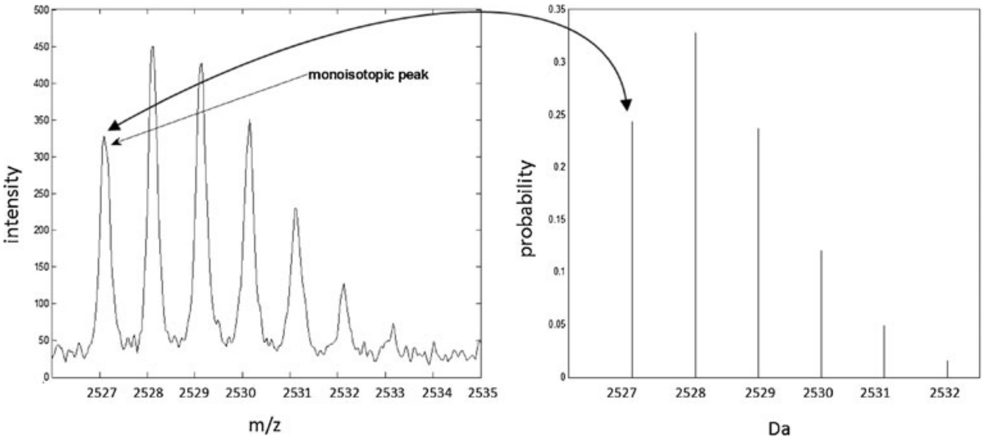


**Figure 1.** Left: an observed isotope pattern; right: a hypothetical isotope distribution. The observed pattern, which could be originating from a peptide, is compared with the hypothetical isotope distribution. The lightest isotopic variant of a peptide (i.e., the variant that is composed out of $^{12}$C, $^{1}$H, $^{14}$N, $^{16}$O, and $^{32}$S atoms) is called the monoisotopic variant. The corresponding peak is called the monoisotopic peak (indicated with an arrow in the left-hand-side panel)

the calibration of the mass spectrometer. According to the data sheets of the suppliers, the mixture contains 17 protein fragments. The amino acid sequences and theoretical monoisotopic masses *(m)* of these fragments are known. The mixture was blended with the matrix molecules and automatically spotted 384 times on one stainless-steel plate by a robot. The plate was processed on a 4800 MALDI-TOF/TOF analyzer (500 Old Connecticut Path, Framingham, MA 01701 U.S.A.) mass spectrometer, which resulted in 384 mass spectra.

First, we focus on the series of four consecutive, 1 Da separated peaks, which are consistently found in more than 90 % of the 384 spectra. We call such series isotopic clusters. The reason for extracting the first four isotope peaks is that the mass of the peptides in the sample are predominantly in the range of 568.1 to 2465.2 Da. Consequently, it is reasonable to assume that the isotope distributions of these peptides are sufficiently characterized by the first four isotope peaks. In total, 35 of such clusters are selected. For 12 clusters, the mass corresponds to the monoisotopic mass of one of the 17 protein fragments known to be present in the mixture. The additional 23 candidates were found to be related to peptides resulting from modifications or artifacts of the proteolytic background [18].

In addition to the 35 putative-peptide isotopic clusters, we select 35 clusters of noise peaks. The selected noise peaks are separated by 1 Da as well, but do not appear consistently across the 384 spectra. The noise peaks are located in mass regions in the neighborhood of the selected peptide isotope clusters. The data from the noise and peptide peak clusters are used as a benchmark to assess the ability of the similarity measures to discriminate noise peaks from peptide peaks.

## Methodology

The comparison between an observed series of peaks with a hypothetical isotope distribution can be performed by considering isotope ratios. An isotope ratio is the ratio between two subsequent peaks in an observed series of peaks or in a theoretical isotope distribution. For example, the first isotope ratio, $O_1$, is equal to the height of the second peak divided by the height of the first peak. The rationale for working with isotope ratios is that ratios are dimensionless and their use allows us to avoid scaling of the expected and observed intensity values. In addition, ratios are not sensitive to multiplicative noise.

