Positive Probability Ltd

Note M5: LCMS Analysis - Proteins

Introduction

LCMS of intact proteins has been used by the FDA to identify biomarkers. The aim is to charge deconvolve the data to generate results peak tables of zero-charge masses, retention times and intensities for both healthy and diseased samples. There are generally several runs for each sample. The results from each sample are then compared to remove noise and, finally, the healthy and diseased samples are compared in an attempt to find significant differences that may represent a biomarker for the disease in question.

This approach for searching for biomarkers is not commonly applied for several reasons:

1. Charge deconvolutions using algebraic methods generate so many artefacts that the results are always suspect.
2. The computation time is unacceptably long for entropic and Bayesian methods – the time for a typical 2 hour run is of the order of a day or more! In addition, numerous artefacts are generated for large output mass windows of, typically, 5-50 kDa.
3. Background ions, which may be numerous for high S/N data, are not identified and removed and therefore confuse the results.
4. Genuine elutions that are present on top of background ions are missed.
5. Map comparison methods are crude and are unable to cope with more than small variations in the chromatography, making it impossible to compare maps run under even slightly different conditions or to compare maps at different points in the age of a column.

In order to make the LCMS analysis of intact proteins generally available, all the above points must be addressed. The problem may be divided into two parts:

1. The analysis of the data to provide reliable maps for each LC run.
2. The reliable comparison of maps and the identification of significant differences.

In this application note we deal with the generation of reliable maps. A separate application note deals with the comparison of maps. The processing is performed in a number of steps:

a) Evaluate the chromatography and co-add scans appropriately to improve the S/N.

b) Baseline correct each co-added block and deconvolve to obtain a reliable peak table.

c) Charge deconvolve to obtain zero-charge masses.

d) Deconvolve the RT dimension and remove background ions, retaining all genuine elutions.

e) Generate a map of zero-charge masses, retention times and intensities for any chosen confidence level.

Data

The data described here are a test mixture of at least 8 proteins of varying purity. Figure 1 displays the data in PPL's data viewer. There are 600 scans covering retention time 20-40 minutes (20 minutes of experiment time).

