Peak Shape Self-Modelling for Low Abundance Analytes in Complex Mixtures

P.Leopold[1], M.R.Alecio[2], L.Pannell[3], R.S.Ray[2], X.K.Zhang[4], A.G.Ferrig[2]

[1] BioAnalyte Inc., Portland, ME 04101;
[2] Positive Probability Ltd, Isleham, U.K., CB7 5RX;
[3] Mitchell Cancer Institute, University of South Alabama, Mobile AL 36688
[4] Genzyme Corporation, Framingham, MA 01701

Overview

As instruments are pushed to their limits and beyond, there is an increasing demand to extract as much information from data as possible. This will almost invariably require the application of advanced data reconstruction methods. For these methods to provide the best possible results demands that the model used to describe the data is of high quality. In this work we describe an iterative ReSpect-based method for modelling severely overlapped peaks.

Introduction

Applications of mass spectrometry to environmental and human health problems are facing twin demands:

1. Maximising information yield from analyses of raw mixtures of low-abundance analytes.

2. Simplifying data extraction for non-experts. Together with continuing demands for high throughput, mass accuracy and resolution, these specifications are driving innovation in data analysis, where parametric optimisation is a crucial part of maximising information but is far from simple to implement.

Almost all advanced data processing methods require a peak model so that some form of fitting procedure may be applied. We present a parametrically-self-optimizing deconvolution algorithm that resolves common ambiguities associated with complex, overlapping spectra in low S/N situations. We find the smallest number of peaks necessary to describe complex spectra, including spectra that include “tough-call” overlapping peaks. Results are validated with nearly degenerate peaks for peptides and multiply-charged glycoforms.

Methods

Peak modelling is ideally performed on a section of a scan that is deemed to comprise a single isolated shoulder-less peak. Where peaks overlap, conventional methods require the number of peaks to be specified, or at least constrained, since goodness-of-fit metrics by themselves do not select an optimal number of peaks. We have developed a method that fits the fewest peaks to the data within the noise. The principle involved is straightforward.

1. In this work, data peak profiles are defined by four parameters – left and right width at half height and left and right shape. Widths are defined in units of sampling intervals. Shape is defined by a number. Programmed shapes range from square wave (infinity but limited to 100) to Gaussian (2) to Lorentzian (1) to super-Lorentzian (0).

2. For noisy data a wide range of frequency components will be present from very low frequencies to those that describe high frequency components of the noise. The frequencies that describe genuine peaks will be of some intermediate value. It is clearly possible to model any of the frequencies present and the only feature about genuine signals that identifies them from everything else is that they are expected to be rather more intense than any of the noise components. Even so, peak widths and shapes can be very distorted at low S/N and the methodology can be made very robust by seeding the method with a crude estimate of the peak width. Mass spectrometry peak profiles tend to be Gaussian or, at least, have a high Gaussian content. The first trial model is therefore a symmetrical Gaussian profile of the seed width at half height.

3. The trial model, whether too wide or too narrow, is used to deconvolve the data as a means of obtaining the underlying sharp information. Depending on the extent of any peak overlap this will mean that:

   a) If the model is too narrow, genuine peaks may be split into more than one component.

b) If the model is too wide, overlapped peaks may not be resolved.

Either way, the important thing is to retain only those features with intensity significantly greater than the noise.

4. The pattern of significant features may now be used as a model in its own right for a second data deconvolution. The data are therefore collapsed into a peak profile that may be modelled using conventional methods to generate a second trial model. This completes the first iteration cycle.

5. The process is repeated until there is no significant change in any of the four peak profile parameters. The final self-optimised model is then used for a definitive data deconvolution to generate the most reliable and plausible peak table. The peak table contains peak positions and intensities along with their associated errors. The peak table may, of course, be used for other application such as deisotoping or charge deconvolutions.

Experimental

Spectra from LCMS ESI-ToF chromatograms of protein digests acquired on a Waters Q-ToF (Professor Lewis Pannell) were used for method development (Case 1). Glycoprotein data used for testing the methodology on severely overlapped glycoform peaks (Case 2) was also acquired on a Waters Q-ToF in a production environment. The data were extracted from injection electrospray run on intact glycoforms. The heterogeneous protein data used to check mass accuracy for severely overlapped peaks (Case 3) was acquired on a Waters Q-ToF (Kate Zhang).