In the proposed methodology, a model is required to predict the hypothetical isotope distribution. To this aim, the polynomial regression model described by Valkenborg et al. [19] can be used. Alternately, an empirical estimate can be applied. For this purpose, the Human HUPO database was digested in silico by using trypsin as a protease. The digest led to 126,376 peptides with masses ranging from 400 to 4000 Da. The program BRAIN [20] was used to calculate the isotope distribution and monoisotopic masses of the resulting peptides. For a given mass of $m$ Da, a set of peptides with monoisotopic masses within the interval of $[m - 5, m + 5]$ Da was selected. Next, the mean value of isotope ratios was calculated for the peptides within the assumed mass interval and stored in vector $\boldsymbol{R}$. Additionally, the variance-covariance matrix of the ratios $\boldsymbol{\Sigma}$ was retrieved from the selected data as well.

To assess the degree of similarity between the observed and expected isotope ratios, a similarity measure is needed. The standard measure (i.e., Pearson's $\chi^2$ statistic) is defined as follows:

$$\chi^2 = \sum\nolimits_{i=1} (O_i - R_i)^2 / R_i \tag{1}$$

where $O_i$ is the observed value of the $i^{th}$ consecutive isotopic ratio $(i = 1,2,3)$ and $R_i$ is the corresponding expected value.

An alternative similarity measure could be the Mahalanobis distance [13]. The distance takes into account the variability and correlation of the ratios, and is defined as follows:

$$M = \left\{ (\boldsymbol{O} - \boldsymbol{R})' \sum\nolimits^{-1} (\boldsymbol{O} - \boldsymbol{R}) \right\}^{1/2} \tag{2}$$

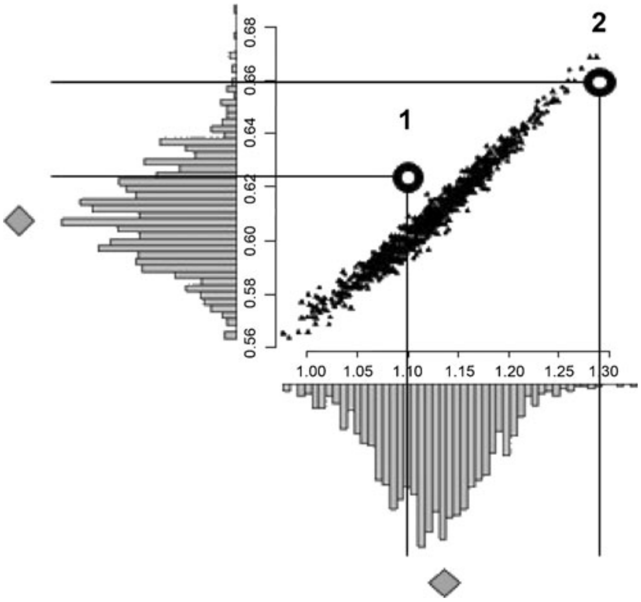


Figure 2. Scatter plot of the first isotope ratio (x-axis) and the second isotope ratio (y-axis) of peptides with a mass between 2000 and 2020 Da based on the Human HUPO database

where $\boldsymbol{O}$ and $\boldsymbol{R}$ denote the vectors containing, respectively, the observed and expected consecutive isotope ratios, and $\boldsymbol{\Sigma}$ denotes the variance-covariance matrix of the expected ratios. The expected ratios and corresponding variance-covariance matrix were calculated using the theoretical isotope dsitributions from the Human HUPO database. Note that the expected values can also be computed by the polynomial model, which is more straightforward from a practical point of view.

The use of the Mahalanobis distance is motivated by the fact that it takes into account the correlation between isotope ratios, which could allow for a better discriminatory performance. The motivation is illustrated in Figure 2. The figure presents the scatter plot of the first and second isotope ratio for peptides with a mass between 2000 Da and 2020 Da from the Human HUPO database. The grey diamonds above the histograms indicate the mean values of 1.1243 and 0.6082 for the first and second isotope ratios, respectively. The plot indicates that there is a substantial amount of correlation between the two ratios.

Consider the two points, marked by the black circles. A similarity measure that does not take into account the correlation would more likely classify point 1 as a genuine peptide because its coordinates are close to the mean values. However, a measure taking into account the correlation, such as, e.g., the Mahalanobis distance, would most likely opt for point 2 because the coordinates of this point reflect the (linear) association resulting from the joint distribution of the two isotope ratios.