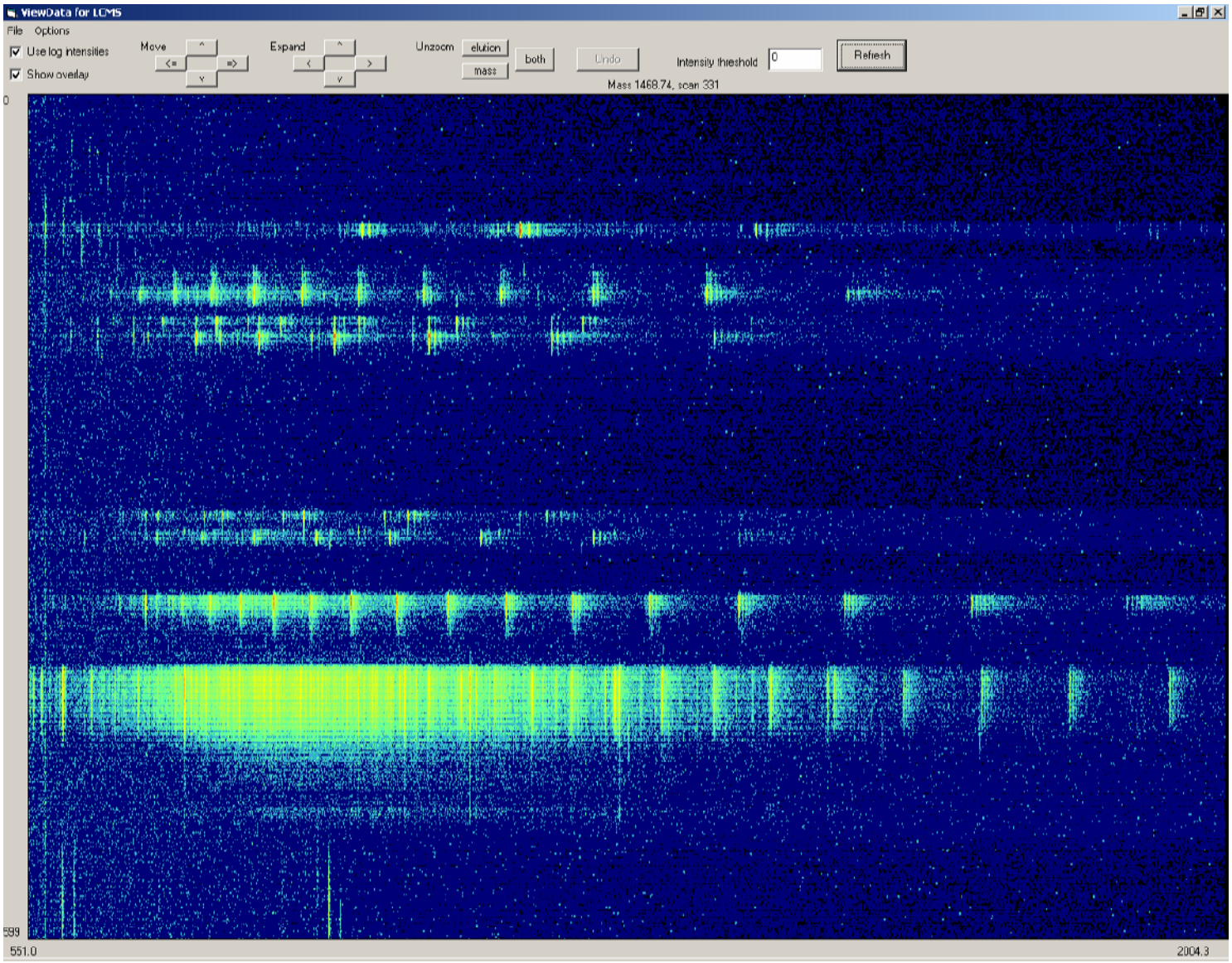

Figure 1. 600 scans (vertical axis) of LCMS data from m/z 551.0-2004.3 (horizontal axis).

Evaluating the Chromatography

Scans may be co-added to generate 'blocks' before processing. The number of scans to be co-added may be varied. To improve S/N and retain retention time information, blocks may overlap. A mass must be present in a chosen number of adjacent blocks before it is considered to be genuine.

For example, if peaks elute over 15 scans, it would be reasonable to co-add 5 scans, advance by 5 scans and co-add the next 5 scans, etc. An appropriate number of adjacent blocks to identify an elution would be 3.

If peaks elute over 7 scans, it would be reasonable to co-add 8 scans, advance by 4 scans and co-add the next 8 scans, etc. In this case, an appropriate number of adjacent blocks to identify elutions would be 2. Alternatively, 3 scans could be co-added, advance by 3 and use 2 or 3 adjacent blocks. However, the S/N for each block would be much lower.

Zooming in to the data shown in Figure 1 (see Figure 2 below) clearly shows the poor quality of the data and that many peaks are at the noise level. Significant peaks in this region (shortest elutions) occur over about 10 scans. Co-adds of 8 scans with an advance of 4 scans and 3 adjacent blocks to confirm an ion will just be adequate to reconstruct all significant masses. In fact, because a charge deconvolution is performed to provide zero-charge masses, it is masses that must be present in the chosen number of adjacent blocks and not ions.