All test spectra were baseline corrected before modelling with the methodology presented here.

Results

Case 1: Poorly resolved isotope clusters at low S/N from protein digest spectra were used as a means of developing and testing the method. One such cluster is shown in Figure 1.

A careful assessment of the data shows the peak width to be approximately 13 sampling intervals at half height. Trial symmetrical Gaussian models with half height widths of 8, 10, 16 and 20 sampling intervals were used as seeds for the method. Figures 2 and 3 illustrate just how inappropriate the models of 8 and 20 are. When the model is far too narrow, genuine peaks will be split in the deconvolved result even though the reconstructed data will still be a good fit to the data. Figure 2 shows the deconvolved result using the narrow model. When the model is far too wide, the deconvolved result may still appear to be acceptable. However, the reconstructed data will reflect the inappropriate model width and it will be a poor fit to the data as shown in Figure 3.

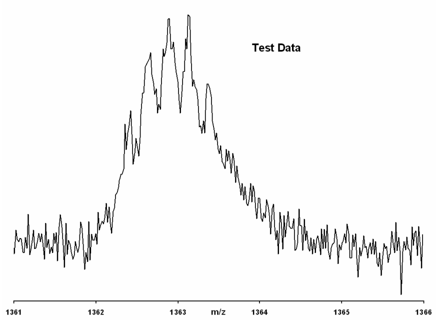

Figure 1. Poorly resolved low S/N isotope cluster from a protein digest

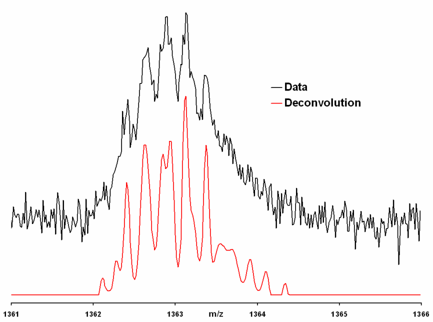

Figure 2. Deconvolved result (red) using a model with a width of 8

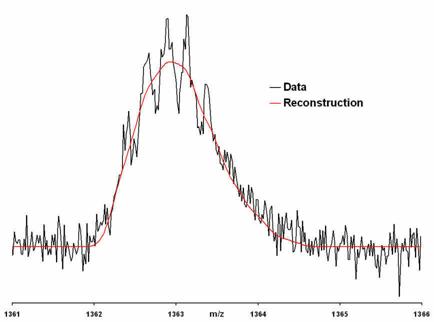

Figure 3. Data reconstruction (red) using a model with a width of 20

By using the described methodology, all four profile parameters are optimised together. The following figures show how the parameters change from the input values to their optimum values as the iterations progress. Figures 4-7 show how the width parameters change and Figures 8-11 show how the shape parameters change. All trial seed models had a Gaussian shape, value 2. LW=left width, RW=right width, LS=left shape and RS=right shape.

