Positive Probability Ltd

Note M2: Deisotoping – Chaperonin Protein Identification

Introduction

Identifying unknown proteins from the high resolution MS of enzymatic peptides digests will be reliable only if the peak picking and deisotoping algorithms provide reliable results. Deisotoping requires fitting theoretical isotope intensity distributions to the observed patterns. This is achieved using an empirical formula representative of the average for the species being studied and a very large number of simultaneous equations. In practice, there is an error bar for the m/z and intensity of each isotope peak but algebraic methods use a fixed error for m/z and no error in the intensities. They are therefore forced to generate many additional peaks and artefacts in order to balance the equations.

The solution is to use a data reconstruction technique that takes intensity uncertainties into account to provide an artefact-free result. Error bars are normally obtained from a spectrum deconvolution but these methods are slow compared with algebraic centroiding. In this example we describe:

1. A quantified centroiding method comparable in speed with traditional methods that provides a peak table with robust error assessments.
2. Confidence filtering of the peak table to minimise the effect of imperfect background corrections and for reporting only those deisotoped masses equal to or greater than the chosen confidence level.
3. A fast ***ReSpect***™–based data reconstruction deisotoping analysis of the filtered peak table that fits the data within the error bars for both MALDI (Z=1) and ESI (multi-charge).

The results from database searching to identify proteins using the new methodology described here are compared with the methods provided by instrument manufacturers.

Experimental

The MALDI spectrum of a tryptic digest of Chaperonin 60k was selected for investigation. The spectrum was baseline corrected (where appropriate) and then centroided and deisotoped using the programs available in MassLynx (MaxEnt3), Data Explorer, and Analyst. The spectra were also processed using the PPL methodology described here. The empirical formula used for the PPL deisotoping was $C_6H_9N_{1.6}O_{1.75}$. The final peak tables were then used as the input to the Mascot search engine to identify the protein.

PPL Method for Peak Identification

1. Quantified Centroids

All data reconstruction methods make use of a peak model and a noise level to generate data deconvolutions. The convolution of the sharp, deconvolved result with the model provides the reconstruction. The difference between this and the starting data is the misfit, a “noise channel” that contains most of the information required to compute the peak position and intensity error bars. Peak overlaps and intensity ratios are also taken into account when computing errors.

We have developed a very fast data reconstruction method that provides substantial S/N enhancement without broadening peaks. The reconstruction may then be efficiently centroided and the error bars computed. Irrelevant features may be removed from peak tables using significance filters – arbitrary thresholds do not apply. The certainty of peak positions and intensities is high for intense and isolated peaks but this certainty decreases with both decreasing S/N and for severely overlapped peaks.

By computing the way the underlying noise level changes across the data, the program can take into account any variation of the noise level and present the true statistical significance of peaks. As a result, weak features in regions of low noise are recovered with the same efficiency as those in regions of high noise. The raw data are shown Figure 1 below.