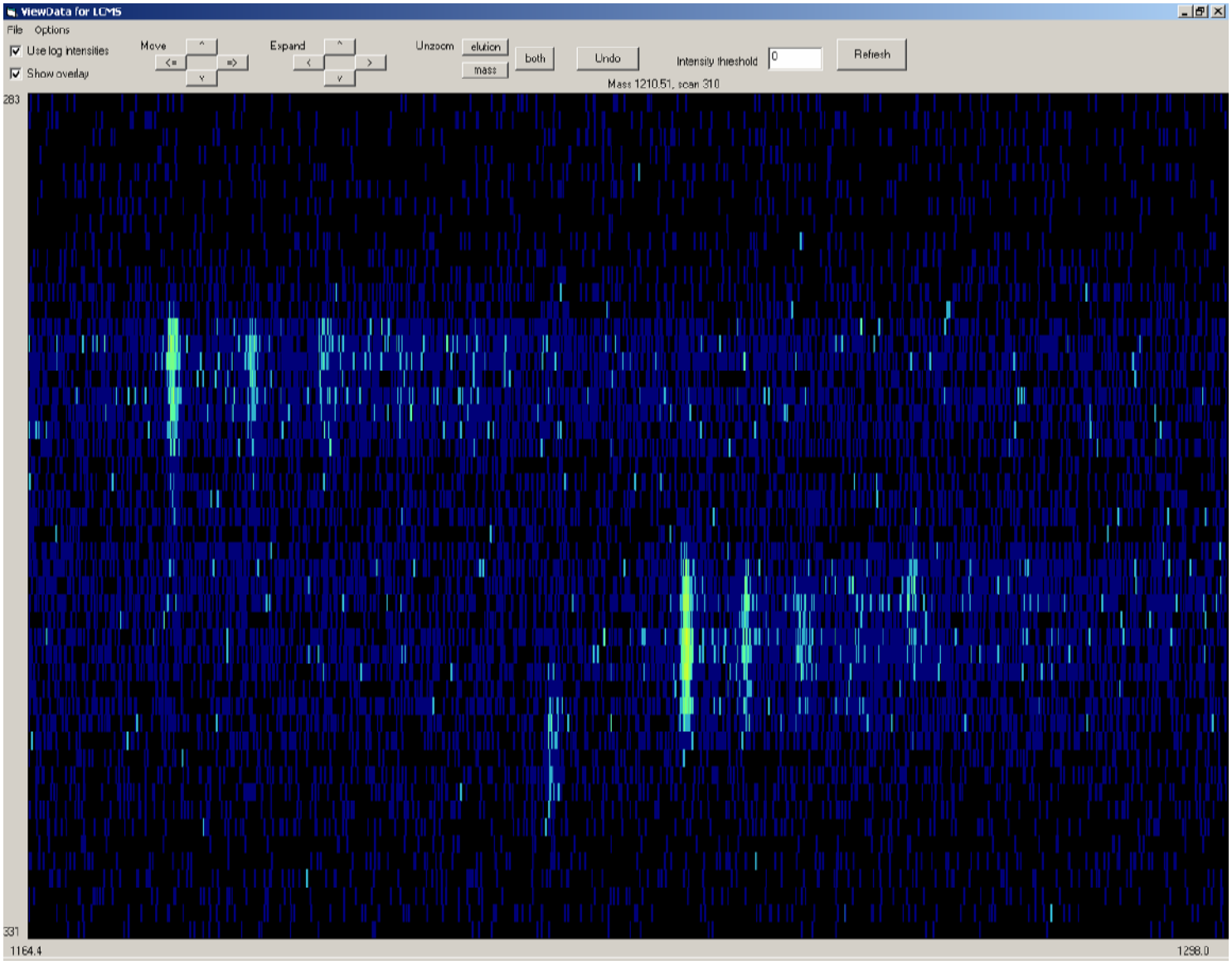

Figure 2. Scans 283-331 and m/z 1164-1298.

Although masses will be reconstructed for very weak ions that elute over just a few scans, it is unlikely that they will pass the minimum adjacent blocks filter if this is set too high.

Data Processing and Results

Preparing the Models

The peak width will be dependent on charge and mass but, depending on the instrument, acquisition conditions and range of masses present, the effect may or may not be severe. For proteins, models should be generated from the lower masses because these peaks will be broader. By co-adding a few suitable scans, models may be generated at different m/z so that any variation in peak width may be established and taken into account during the deconvolution of each block prior to charge deconvolution. The models may be saved and used for any experiment run under similar conditions. Each model may be displayed and their widths displayed as a function of m/z so that the list may be edited.

Figure 3 shows the Standards Page (model standards). Models have been generated for the lowest mass near the top of Figure 1. For this low mass the ions are resolved into isotopes. This is not the case for higher masses. It is therefore important to model the entire isotope envelopes and not individual isotope peaks. When working with proteins, modelling is performed over each charge as though they were single peaks. Therefore, when isotopes are resolved as seen here, they do not interfere with the modelling process.