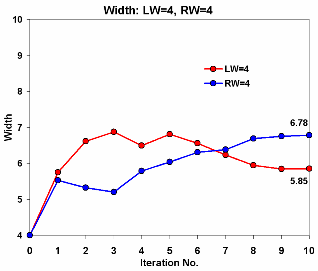

Figure 4. Trial width = 8

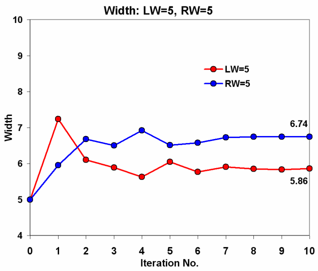

Figure 5. Trial width = 10

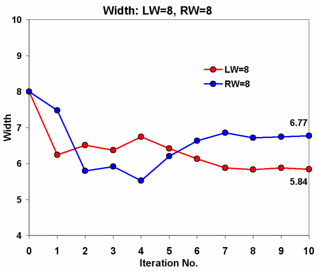

Figure 6. Trial width = 16

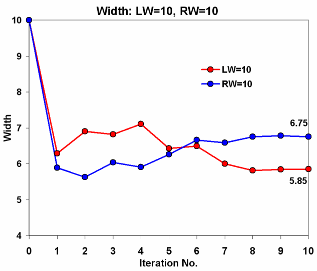

Figure 7. Trial width = 20

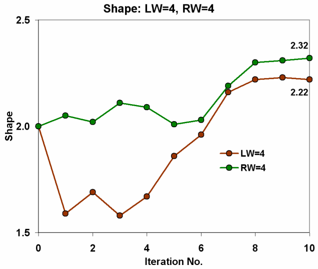

Figure 8. Trial width = 8

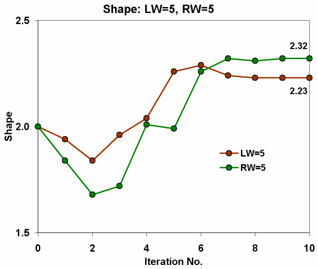

Figure 9. Trial width = 10

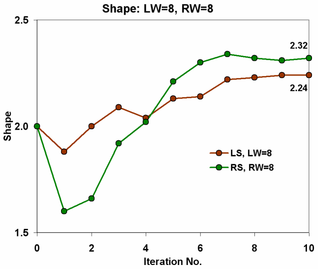

Figure 10. Trial width = 16

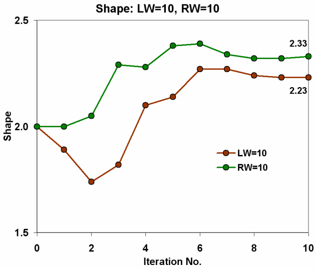

Figure 11. Trial width = 20

Regardless of the input peak width, all four peak shape parameters converge to virtually identical values, certainly to within 1%, after only 10 iterations. The improvement in the quality of both the deconvolved result and of the reconstruction of the data using the optimised model is very clear and is shown in Figure 12. The deconvolution is very clean and the reconstructed data are a noise-free version of the original data.The ideal model and the two extreme trial models are shown in Figure 13.

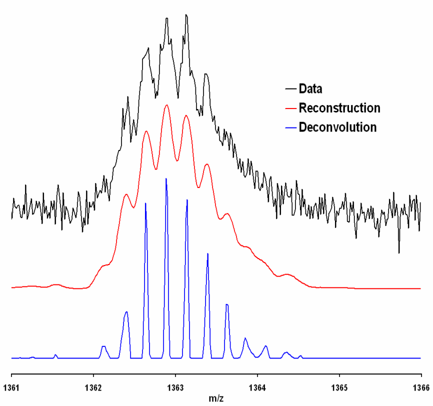

Figure 12. Reconstruction and deconvolution using ideal model

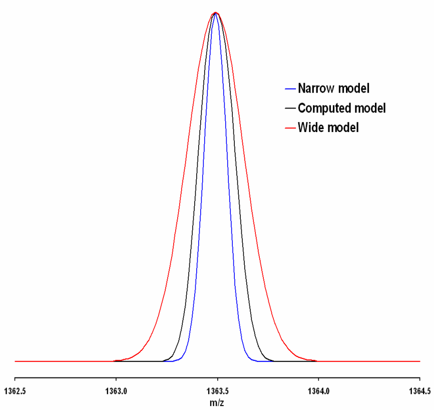

Figure 13. Two extreme models and the computed model

Case 2: The next test was to see how the methodology performed on seriously overlapped peaks. For this test we studied a ~150 kDa glycoprotein with four glycoforms separated by approx 140 Da, or 0.1%, surpassing the ToF resolution. The ToF MCP detector has marginal signal to noise characteristics for masses at this limit. The m/z peaks are nearly degenerate and there are no isolated peaks that may be used for model design. All that could be determined was that the peak width was somewhere in the order of 90 sampling intervals and that this appeared to be reasonably constant with m/z. Therefore, the requirement was to design a model from one of the strongest charges, deconvolve the data and then perform a charge deconvolution. Interpretation of the charge-deconvolved result would allow the quality of the processing to be objectively evaluated from the mass accuracy of the species identified.

The whole data and the Z=39 charge state (the most intense) are shown in Figures 14 and 15. In particular, Figure 15 shows how severe the peak overlap is and how poor the S/N is even on the most intense charge.

