Positive Probability Ltd

Note M3: Deisotoping – HSA Protein Identification

Introduction

In this example, we investigate the ability of several manufacturers' methods with the PPL data reconstruction methodology for protein identification. The data are taken from the LCMS of a digest of HAS. The methodologies discussed are available in MassLynx and Analyst. They are: Waters MaxEnt, ABI Bayesian, ABI peak scoring and the PPL ***ReSpect***™ algorithm.

Data

Each individual scan of the LCMS run is noisy and sparsely populated but much of the noise is averaged on co-adding. However, mass errors will be much higher than those from higher quality data. Figure 1 compares a single scan with the co-added scans. Charges range from Z=1 to Z=4 and multi-charge deisotoping is required.

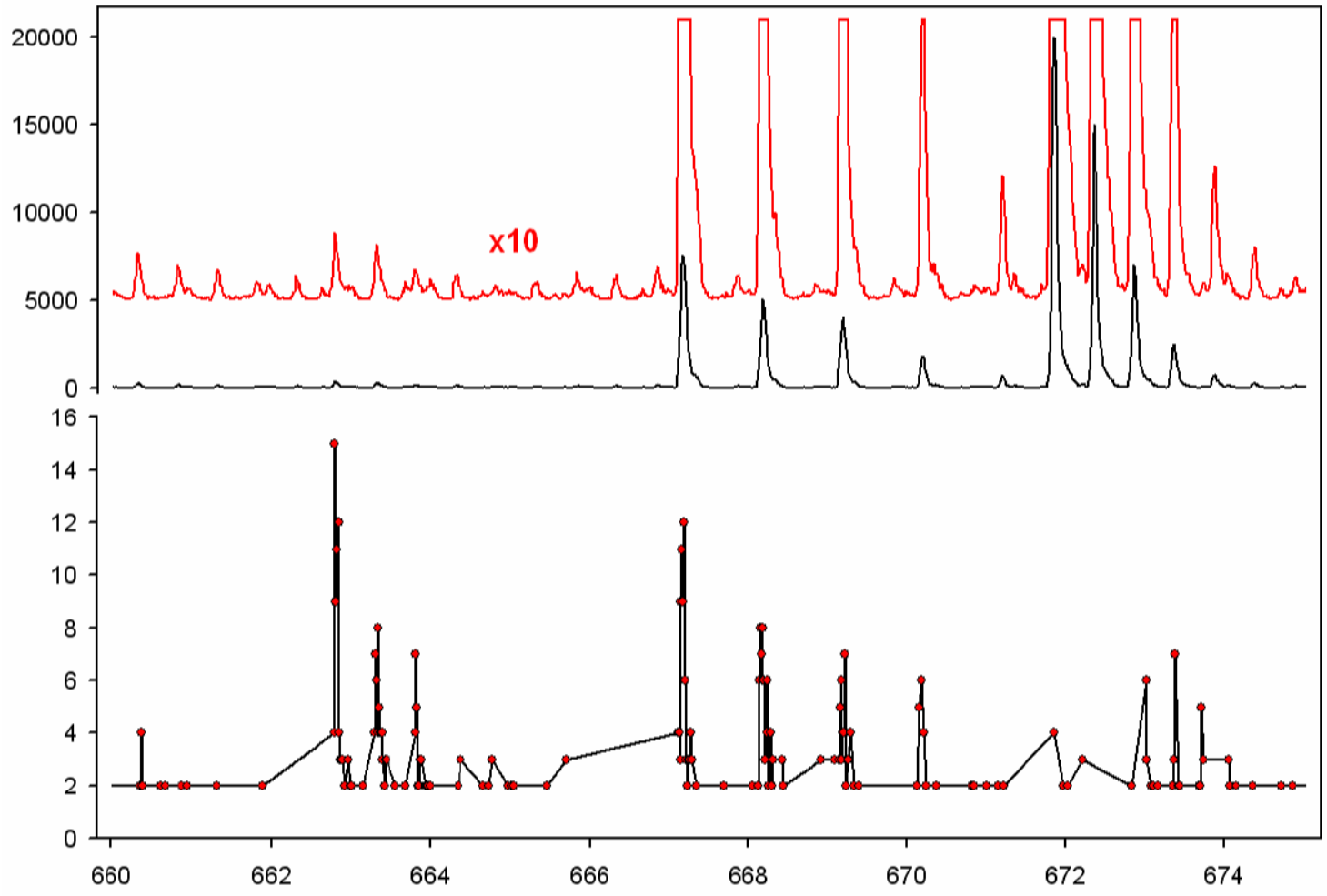

Figure 1. Comparison of a single scan (bottom) with the co-added scans (centre & top) for the HSA data

The data to be processed are 40 minutes of experiment time from a 2 hour run. All scans in the 40 minute window were co-added and the resulting spectrum is shown in Figure 2 below.