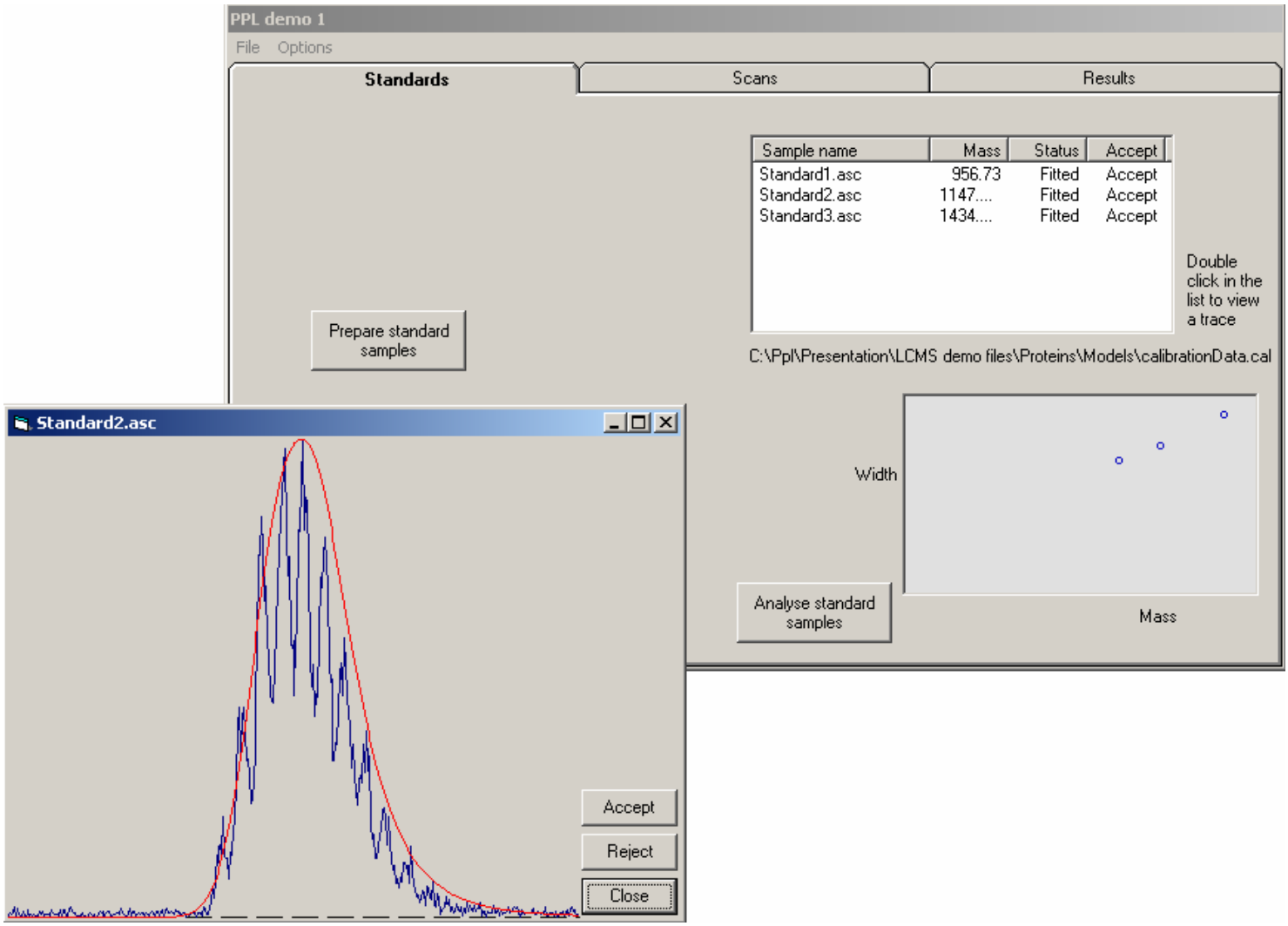

Figure 3. Modelling peaks to determine the way the peak width changes with m/z.

Setting Processing Parameters

The “Scans page” is for setting the input parameters used for processing the data. The chromatography options uses the conditions determined from Figure 2. The electrospray options allow data and output ranges to be set, along with the charge information. The minimum and maximum charges are set at their defaults, which are computed from the data and output ranges. The analysis section allows the deconvolved peak table for each block to be filtered at different confidence levels to remove noise. The panel in the analysis section shows the progress of the computation.