The modelling methodology was applied to the data shown in Figure 15 and all four profile parameters are optimised together. Figure 16 and 17 show how the parameters changed from the input values to their optimum values as the iterations progressed. Figure 16 shows how the width parameters changed and Figure 17 shows how the shape parameters changed.

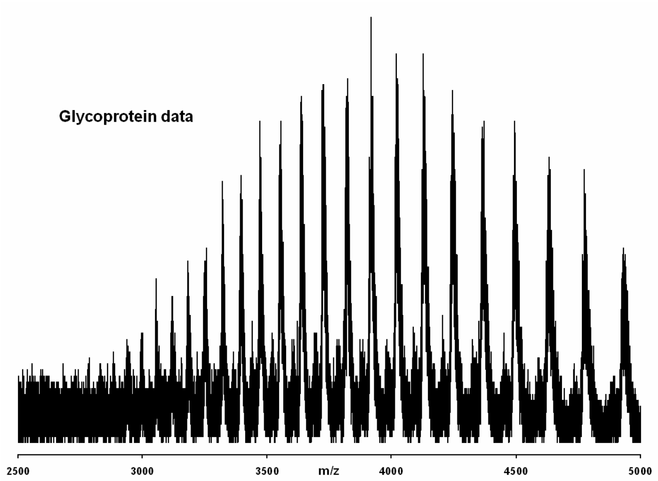

Figure 14. Glycoprotein data – full spectrum

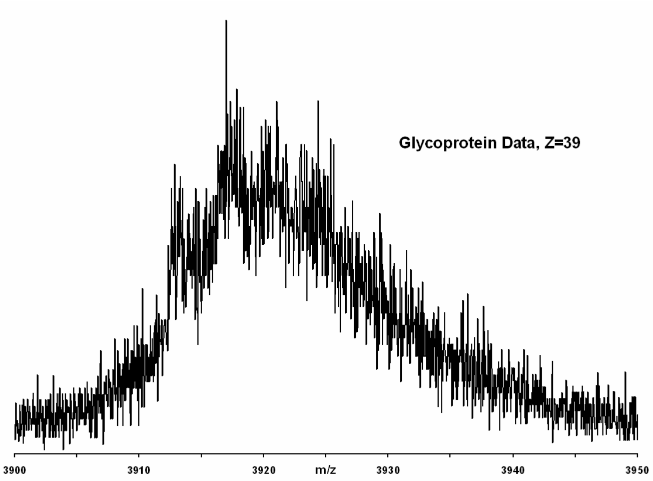

Figure 15. Z=39 for the glycoprotein data

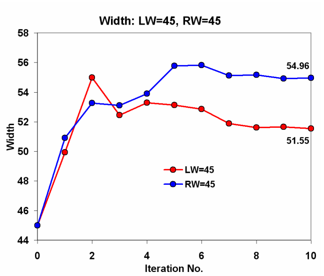

Figure 16. Trial width = 90

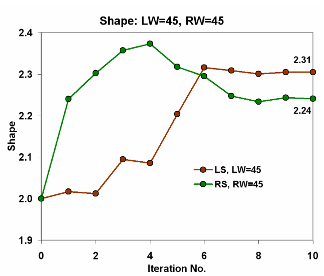

Figure 17. Trial width = 90

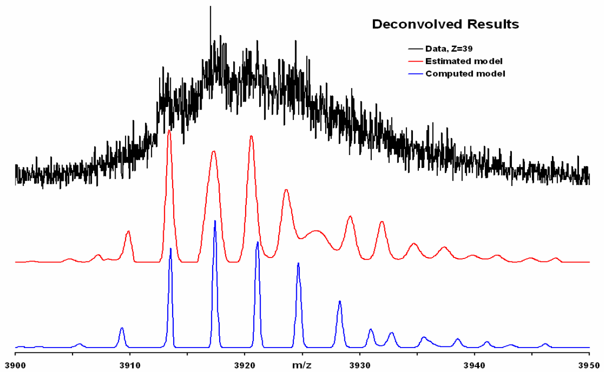

Figure 18. Deconvolved results for Z=39

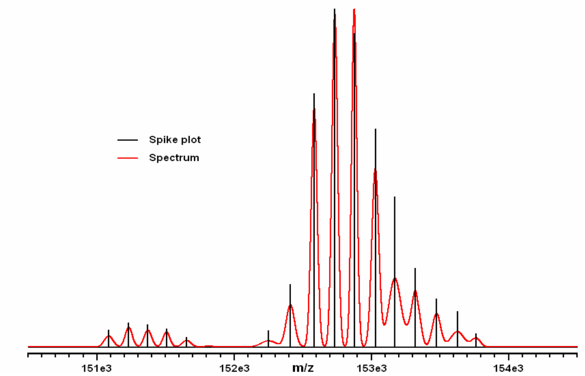

Figure 19. Charge-deconvolved result

Figure 18 shows the deconvolved results: red trace using the estimated model (left & right width = 45; left & right Gaussian shapes = 2) and the blue trace using the optimised model. The peak at m/z 3925 is split into two components using the estimated model and the peak at 3921 is shifted. ReSpect was then used to charge deconvolve the peak table. Figure 19 shows a horizontal expansion of the result. The peak widths (red trace) directly reflect the size of the mass errors. Peak intensities (areas) are correctly maintained and Table 1 shows the quantified results. Mass differences between adjacent glycoform masses are shown. Errors are ±1 standard deviation.