## Results

For the selected peptide and noise peak-clusters, we calculate Pearson's $\chi^2$ statistic and the Mahalanobis distance. Figure 3 summarizes the performance of the two similarity measures for a set of randomly selected spectra. To this aim, the receiver operating characteristic (ROC) curve is used. Each point on the ROC curve represents a sensitivity/

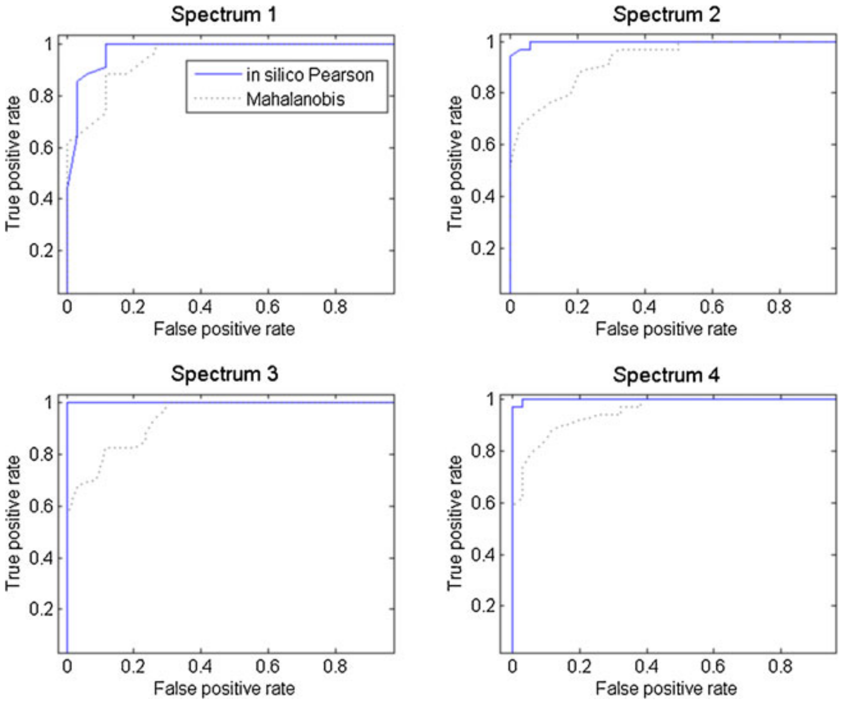


Figure 3. ROC curves for each of the four analyzed spectra

specificity pair corresponding to a particular cut-off point for the similarity measure. A perfect discrimination is reflected by a curve passing through the (0,1) point at the upper-left side, as seen in the plot for Spectrum 3 for Pearson's $\chi^2$.

The ROC curves presented in Figure 3 indicate that the discriminative performance of Pearson's $\chi^2$ statistic is uniformly better than that of the Mahalanobis distance. Similar results (not shown) are obtained for Pearson's $\chi^2$ statistic when using the polynomial model.

## Conclusions

Our analysis indicates that Pearson's $\chi^2$ statistic offers a better discriminative power for detecting the peptide clusters than the Mahalanobis distance. This result is most likely due to the fact that the Mahalanobis distance is very much based on the assumed form of the variance-covariance matrix $\Sigma$. The matrix derived from the in-silico tryptic digest database may not be adequate for the isotope ratios observed in a spectrum. Thus, Pearson's $\chi^2$ is the preferred statistic for evaluating the isotope distribution in mass spectrometry data.

An important practical point related to the use of Pearson's $\chi^2$ statistic is the choice of the threshold for deciding whether the observed isotope cluster is similar enough to the expected isotopic distribution. Based on our experiment, a threshold value of, e.g., 0.2 would be suitable. It is difficult to propose any concrete value for the threshold in general, though, as it most likely depends on the technological platform used to generate spectra. To empirically obtain a value of the threshold, an experiment and analyses similar to the ones presented in the paper could be performed.

## Acknowledgments

D.V. acknowledges the support of the SBO grant 'InSPECtor' (120025) of the Flemish agency for Innovation by Science and Technology (IWT).