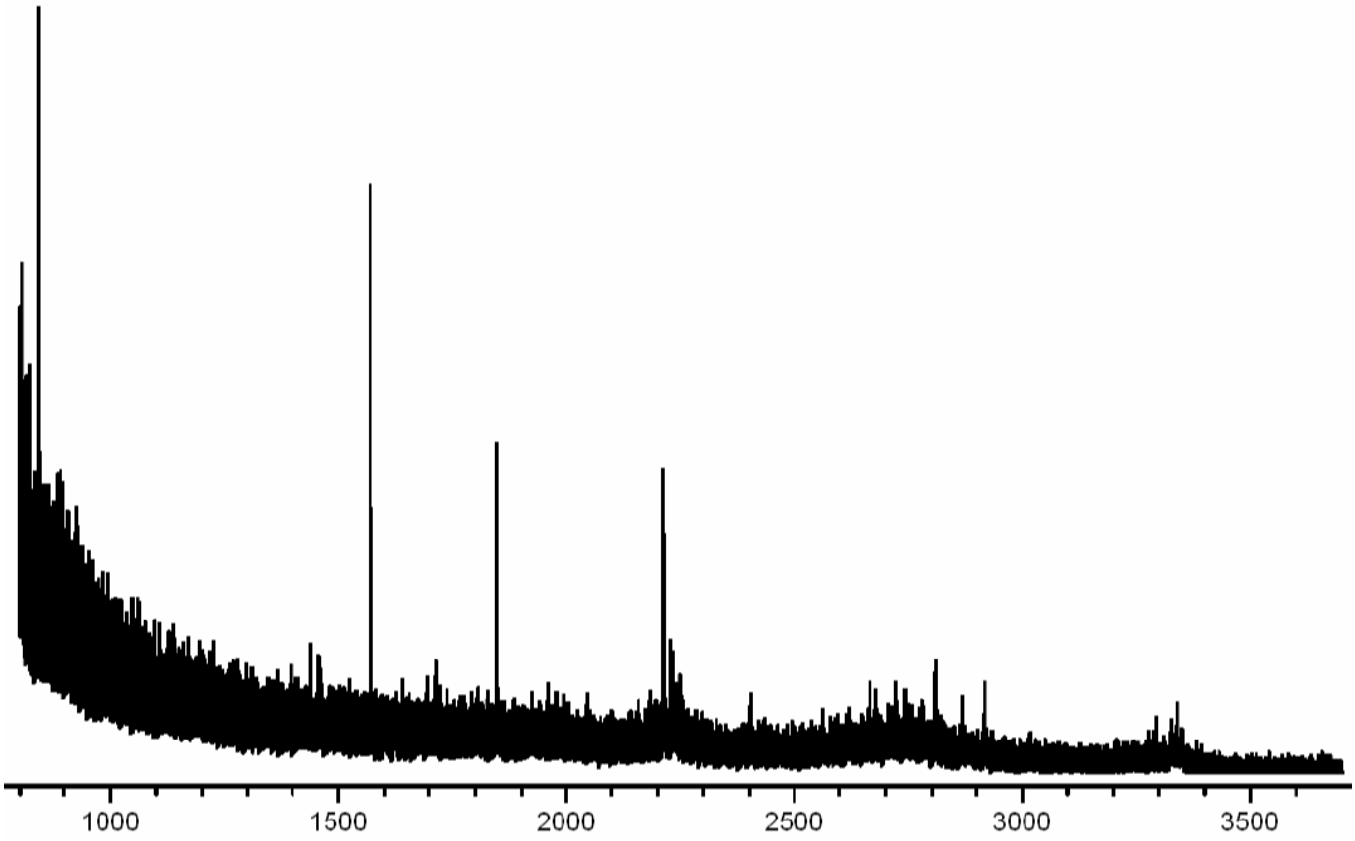

Figure 1. The MALDI-ToF spectrum of Chaperonin 60k digest. The varying noise level is obvious.

2. Confidence Filtering

Given ideal data, significance and confidence filters are directly equivalent so that a 2 s.d. significance filter is identical with a 95% confidence filter. However, the statistical significance of peaks will only be correct for a correctly set baseline. If the baseline is too low, both signals and noise features have more intensity than they should and their significance increases. Filtering a peak table by significance alone will retain insignificant and unwanted features that may have undesirable consequences for any subsequent processing. We have therefore developed a novel method for analysing peak tables and extracting the distribution of noise contained in them. It is then possible to compute confidence thresholds, again taking into account any varying noise level, that are independent of the background and conventional statistical significance levels. The PPL method used 1 s.d. and 68% confidence to remove obvious noise features from the peak table.

3. Deisotoping

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. We have therefore developed a ***ReSpect***™-based deisotoping program that performs its fitting within the noise level and intensity errors. This ensures that there is positive evidence in the data for any reconstructed deisotoped peak (or zero-charge mass) and that the results are free of artefact, other than those that arise from the applied empirical formula being an average and therefore a compromise for any particular peptide.

Results

Table 1 shows the peptide masses identified by the various processing methods along with their mass errors. The methods are: ***PPL*** – Positive Probability, ***ME3*** – MassLynx (MaxEnt3), ***DE*** – Data Explorer. PPL1 shows the intensities of the peptides relative to their S/N and PPL2 shows the list using absolute intensities for direct comparison with the other methods. ***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 common masses. The highlight shows which Chaperonin 60k 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. DE reports intensities on an arbitrary scale and these are shown in italic text.