Table 1. Quantified Zero-charge Results

Mass	Mass error	Intensity	Mass Dif.	Dif Error
151085.9	32.1	3134		
151232.9	25.6	4434	147.0	41.1
151372.1	27.4	4061	139.2	37.5
151507.2	24.8	3298	135.0	36.9
151653.4	31.4	1821	146.2	40.0
151812.7	29.8	257	159.4	43.3
152250.2	54.3	2967		
152411.3	30.7	11427	161.1	62.4
152584.5	21.8	46103	173.3	37.7
152734.7	20.9	61484	150.2	30.2
152876.3	19.0	56963	141.6	28.2
153029.8	25.2	39652	153.5	31.6
153172.5	45.0	27321	142.8	51.6
153313.9	34.9	14430	141.4	57.0
153475.8	29.5	8741	161.8	45.7
153627.3	47.4	6486	151.6	55.8
153763.0	31.0	2460	135.6	56.7

Case 3: In a third trial we examined a heterogeneous protein of 60 kDa. The data and an expansion of the most intense charge are shown in Figures 20 & 21 before baseline correction. All peaks have other peaks in their wings and so no individual peak is suitable for modelling. In addition there is a small increase in peak width with increasing m/z. For this test we therefore decided to model the three most intense charges and then deconvolve all the data with the automatically generated model. The peak table could then be charge deconvolved and the result compared with the expected heterogeneity.

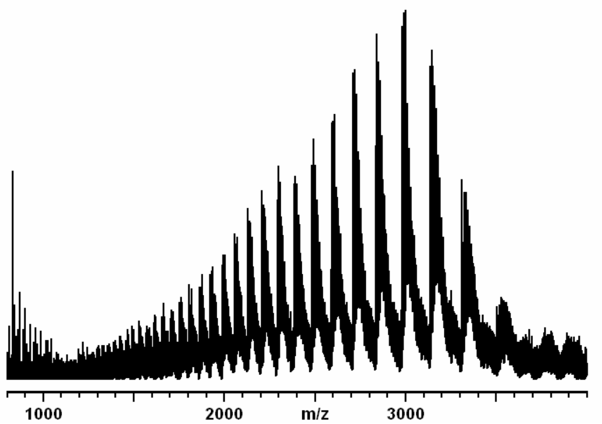

Figure 20. Heterogeneous protein data

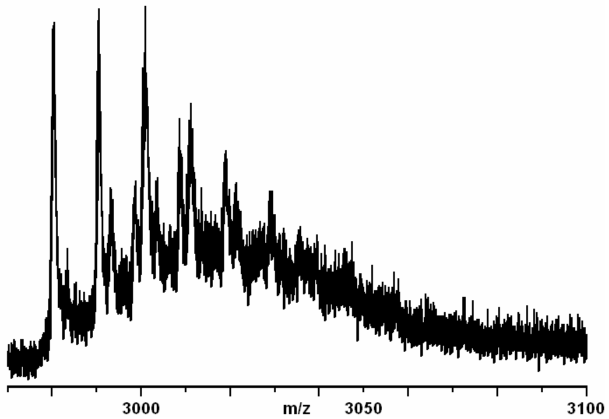

Figure 21. Most intense charge

Figures 22 & 23 show the way the peak profile parameters change as the iterations progressed. Figure 24 shows the charge deconvolved result and Table 2a shows the subsequent analysis.