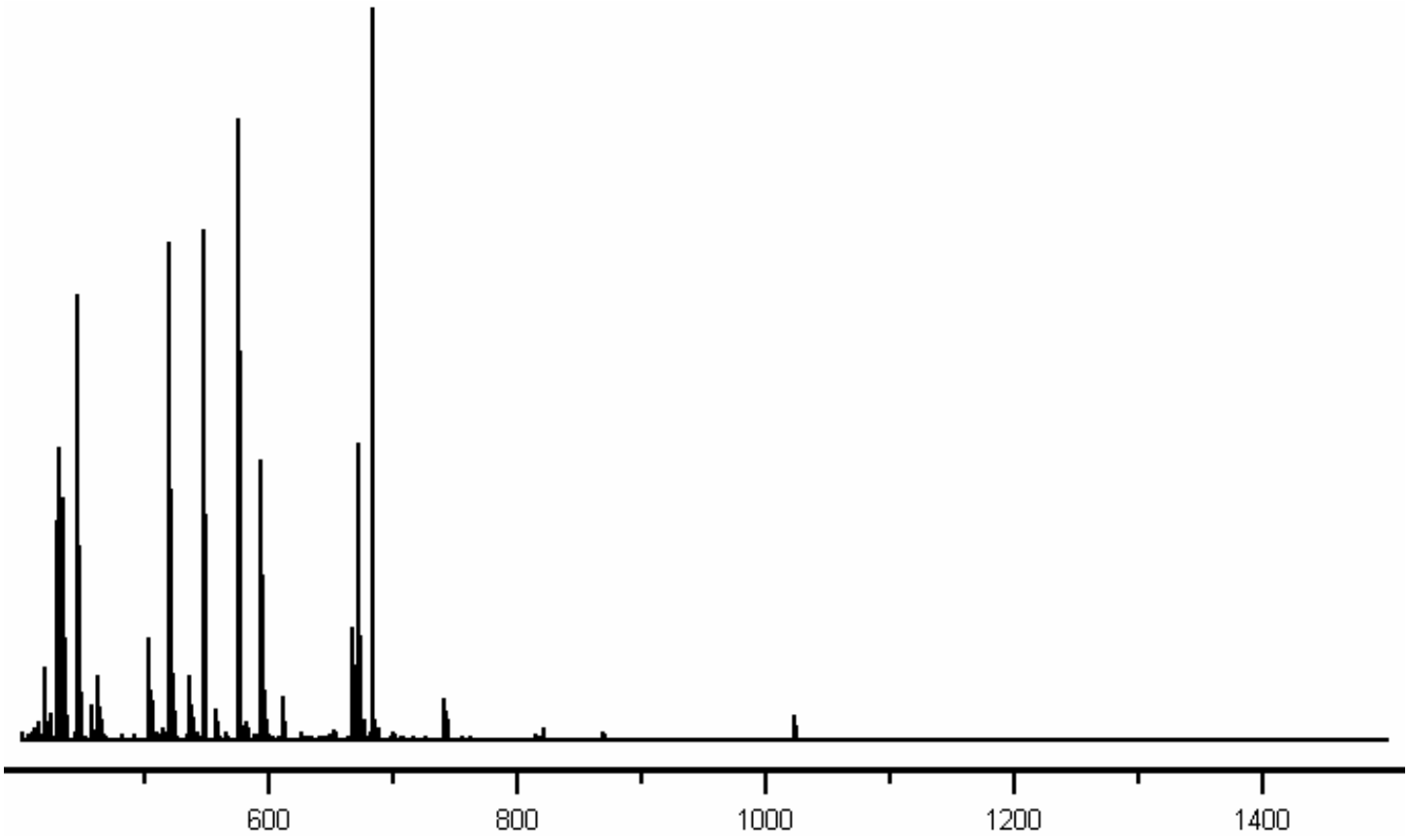

Figure 2. HSA data – the sum of 40 min of experiment time

Data Processing

The spectrum shown in Figure 2 was first baseline corrected. In this particular case the change in the underlying noise level is small and there is little benefit to be gained by taking the noise variation into account for the PPL data reconstruction methodology. The data were then centroided using each manufacturer's method and the PPL fast data reconstruction centroiding. Absolute peak intensities were used to compare the various methodologies.

Algebraic deisotoping assumes that there is no intensity error for each isotope peak. This places an extremely severe constraint on the fitting process and generates artefact peaks. The ***ReSpect***™–based deisotoping program therefore performs its fitting within both the noise level and the intensity errors. This freedom – absent for algebraic methods – ensures that there is positive evidence in the data for any reconstructed deisotoped m/z peak (or deisotoped to zero-charge) and that the results are free of artefact peaks. Of course, the applied empirical formula is an average and is therefore a compromise for any particular peptide. This applies to all methods and so peaks may be present in the result that arise from imperfect fitting. The empirical formula used for the PPL deisotoping was $C_6H_9N_{1.6}O_{1.75}$.

