# An Improvement Over the CODA Filter for LC-MS Analysis.

Daniel Haeser Rech[1], Susana Linskens[2], Apuã Paquola[3], Antônio Augusto M. Fröhlich[4], Edmundo C. Grisard[1], Sirlei Daffre[5], and Daniel Macedo Lorenzini[5,1]

[1] Laboratório de Bioinformática - MIP – UFSC
haeser@lisha.ufsc.br, grisard@ccb.ufsc.br
[2] Lanais-Pro – UBA - CONICET
linskens@qb.ffyb.uba.ar
[3] Laboratório de Reparo de DNA - ICB – USP
apua@ime.usp.br
[4] Laboratório de Integração Software / Hardware – INE - UFSC
guto@lisha.ufsc.br
[5] Laboratório de Bioquímica e Imunologia de Artrópodes - ICB – USP
sidaffre@icb.ufsc.br, dloren@usp.br

**Abstract.** Liquid chromatography combined with mass spectrometry (LC-MS) analysis has been successfully employed in proteome research. Data generated from LC-MS equipments usually comes with a considerable amount of noise. Thus, there is a need for the use of filters to enhance the quality of mass spectrometry chromatograms. An algorithm developed recently (CODA[2]) achieves good results in removing noise in a short time. On the present work it will be shown some modifications to the CODA algorithm that improve considerably its results. Further, a set of testing chromatograms will be presented to demonstrate the efficiency of the modified algorithm.

## 1 Introduction

Liquid chromatography combined with mass spectrometry (LC-MS) analysis has been successfully employed in proteome research. Chromatograms generated by this method present high sensibility and resolution. This method can also be automated easily, enabling large-scale proteome analysis [1].

However, chromatograms generated by mass spectrometers in LC-MS analysis commonly contain an undesirable amount of noise. There are usually two types of noise present in these chromatograms: background and spikes. Background is caused by impurities present in the solvents used in the liquid chromatography. Spikes are single scan artifacts produced by the mass spectrometer that doesn't represent real data.

Depending on the amount of noise on a chromatogram, the direct analysis of the data may become very difficult. To ease the process of identifying real peaks, Windig [2] developed the CODA (Component Detection Algorithm) filter for removal of

background and spikes from chromatograms. However, this algorithm still allows a good deal of noise to pass its filtering.

## 2 Theory

The CODA algorithm uses a scoring method for removing low quality chromatograms (the entire algorithm is explained in [2]). It works by calculating the global similarity between a chromatogram and its dynamically calculated smoothed version. In regions where the difference is high it means there are spikes in it, since spikes don't have a smooth form. Background is removed through the calculation of the mean of the intensities of the chromatogram. A high mean value characterizes high levels of background noise. Although the CODA filter removes all the noise from low scoring chromatograms, the high scoring chromatograms still could contain spikes and background.

A solution to this problem is to perform the calculation of local similarities in addition to the global similarity of a chromatogram and its smoothed version. This approach makes it possible for the algorithm to analyze a chromatogram in sub-regions, filtering out low scoring regions from high scoring chromatograms. We have extended the CODA algorithm to do such local scans. After the calculation of the global similarity, the modified algorithm also performs a window scan that calculates the local similarity of each window in a chromatogram. Depending on the score, the window is filtered as noise. The same formula for global similarity is applied to calculate the local similarities, but the indexes need to be adjusted to fit the window properly.

Another issue with the original CODA algorithm is the way it does spike filtering. The algorithm performs a window scan on a chromatogram, calculating the mean of the intensities of each scan contained on that window. If the window contains a spike, the mean will be greater than zero and the spike will be diminished, but will not be totally eliminated.

We have found a solution to this problem by calculating the median instead of the mean of the intensities. If inside a given window there is only one spike, the median of the intensities of the scans contained in that window will be zero, correctly eliminating the spike. This is also true for windows containing more than one spike, but on which zero intensity scans outnumber by at least one the scans containing spikes. This means that even a window with several spikes could still be eliminated from the chromatogram.

The use of the median requires the window to have at least a width of 3 scans, which is not a problem since a window shorter than 3 scans isn't enough for identifying real peaks. Also, the window should be no larger than double the minimum estimated width of the real peaks. If the scans containing a real peak are outnumbered by zero intensity scans on a given window, the algorithm will filter it as noise.

As presented in [WINDIG], the calculation of $A(w)^R$ should now be as follows:

$$a(w)_{ij}^{R} = \text{median}(a_{kj}) \,. \tag{1}$$

The calculation of local similarities and median of intensities increase the computational cost of the original algorithm, although in tests performed the total processing time was always less than 20 seconds in relatively low cost computers. This increase in the computational cost should be acceptable, since the improvements in the algorithm allow for better elimination of spikes and background in high quality chromatograms, something that the original algorithm could not perform at all.

## 3 Experimental

For testing the improved CODA algorithm, the following resources were used:

### 3.1 LC-MS Data

A mixture of six peptides from tryptic digestion of human growth hormone was prepared by adding equal amount of each purified peptide (around 0,5 pmol) .The sample was injected to a LC-MS system consisting of a Surveyor Pump (ThermoFinnigan) and an LCQ Duo Mass Spectrometer (ThermoFinnigan). A Vydac C18 column (250 x 1.0 mm) was used at a flow rate of 40 ul/min. The mobile phases used for gradient elution consisted of (A) 0,1 % formic acid and (B) acetonitrile/water (80:20 v/v) containing 0,1 % formic acid. A 0-50% linear gradient was applied during the first 60 min, followed by a 50-100 linear gradient for a further 10 min. MS system operated in the positive profile scan mode.

### 3.2 Algorithm Implementation

The modified CODA was implemented in C++ and tested on a computer with an AMD Athlon 1400MHz processor and 1 GB of RAM. Total processing time was always under 20 seconds in the tests performed.

## 4 Results and Discussion

The presence of noise is substantial in the unfiltered testing data. The original CODA filter efficiently removed low quality chromatograms, but failed to filter the noise on high quality ones. In the modified CODA, the reduction of noise in high quality chromatograms was clearly perceptible in the testing data. The process of identifying peaks becomes much easier after the data has passed through the new filter.

Both filters did not distort the mass spectrum and chromatogram data

significantly. In mass spectrometry it's very important that the real peaks remain with the same characteristics and the same positions after the filtering. Distorted data could lead to improper analysis.

Another way of comparing both filters is by how many non-zero intensity mass values they have in the output. As seen on table (1), the original CODA reduced the unfiltered data about 6 times. In contrast, the modified CODA reduced the same unfiltered data more than 30 times. This is an important advantage if the filtered data would be processed by a peak localization algorithm, meaning that it would have much less data to process.

**Table 1.**Total of non-zero scans for the unfiltered data, the original CODA filter and its modified version:

| | Unfiltered data | CODA | New CODA |
|---|---|---|---|
| non-0 scans | 1,464,588 | 245,981 | 46,078 |

## References


1. Peng, J., Gygi, S.P.: Proteomics: the move to mixtures. J Mass Spectrom. 36 (2001) 1083-91.
2. Windig, W., Phalp, J. M., Payne, A. W. A.: Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. Anal. Chem. 68 (1996) 3602-3606.