In Table 2a the predicted masses were calculated from the constants in Table 2b below and the calibrated masses were adjusted so that their average difference with respect to the prediction was zero.

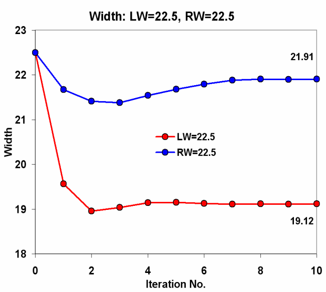

Figure 22. Trial width = 22.5

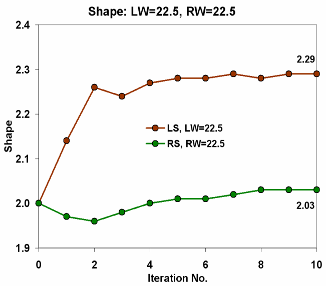

Figure 23. Trial width = 22.5

Table 2a. Assigned Masses

							Masses & Mass Differences					
Series 1				Ideal Model					Estimated Model			
Core	Fuc	Hex	HexNAc	Pred M	Calibrated	M Err	M Diff	Cal-Pred	Calibrated	M Err	M Diff	Cal-Pred
1	3	0	0	59584.3	59584.5	0.3		0.2	59584.4	1.6		-0.1
1	3	0	1	59787.5	59787.9	0.5	203.3	0.4	59787.6	3.7	203.2	-0.2
1	3	0	2	59990.7	59991.3	1.5	203.5	0.6	59993.7	3.6	206.1	2.4
1	3	0	3	60193.9	60193.6	2.3	202.3	-0.3	60196.5	7.1	202.7	2.8
1	3	0	4	60397.1	60398.9	4.4	205.3	1.8	60408.9	6.2	212.4	9.9
1	3	1	1	59949.6	59950.3	0.7		0.7	59950.5	2.9		0.2
1	3	1	2	60152.8	60152.8	0.9	202.5	0.0	60152.9	3.3	202.4	0.1
1	3	1	3	60356.0	60355.7	1.1	202.8	-0.4	60356.7	4.0	203.8	1.0
1	3	1	4	60559.2	60559.2	1.9	203.6	0.0	60560.1	5.6	203.4	0.9
1	3	1	5	60762.4	60763.0	3.5	203.8	0.6	60765.6	6.5	205.6	2.6
1	3	2	0	59908.6	59908.5	3.4		-0.1	59905.9	6.2		-2.5
1	3	2	1	60111.8	60110.6	3.5	203.2	-1.1	60109.1	4.8	203.2	-1.5
1	3	2	2	60315.0	60315.8	3.5	205.2	0.9	60314.8	5.7	205.7	-1.0
1	3	2	3	60518.2	60520.2	2.8	204.3	2.0	60519.9	5.4	205.1	-0.3
1	3	2	4	60721.4	60720.0	2.5	199.8	-1.3	60720.1	6.1	200.2	0.0
Series 2												
Core	Fuc	Hex	HexNAc	Pred M	Calibrated	M Err	M Diff	Cal-Pred	Calibrated	M Err	M Diff	Cal-Pred
1	2	0	2	59844.5	59844.5	0.6		0.0	59843.5	4.3		-1.0
1	2	0	3	60047.7	60048.8	1.5	204.3	1.0	60046.4	5.4	202.9	-2.3
1	2	0	4	60250.9	60251.2	0.9	202.4	0.2	60248.6	6.2	202.1	-2.6
1	2	1	2	60006.7	60007.1	2.3		0.4	60008.0	5.0		0.9
1	2	1	3	60209.9	60209.1	1.7	202.0	-0.7	60207.0	4.7	199.0	-2.2
1	2	1	4	60413.1	60412.1	3.8	202.9	-1.0	60408.9	6.2	201.9	-3.2
1	2	2	4	60575.2	60577.2	3.2		2.0	60574.7	4.4		-2.5
1	2	5	3	60858.4	60853.7	2.8		-4.8	60853.5	5.1		-0.1
1	2	5	4	61061.6	61061.1	4.4	207.4	-0.6	61059.3	5.4	205.7	-1.8
					SD			1.3				2.7
					Mean	2.2	203.5	0.0		5.0	203.8	0.0

Table 2a also shows the mass errors computed by ReSpect and the major repeating mass difference as a means of comparing results with those obtained using the estimated model. At the bottom of the table, relevant mean and SDs are shown.