## References

1. Valkenborg, D., Mertens, I., Lemière, F., Witters, F., Burzykowski, T.: The isotopic distribution conundrum. Mass Spectrom. Rev. **31**(1), 96–109 (2011)
2. Rockwood, A.L., Palmblad, M.: Isotopic distributions. Methods Mol. Biol. **1007**, 65–99 (2013)
3. Renard, B.Y., Kirchner, M., Steen, H., Steen, J.A.J., Hamprech, F.A.: NITPICK: Peak identification for mass spectrometry data. BMC Bioinforma **9**, 355 (2008)
4. Nicolardi, S., Palmblad, M., Dalebout, H., Bladergroen, M., Tollenaar, R.A., Deelder, A.M., van der Burgt, Y.E.: Quality control based on isotopic distributions for high-throughput MALDI-TOF and MALDI-FTICR serum peptide profiling. J. Am. Soc. Mass Spectrom. **21**(9), 1515–1525 (2010)
5. Senko, M.W., Beu, S.C., McLafferty, F.W.: Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. J. Am. Soc. Mass Spectrom. **6**, 229–233 (2005)
6. Hsieh, E.J., Hoopmann, M.R., Maclean, B., MacCoss, M.J.: Comparison of database search strategies for high precursor mass accuracy MS/MS data. J. Proteome Res. **9**(2), 1138–1143 (2010)
7. Palmblad, M., Buijs, J., Hakanson, P.: Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. J. Am. Soc. Mass Spectrom. **12**, 1153–1162 (2001)
8. Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K., Burzykowski, T.: Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. Rapid Commun. Mass Spectrom. **21**, 3387–3391 (2007)
9. Valkenborg, D., Thomas, G., Krols, L., Kas, K., Burzykowski, T.: A strategy to analyse data from high performance liquid chromatography combined with high resolution mass spectrometry. J. Mass Spectrom. **44**, 516–529 (2009)
10. Senko, M.W., Beu, S.C., McLafferty, F.W.: Automated assignment of charge states from resolved isotopic peaks for multiply-charged ions. J. Am. Soc. Mass Spectrom. **6**, 52–56 (1995)
11. Breen, E.J., Hopwood, F.G., Williams, K.L., Wilkins, M.R.: Automatic poisson peak harvesting for high throughput protein identification. Electrophoresis **21**, 2243–2251 (2000)
12. Gay, S., Binz, P.A., Hochstrasser, D.F., Appel, R.D.: Modeling peptide mass fingerprinting data using the atomic composition of peptides. Electrophoresis **20**, 3527–3534 (1999)
13. Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **2**(1), 49–55 (1936)
14. Matzke, M.M., Waters, K.M., Metz, T.O., Jacobs, J.M., Sims, A.C., Baric, R.S., Pounds, J.G., Webb-Robertson, B.J.: Improved quality control processing of peptide-centric LC-MS proteomics data. Bioinformatics **27**(20), 2866–2872 (2011)
15. Schulz-Trieglaff, O., Machtejevas, E., Reinert, K., Schlüter, H., Thiemann, J., Unger, K.: Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. BioDataMining **2**(1) Article 4 (2009)
16. Cairns, D.A., Perkins, D.N., Stanley, A.J., Thompson, D., Barrett, J.H., Selby, P.J., Banks, R.E.: Integrated multi-level quality control for proteomic profiling studies using mass spectrometry. BMC Bioinforma **9**, 519 (2008)
17. Liu, Q., Sung, A.H., Qiao, M., Chen, Z., Yang, J.Y., Yang, M.Q., Huang, X., Deng, Y.: Comparison of feature selection and classification for MALDI-MS data. BMC Genomics **10**(Suppl 1), S3 (2009)
18. Picotti, P., Aebersold, R., Domon, B.: The implications of proteolytic background for shotgun proteomics. Mol. Cell. Proteomics **6**(9), 1589–1598 (2007)
19. Valkenborg, D., Jansen, I., Burzykowski, T.: A model-based method for the prediction of the isotopic distribution of peptides. J. Am. Soc. Mass Spectrom. **19**(5), 703–712 (2008)
20. Dittwald, P., Valkenborg, D., Claesen, J., Burzykowski, T., Gambin, A.: BRAIN: A universal tool for high-throughput calculations of isotopic distribution for mass spectrometry. Anal. Chem. **85**(4), 1991–1994 (2013)