Positive Probability Ltd

Note P5: Resolving Overlapped peaks – MS DNA Sequencing

Introduction

Traditional data processing involving filters are mathematically simple linear methods. They are generally fast to compute and commonplace in manufacturers' software. Data reconstruction methods are iterative, non-linear and computationally intensive. They are relatively time-consuming to perform but the additional information that is recovered can be dramatic and will often outweigh the computational time penalty.

What is required is a method that does not broaden peaks when the S/N is enhanced and does not introduce noise when the resolution is enhanced. In addition, it should be possible to quantify the results and provide realistic assessments of position and intensity errors. The peak profile can be determined from the data and it is only necessary to devise a fitting method that provides peak positions and intensities. Data reconstruction methods are well suited to large-scale problems of this type and the ***ReSpect***™ data reconstruction technique has been designed to mathematically reconstruct the data to within the overall noise level of any feature. Naturally, the method can only recover the information that is present in the data.

It is important to remember that there is never any attempt to operate directly on the data. Instead, the method always works forwards, reconstructing the information in the data. The only input is a peak profile – the Model – which must adequately match the profile of the data peaks.

The example described here is used to illustrate the benefit of the ***ReSpect***™ data reconstruction methodology for extracting the underlying detail contained in data.

Data

The data described here are a small region of a TOF-MS DNA sequencing experiment. For this particular sample, the sequence contains three sites of degeneracy where the DNA synthesis products varied at three particular sites. One site has a G/T split, the second another G/T and the third a C/T split. The higher mass peaks are therefore clusters of 6 major peaks corresponding to variations that contain combinations of C, T and G.

The aim was to unambiguously assign mass differences to the presence of specific bases so that sequences could be determined. It was therefore important to obtain accurate masses for all overlapped peaks.

Data Processing

In order to obtain the best possible result, the baseline was first corrected using the ***Nadir***™ baseline correction program (Baseline3). The peak width was estimated to be close to 8 sampling intervals and this was used as the only input, the other inputs (Feature Width and Degree of Fit) being left at their default values. The peak width was found to be almost constant over the range of the data and so the ***Sleuth***™ program was then run in standard mode but using the best resolution option.

The starting data and baseline corrected result are shown in Figure 1 below.

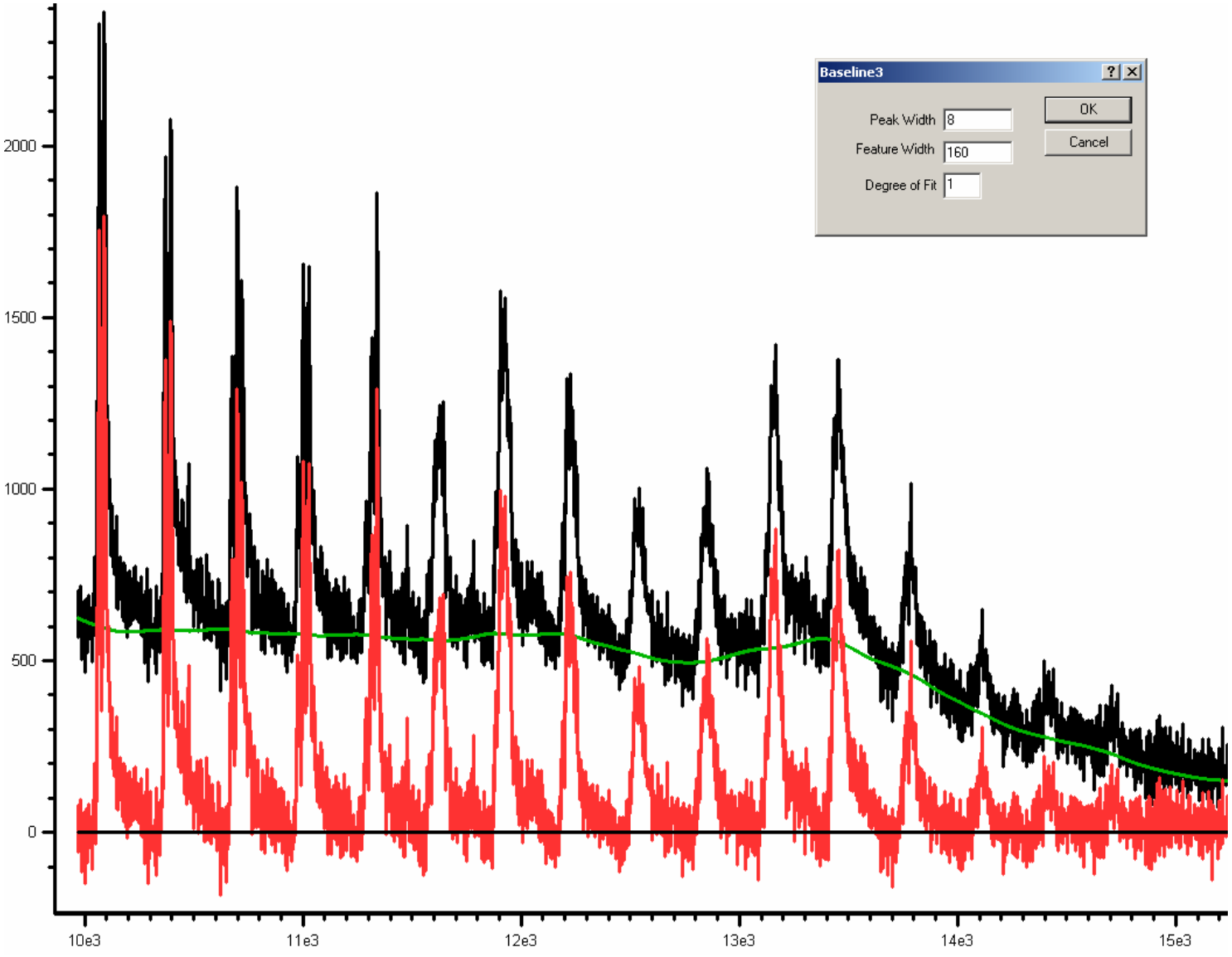

Figure 1. Data (black), computed baseline (green) and corrected result (red).