Table 2b. Constants

Mass Constants	
55574.60	Protein
3571.26	4xCore
146.14	Fuc
162.14	Hex
203.20	HexNAc

The different highlight colours match those in Figure 24 for ease of identifying masses in the charge-deconvolved result.

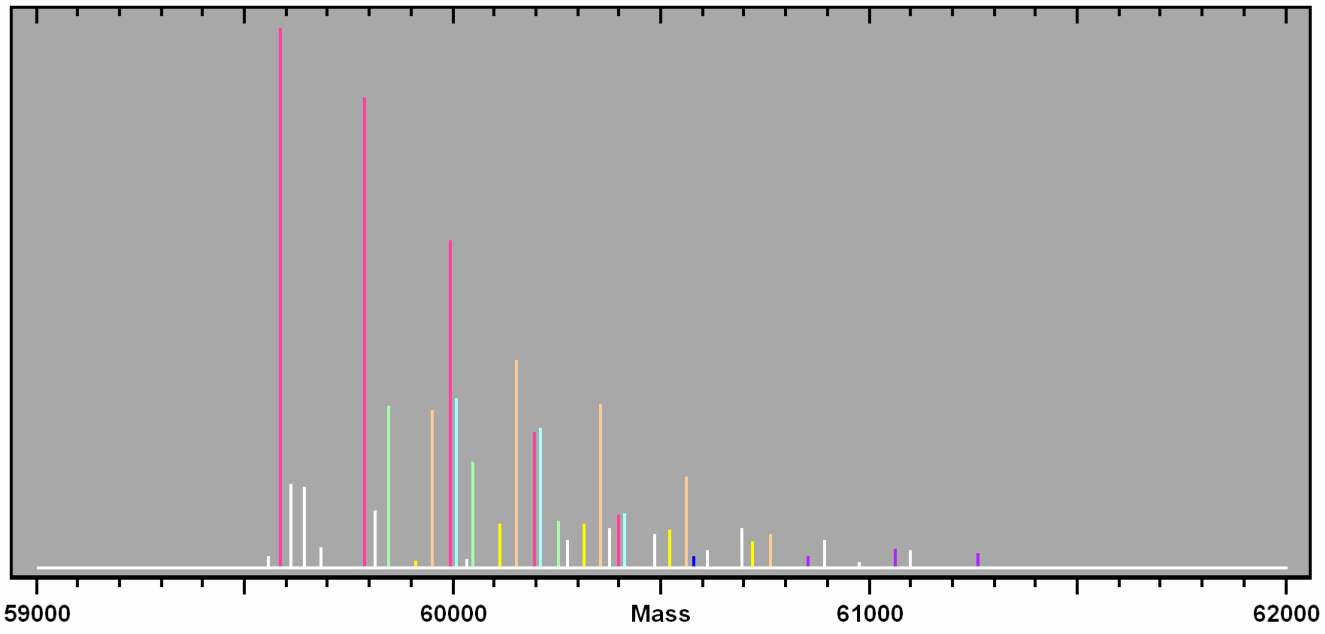

Figure 24. Zero-charge result using optimised model

The zero-charge results for both the optimised and the estimated model are almost identical and it is only the quantified results shown in Table 2a that differentiate between them. A lack of space prevents showing both zero-charge results. It is worth noting that the total number of peaks reconstructed by the charge deconvolution increased from 194 to 207 for the estimated and optimised models respectively due to the optimised model being narrower than the estimated model.

Discussion

When analysing complex ESI data, whether it be isotopically resolved digests or unresolved high mass intact proteins, the best results will only be obtained from data reconstruction methods when the applied model is a close fit to the data peak profiles.