The minimum adjacent charges [MAC] is used to define the number of charges required to identify a mass. Setting it too high risks not reconstructing low masses where only a few charges may be present. An examination of the data shows that the intense “dirty” elution (scans 400-465) contains numerous peaks and that many arise from masses approximately ¼ of the main mass. There are also around 30 charges for the main mass (~29 kDa) and around 7 charges for the associated masses (~7-8 kDa). With so many ions at very similar RT, there will be numerous coincidental correlations that fit the data for MAC=3 was used to identify and reconstruct higher masses. Algebraic and other data reconstruction methods have the same problem. However, ***ReSpect***™ can take into account that higher masses will have more charges and MAC is set in accordance with the output mass range. The low and high limits for MAC represent the number of charges required for output mass range limits. For these data it is reasonable to set MAC from 3 (5 kDa) to 30 (50 kDa). The input parameters is shown in Figure 4 below.

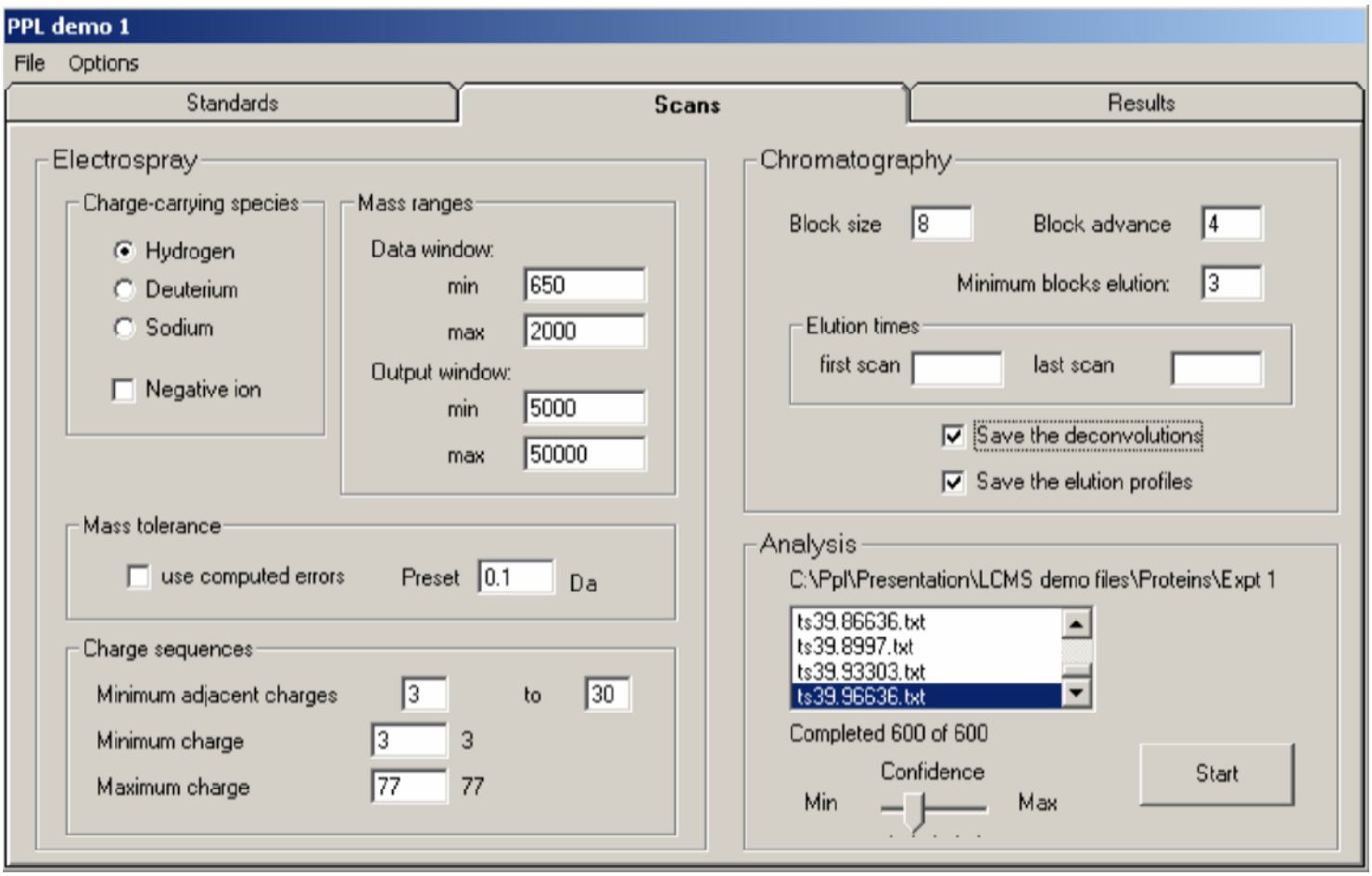

Figure 4. Input parameters for processing the data described here.