Table 1 – Identified Chaperonin 60k Peptides

MassTh	PPL1 Error (ppm)	Int	Pk	PPL2 Error (ppm)	Int	Pk	ME3 Error (ppm)	Int	Pk	DE Error (ppm)	Int	Pk
843.5046	3.8	5481	56	3.7	17297	12						
875.4362	9.9	6159	45	9.9	17979	11	-32.0	6438	36	17.3	*195484*	25
1011.5223	-17.7	6567	37	-17.7	12537	19	-9.5	5989	49			
1045.5530	2.9	5638	53	2.9	9735	25	-5.4	7381	25	-3.6	*292679*	11
1201.6541	7.9	6270	42	8.0	7899	45				8.6	*205312*	22
1454.6399	-3.0	9333	25	-3.0	9529	31	5.4	5422	75	-0.4	*222141*	18
1567.8808	2.0	73479	2	2.0	66726	2	-11.4	45270	1	5.0	*1544180*	1
1711.7775	2.1	9549	23	2.1	8085	42	16.9	6508	35	-6.1	*203962*	24
1799.0101	2.1	6291	41	2.1	5094	84				-3.2	*131471*	75
1845.9194	-2.7	62049	3	-2.7	49614	3	14.7	34250	4	1.2	*1089940*	4
2042.9415	-17.2	11646	19	-17.2	8536	38						
2399.2782	0.8	6682	35							-2.8	*122331*	89
2402.2415	12.0	24889	10	12.0	16929	14	-7.1	5509	72	-3.5	*230883*	15
2739.4966	-11.0	12115	17	-11.0	8927	33				-8.3	*162374*	39
2867.3257	-4.1	29321	7	-4.1	18022	11	15.1	5583	70	-11.0	*216945*	19
3239.7244	-7.5	18950	14	-7.4	10122	24						
AAE	4.8			4.8			13.5			6.0		
Top 25			**9**			**9**			**3**			**9**
Top 50			**14**			**14**			**6**			**10**
Top 100			**16**			**15**			**9**			**12**

Note: *PPL and ML processed raw data. DE used smoothed data as this gave a substantially improved result.*

Chaperonin 60k Search Results

The top 25, 50 and 100 peaks for the different methods were input to the Mascot search engine and the results are shown in Table 2 using 25 and 50 ppm errors. ***Mowse*** scores >75 are considered significant (green) and those <75 ambiguous (red). Chaperonins originate from E.coli. Where E.coli is the 1st hit and Chaperonin 60k the 2nd, the hit is shown in green. NF = Chaperonin 60k not found. ***Matched*** is the number of identified peptides. ***Coverage*** is the percentage of sequence covered by identified peptides.

Table 2 – Search Results for Chaperonin 60k

Peaks		50 ppm PPL1	50 ppm PPL2	50 ppm ME3	50 ppm DE	25 ppm PPL1	25 ppm PPL2	25 ppm ME3	25 ppm DE
Top 25	Hit	1	1	NF	1	1	1	NF [a]	1
	Matched	9	9	-	9	9	9	-	9
	Coverage	31%	23%		21%	31%	23%		21%
	Mowse	111	96	-	96	114	98	-	98
Top 50	Hit	2	2	NF	2	2	2	8 [b]	2
	Matched	14	14	-	10	14	11	6	10
	Coverage	38%	36%		26%	38%	36%	13%	26%
	Mowse	142	133	-	82	145	136	32	84
Top 100	Hit	1	2	3 [c]	2	2	2	2	2
	Matched	16	15	9	13	16	15	9	12
	Coverage	40%	39%	23%	31%	40%	39%	23%	30%
	Mowse	121	103	43	85	125	106	44	80

[a] Hit 2 was Escherichia coli; [b] Hit 6 was Escherichia coli; [c] Hit 1 was Escherichia coli

Discussion

DE centroiding is very prone to noise. Without substantial smoothing only 9 peptides are identified in the top 100 peaks. Similarly, ME3 only identifies 9 peptides in the top 100 peaks due to artefact intensities. The PPL centroiding and deisotoping identify many more peptides. In addition, ppm errors are much reduced and more intensity is recovered for identified peptides (Table 1, PPL2 & ME3). No comparison can be made with the DE arbitrary intensities. Chaperonin 60k is always Hit 1 or 2 for PPL1, PPL2 and DE for Mascot searches. It is rarely found for ME3, giving low coverage and Mowse scores. Greater coverage and significant Mowse scores are obtained using the PPL methodology and coverage and Mowse scores are higher when the S/N is taken, highly diagnostic larger peptides gaining significance.

Conclusions

The new methodology for centroiding and artefact-free deisotoping described here offers the following advantages of 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.
5. The above are all improved further by taking the S/N into account.