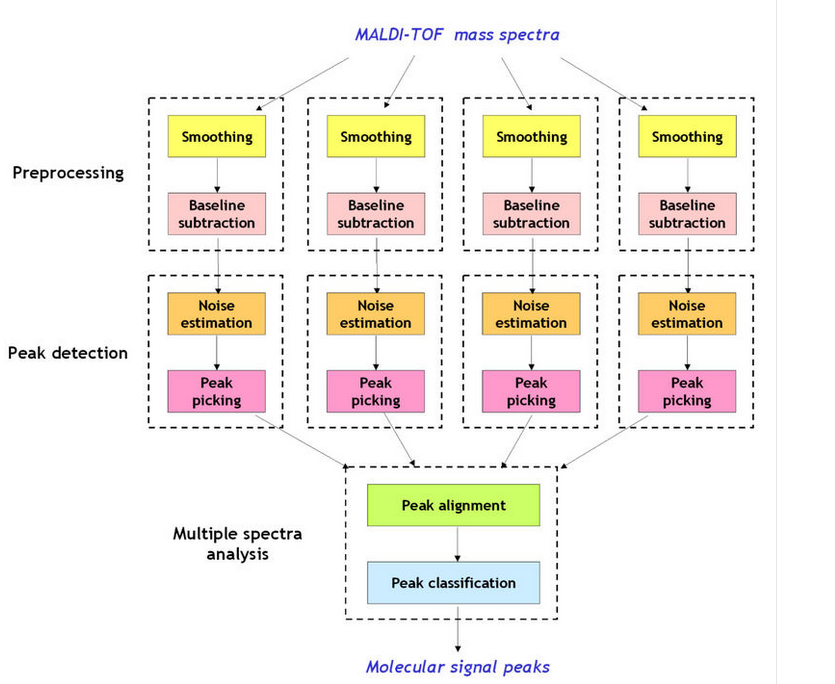
LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise

Commercial:APEX,CENTROID(BURKER)

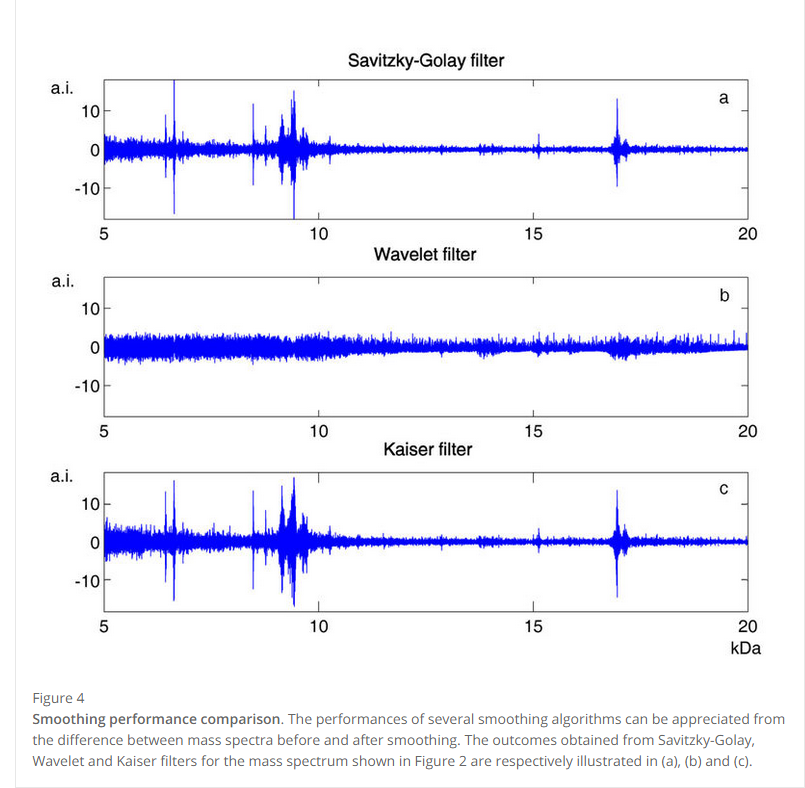
Open source: CROMWELL

 The acquired spectra always present complex features, because the protein signals, characterized by "true" peaks in the mass spectrum, can be contaminated by several chemical and/or physical processes of the measurement procedure [[6](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR6), [7](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR7)]. Two different kinds of disturbance can be revealed in the spectra: baseline drift and background noise. The baseline is the trend of the signal that would be generated by the mass spectrometer if no material was introduced into it. The background noise is a signal produced by electronic disturbances and fragments of material, with rapid fluctuations randomly varying over small mass ranges.

In-vitro purified protein samples and human plasma samples were used for the validation of : in both conditions, the proposed method showed a significant accuracy in the detection of the protein signals, and it was able to provide a superior sensitivity and specificity than two well-established commercial algorithms, APEX and CENTROID, and a freely available algorithm, CROMWELL.