At the end of the deconvolution of each block the peak table is filtered according to the set confidence level and the charge deconvolution performed. Background ions are identified and removed and elution profiles computed along with their retention time and associated errors. Only masses that are present in at least the set minimum number of adjacent blocks are retained. The confidence of each retained mass is known so that the results table may be further filtered to remove obvious noise.

Results

At the end of the computation the results are presented in the “Results Page” as shown in Figure 5.

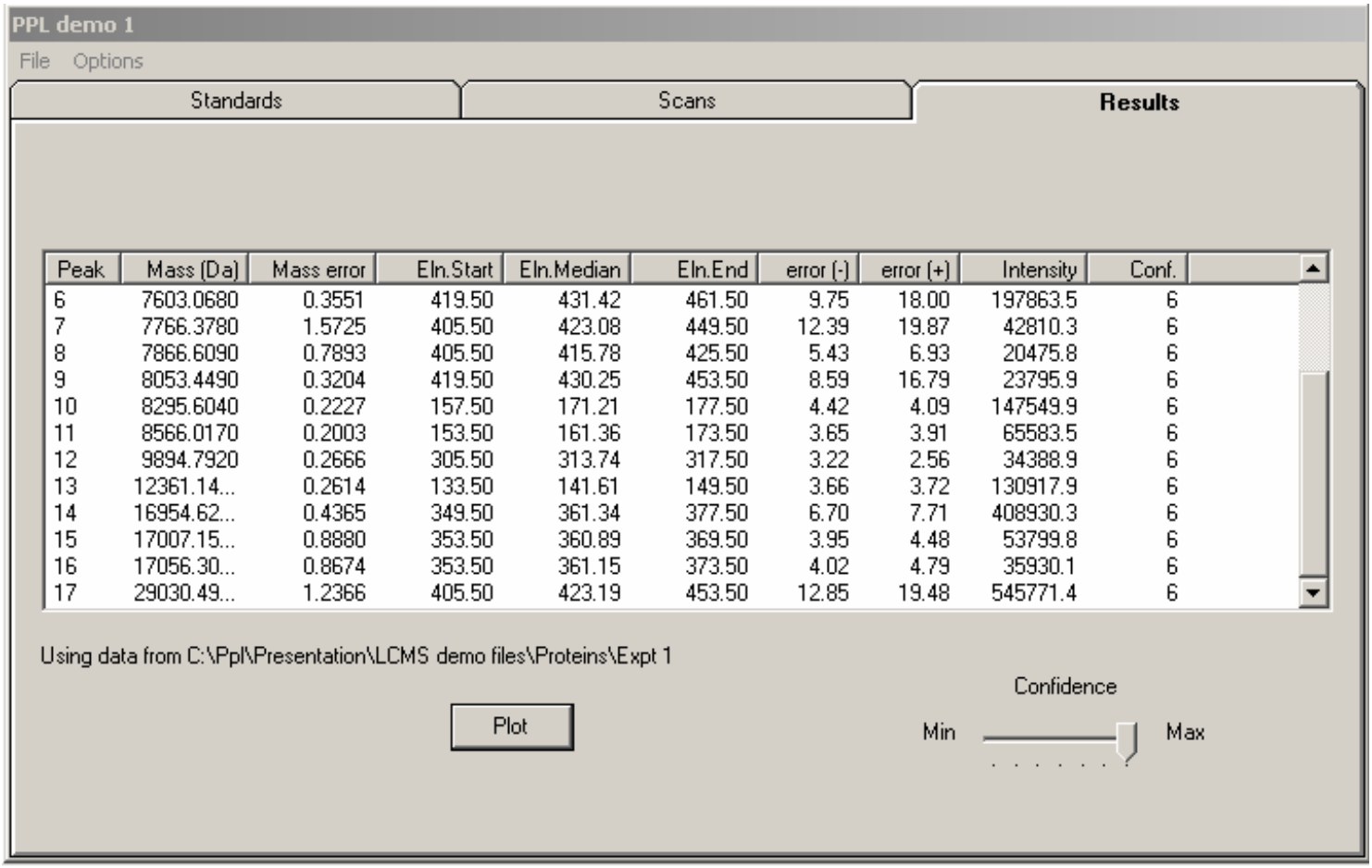

Figure 5. The “Results Page”.