It is generally necessary to perform some form of spectrum deconvolution to resolve peaks and to determine position and intensity errors. This allows a rigorous assessment of the results.

We have examined a range of examples as a means of assessing the new methodology.

As described earlier, noise contains a wide range of frequencies. However, the common practice of filtering data prior to any processing will destroy the original noise characteristics and enhance the lower frequencies, reducing the ability of any method to separate genuine signals from noise. For this reason, the method requires a seed width that should be as close as the user can estimate to the truth peak width. Even so, given a good S/N the method is sufficiently robust that any realistic estimate of the peak width will ensure that the method converges to the best possible peak profile parameters within the freedom given by the presence of any noise.

In the work presented here we have shown how the peak parameters change over 10 iterations. However, the number required for convergence is dependent on the noise level, the peak width and the severity of any peak overlaps. Therefore, in the continuing methodological development we are making the method self-monitoring so that convergence is reached when the reconstruction of the data fits the data within the noise.

Case 1: In this example, we have shown that the seed width for the new modelling methodology may be in serious error, certainly by >±30% but that the method converges to the correct peak profile parameters regardless of the input seed. This example also shows the effect of a narrow model on the deconvolved result and a wide model on the data reconstruction, emphasising the need for using a model that is at least close to the truth. The new self-optimising method achieves this.

Case 2: In this test we have taken a particularly poor S/N example where peak overlap is severe. It is clear from Figure 15 that it difficult to design a suitable model manually by manipulating a peak profile overlaid with the data. Hand developed models result in the detection of approximately four glycoforms. Traditional curve fitting routines are also complicated by the fact that the number of peaks and their approximate positions needs to be known in advance. In Figure 15 it is also clear that low frequencies are present so that the peaks close to m/z 3914 and 3921 each appear to be two peaks. Even at this poor S/N the method "homes in" on peak profile parameters that fit all the data rather than just two split peaks.

In the zero-charge result the mass errors are high due to the wide peak width and the excessive noise in the data. Even so, the model detects 17 glycoforms, a stunning result given the overall quality of the data. Moreover, the mass difference between the zero-charge glycoforms was not a constraint in the model. Nevertheless, the deviations from 140 Da are relatively small, the mean difference being 149.3±10.6. This finding provides an independent confirmation of the robustness and accuracy of the method. It also far outstrips the number of glycoforms detected using conventional data analysis.

Case 3: Here, we wanted an example that would demonstrate the improved mass accuracy that is obtained from the ReSpect data reconstruction methodology when the "ideal" model is used. For this it was necessary to use data of reasonable S/N but where many peaks were severely overlapped. In this case it was possible to estimate the peak width to better than 10% and so the estimated model was actually quite close to the self-optimised one. Even so, the number of peaks in the spectrum deconvolution peak table was dramatically higher using the slightly narrow estimated model.

Because the ReSpect data reconstruction algorithm analyses all peaks and their errors, it is not surprising that the charge-deconvolved results for both models were very similar and it is necessary to look at the result statistics. In Table 2a the average of the computed mass errors for the optimised model are ±2.2 Da. This is less than half the value of ±5.0 Da for the estimated model. However, the most appropriate test is to compare the standard deviation for the differences between the known and reconstructed masses. The values for the optimised and estimated models are respectively ±1.3 Da and ±2.7 Da. The mass accuracy has therefore improved by just over a factor of 2 by using the optimised model.

It should be noted that there are a number of masses on the zero-charge result that are not identified by the presented analysis – the white spikes in Figure 24. In fact almost all of these have been identified but there is insufficient space to present a detailed analysis.

Conclusions

The three tests presented here demonstrate that:

- Spectrum deconvolutions and reconstructions of the data are greatly improved by using optimised models from the new modelling methodology (***Case 1***).
- The new modelling methodology is robust even for very low S/N data in which peaks are severely overlapped and where low frequency noise is apparent (***Case 2***).
- There is a significant improvement in mass accuracy using self-optimised models from the new method even when an estimated model has very similar peak profile parameters to the optimised model (***Case 3***).
- The self-optimising method successfully optimises all four peak profile parameters, namely the left and right widths and their shapes.

Time trials have shown that model optimisation takes typically 2-5 seconds. This is because only small segments of the data are required for modelling.