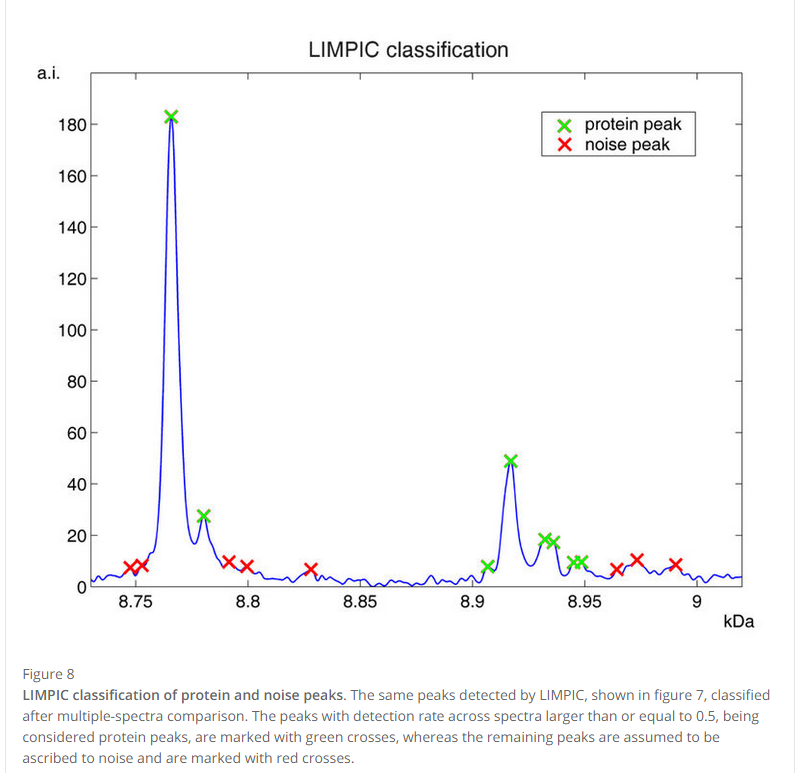
With regard to spectrum denoising, we analyzed the performances of Kaiser filter with different window lengths, in order to find the general criteria for setting this parameter. In fact, small window lengths might result in an insufficient denoising, whereas large ones might distort the "true" signa

The performances of Kaiser filter in terms of data smoothing were compared with those of Savitzky-Golay filter [[20](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR20)], the solution generally adopted in FlexAnalysis with APEX and CENTROID, and Wavelet filter [[10](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR10)], the smoothing technique used in CROMWELL.



Wavelet filter与另外两种filter算法效果不同

With regard to baseline removal, we compared the peak-elimination method used in LIMPIC with the classical method based on minimum value interpolation [[6](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR6), [10](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR10)], used by APEX, CENTROID and CROMWELL. The two estimates of the baseline drift are superimposed in Figure [5](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#Fig5)



Using the selection criterion based on the PDR value, it was also possible to measure the *hit-rate*, defined as the ratio between the number of "true" peak classes and the average number of the peaks revealed in the single spectra: the larger this ratio the better the performances of the peak-picking method. According to this parameter, a larger number of consistent peaks were generally detected with LIMPIC, confirming the superior reliability of this method for mass spectra from both protein mixtures and plasma samples (Table [3](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#Tab3)).

Table 3

and it seemed to be more effective than Savitzky-Golay and Wavelet filters with MALDI-TOF-MS data.

the LIMPIC baseline correction was able to work with all kinds of spectra, even with those characterized by a non-increasing baseline drift, for which the standard method based on fitting a monotone minimum curve might be unsuccessful [[10](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR10)

 According to the method proposed by Yasui [[8](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR8)], the positions of the peaks in the mass spectrum were determined on the basis of the first derivative of the signal, whereas other techniques based on template matching(PS:wavelet,文献里并没这点) were considered unreliable, because the presence of residual noise could corrupt the ideal peak shape and could determine incorrect outcomes

 a noise reduction technique was performed at first: the signal-to-noise ratio (SNR) was enhanced by using a smoothing procedure based on a Kaiser filter [[13](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR13)]. In short, a Kaiser filter is a Finite-Impulse-Response (FIR) filter that approximates an ideal low-pass filter, while attempting to minimize the ripples in the frequency response caused by the signal truncation.

After centering a generic signal block v, the kurtosis kurt could be computed as

kurt(v) = E{v4} - 3(E{v2})2     (3)

where E{•} is the expectation operator [[26](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR26)].

SNR=（r−c）/σ(4)

Subsequently, the processed spectrum was analyzed for the first step of the peak-picking procedure: if the point intensity was the highest among its nearest ± f points, a peak was detected in that position [[8](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR8)]. The number f could be varied to best fit the specific resolution of the mass spectrometer; in this study, f was set equal to 2, in order to cover a range of 0.5 Da. The second step of the peak-picking procedure was the elimination of the detected peaks with a SNR lower than a preset threshold. In accordance with the IUPAC guidelines [[14](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-101#CR14)], the limit of detection was set to 3σ, corresponding to SNR = 3.

a maximum tolerance distance d equal to 300 ppm of the m/z value was accepted for the comparison

single\_spectrum

1.Kaiser smoothing %需传一个平滑参数，与S-G很像

2.数据重采 %需传一个步长参数

3.将数据分成n个区域，求出每个区域y的最小值，kurt值,y的平均值，y的标准差以及x的平均值，然后求出整体信号y的平均值做为一阀值。

对这n个区域做一个筛选（y平均值大于整体信号平均值，且kurt值大于给定参数）

4.减基线：以各个区域x的平均值，以及y的最小值作为基点进行插值。

5.信号归一(需给定一个参考x)

在参考x附近区间找出一个峰值最大的峰作为参考峰进行归一

6.估计噪声值（）

以各区域x的平均值，y的标准差作为基点,对x进行插值得到相应的y1；以x的平均值，y的平均值-y的最小值作为基点，对x进行插值得到相应的y2；噪声 =归一因子\*（y2+常量因子\*y1)

7.找峰（峰宽，峰信号值，峰高）

1.flag1（凸起和凹起，感觉flag1可省略），flag2（左边的k个点斜率都要大于0（k由半峰宽决定））,flag3(凸起)，flag4（右边的k个点斜率都要小于0（k由半峰宽决定））

if flag1 & flag2 & flag3 & flag4

那就是一个潜在峰点，设为pos,长度为n

2.对这n个峰点过滤，如果其对应峰值小于其噪声，pass

3.估计每个峰的峰高，如果峰高小于阀值，Pass