The information contained in the peak table is the reconstructed zero-charge mass and its associated error, the elution median as scan number (or RT), its start and end along with RT errors. These are presented as separate (-) and (+) errors because elutions are rarely symmetrical. The reconstructed intensities and confidence levels are also output.

For these data, 55 masses are reconstructed. The confidence for each mass is known and the table may be filtered using the confidence slider. Some masses are of low confidence and the number reduces to 17 at the highest confidence level. Confidence levels are from 0 to 6 and respectively represent 0% (all peaks), 50%, 68% (1SD), 95% (2SD), 99% (3SD), 99.9% (4SD) and 99.99% (5SD).

Results peak tables may be loaded into Excel for formatting and sorting according to user requirements.

The result is very clean result and, as expected, a range of masses at ~¼ the main mass are apparent.

Results may also be presented as a 2-D plot. The data and results are compared in Figure 6 below.

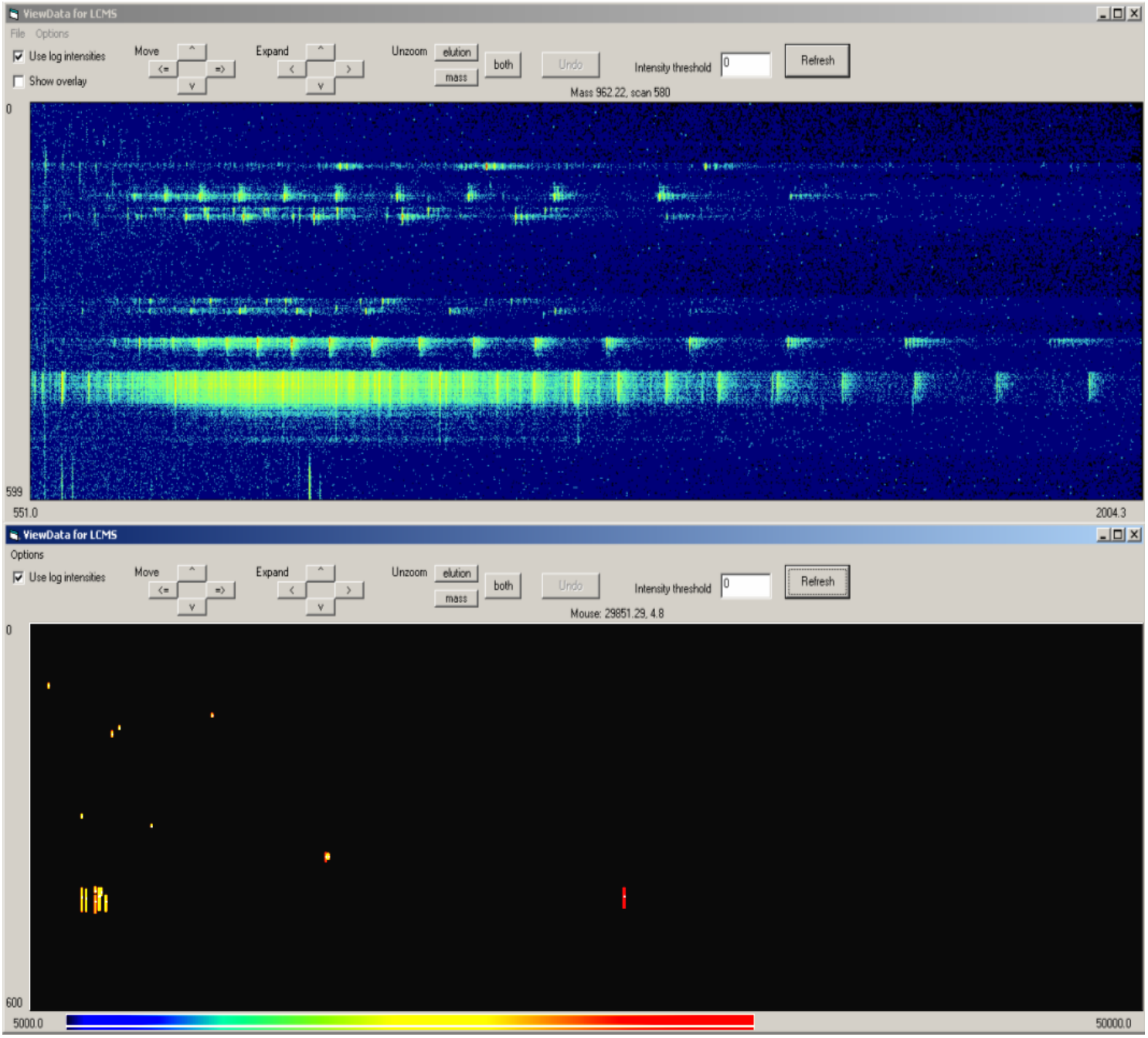

Figure 7. Data (top) and result (bottom) with minimum adjacent charges set from 3 to 30.

For these data the total processing time for a 3 GHz single processor P4 computer was 2 min. 3 sec. for the 20 minutes of data acquisition time. The processing time is therefore only about 10% of the acquisition time. Such speeds are made possible by optimising the ***ReSpect***™ algorithm and its LCMS interface. Given appropriate integration, the processing could be performed on-line so that results were available at the end of the experiment.

The full results at the highest confidence level are shown in the table below. Approximate masses for the components of the mixture and their expected intensities were provided. As can be seen from the table, all expected masses were reconstructed.