Of the methods explored here, Waters MaxEnt, ABI Bayesian and PPL ***ReSpect***™ are all non-linear data reconstruction methods. The ABI peak scoring method is algebraic.

The final peak tables were then used as the input to the Mascot search engine to identify the protein. Table 1 below compares the peptide masses identified by the various processing methods along with their mass errors.

Results

In Table 1 below, the methods are: ***PPL*** – Positive Probability, ***ME3*** – MassLynx (MaxEnt3), ***AN1*** – Analyst (Bayesian), ***AN2*** – Analyst (peak scoring). AN2 is known to be inferior to AN1 but it is much faster than the very slow Bayesian method.

Column headings and highlight are: ***MassTh*** shows the theoretical masses from a theoretical digestion. ***Pk*** is the peak number in decreasing intensity. ***AAE*** is the average absolute ppm error for the identified masses. The highlight shows which HSA peptide masses are found in the top 25 (green), 50 (orange) and 100 (pink) peaks and the totals are shown at the bottom of the table. It is somewhat surprising to note that the rather crude ABI peak scoring methods identifies significantly more peptides than the Waters MaxEnt method.

Table 1: Identified HSA Peptides

	PPL			ME3			AN1			AN2		
MassTh	**ppm**	**Int**	**Pk**	**ppm**	**Int**	**Pk**	**ppm**	**Int**	**Pk**	**ppm**	**Int**	**Pk**
410.2165	-85.2	7640	64				-95.1	7706	70			
430.2540	75.5	35029	24				83.2	44510	20	83.4	33933	16
447.1965	-23.5	12934	43	-17.8	10570	62						
462.2438	25.0	11031	49	34.7	9317	71	14.7	18163	38			
463.2101	-40.8	6857	70				-48.1	12273	54	-44.3	7933	48
508.3121	-20.2	7150	67	-30.4	12380	52						
515.3431	12.0	6227	75	20.2	6838	87				3.9	5852	60
521.2155				-23.4	10070	67						
714.4098	66.4	20718	36				68.4	22326	34	66.6	13734	31
1016.5291	-92.4	15619	39									
1148.6077	-64.6	523469	2	-72.2	440300	3	-80.0	483363	2	-73.1	316392	3
1156.6465	-25.0	4012	98									
1295.6973	-13.9	4436	87									
1341.6274	16.9	299455	5	17.5	239100	6	14.6	300227	5	13.8	196543	5
1547.6748	27.1	4304	88	87.8	6834	88	73.7	7889	69	83.6	6448	52
1638.9304	-90.4	498937	3				99.6	15671	45	99.0	10121	37
1839.9076	98.3	4478	86							95.9	3096	100
2044.0880	36.8	944065	1	21.0	757100	1	21.5	864788	1	20.8	601933	1
2088.7823							-95.5	5462	97			
2201.9939	32.0	4845	81				29.7	5476	95	31.3	4026	76
2514.1254	-64.3	4605	84				-62.4	6419	82			
AAE	39.5			46.6			45.5			51.6		
Top 25			**5**			**3**			**4**			**4**
Top 50			**9**			**3**			**7**			**7**
Top 100			**19**			**9**			**13**			**11**

HSA Search Results

The top 25, 50 and 100 peaks for the different methods were used as the input to the Mascot search engine and the results are shown in Table 2 using 100 ppm error limits. ***Hit*** is the number in the Mascot hit list of possible proteins and NF indicates that HSA was not found. ***Matched*** is the number of identified peptides and ***Coverage*** is the percentage of sequence covered by identified peptides. ***Mowse*** scores >75 are considered significant and are shown in green. Those <75 are considered ambiguous and are shown in red.

Table 2 – Search Results for HSA

Peaks		PPL	ME3	AN1	AN2
Top 25	Hit	1	NF	NF	NF
	Matched	5	-	-	-
	Coverage	10%			
	Mowse	36	-	-	-
Top 50	Hit	1	NF	NF	15
	Matched	9	-	-	7
	Coverage	14%			11%
	Mowse	46	-	-	34
Top 100	Hit	1	NF	1	8
	Matched	19	-	13	11
	Coverage	29%		21%	19%
	Mowse	98	-	51	40

Discussion

PPL identified more peptides with smaller mass errors than the other methods. All peaks for each method were used to calculate AAE, as there were so few common peptides. There is little to conclude from the recovered intensities other than ME3 and AN2 are generally lower than those of PPL and AN1.

Only PPL returned HSA as the top hit for all searches. HSA was not identified in any search for ME3 and only AN1 using the top 100 peaks found HSA as the top hit. Greater coverage and substantially higher Mowse scores are obtained in all cases using the PPL data reconstruction methodology.

Conclusions

The new methodology for centroiding and artefact-free deisotoping described here offers the following advantages over other established methods for protein identification from digest data:

1. Enhanced peptide identification.
2. Improved mass accuracy.
3. Greater coverage for the protein.
4. Improved Mowse scores.