Results Table: Comparison of expected and reconstructed masses and intensities for masses with the highest confidence.

Approx expt. M	Exp Intens.	Peak No.	Found mass	Mass error	Eln start	Eln median	Eln end	Eln err (-)	Eln err (+)	Intens. (area)	Conf. level
5735	Medium	1	**5734.16**	0.18	85.5	95.57	105.5	3.87	3.97	99293	6
		2	7046.08	0.12	405.5	**423.93**	457.5	13.75	23.06	141420	6
7075	Weak	3	**7076.94**	0.23	293.5	299.12	305.5	3.43	3.08	38809	6
		4	7248.25	0.40	405.5	**425.84**	457.5	14.84	21.02	47164	6
		5	7603.06	0.57	401.5	**413.89**	419.5	6.60	4.56	86144	6
		6	7603.07	0.36	419.5	**431.42**	461.5	9.75	18.00	197864	6
		7	7766.38	1.57	405.5	**423.08**	449.5	12.39	19.87	42810	6
		8	7866.61	0.79	405.4	**415.78**	425.5	5.43	6.93	20476	6
		9	8053.45	0.32	419.5	**430.25**	453.5	8.59	16.79	23796	6
8295	Medium	10	**8295.60**	0.22	157.5	171.21	177.5	4.42	4.09	147550	6
8565	Weak	11	**8566.02**	0.20	153.5	161.36	173.5	3.65	3.91	65584	6
9895	Weak	12	**9894.79**	0.27	305.5	313.74	317.5	3.22	2.56	34389	6
12360	Medium	13	**12361.14**	0.26	133.5	141.61	149.5	3.66	3.72	130918	6
16955	Strong	14	**16954.62**	0.44	349.5	**361.34**	377.5	6.70	7.71	408930	6
		15	17007.15	0.89	353.5	**360.89**	369.5	3.95	4.48	53800	6
		16	17056.30	0.87	353.5	**361.15**	373.5	4.02	4.79	35930	6
29030	Strong	17	**29030.49**	1.24	405.5	**423.19**	453.5	12.85	19.48	545771	6

Found masses corresponding with the known proteins present are highlighted in green. Those reconstructed from the “dirty” protein are highlighted in orange. The myoglobin elution is accompanied by two weak peaks that are related (blue highlight).

Conclusions

The ***ReSpect***™ data reconstruction methodology has the following benefits over other methods:

1. It is very fast and can therefore operate in real-time.
2. Background ions are easily identified and removed.
3. Results may be presented at any user-chosen confidence level.