The model for the deconvolution was automatically determined from the first peak in the data using the ***Profile***™ program.

Results

At the end of the deconvolution three results are available – the deconvolved result, the data reconstruction and the misfit which is the difference between the data and its reconstruction. Figure 2 shows the detail for the clusters of interest from m/z 13000-14000 Da. The traces are: Data (red), Reconstruction (blue) and Misfit (gold). For clarity, the deconvolution is shown separately in Figure 3.

The second and third peaks of each cluster are two severely overlapped peaks. The severely overlapped peaks have been cleanly resolved in each cluster so that there are now 6 peaks and not 4.

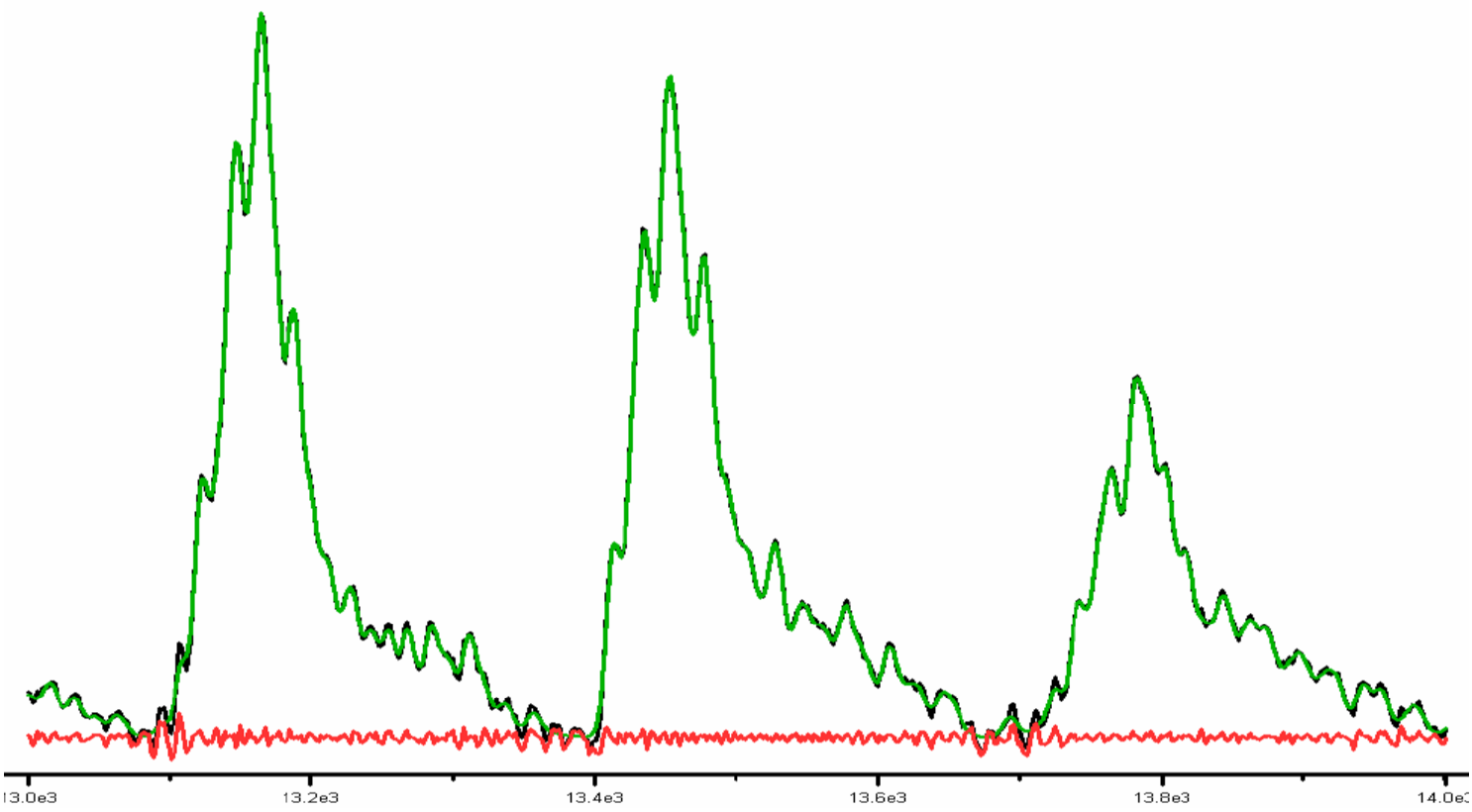

Figure 2. Baseline corrected data (black), Reconstruction (green) and Misfit (red).

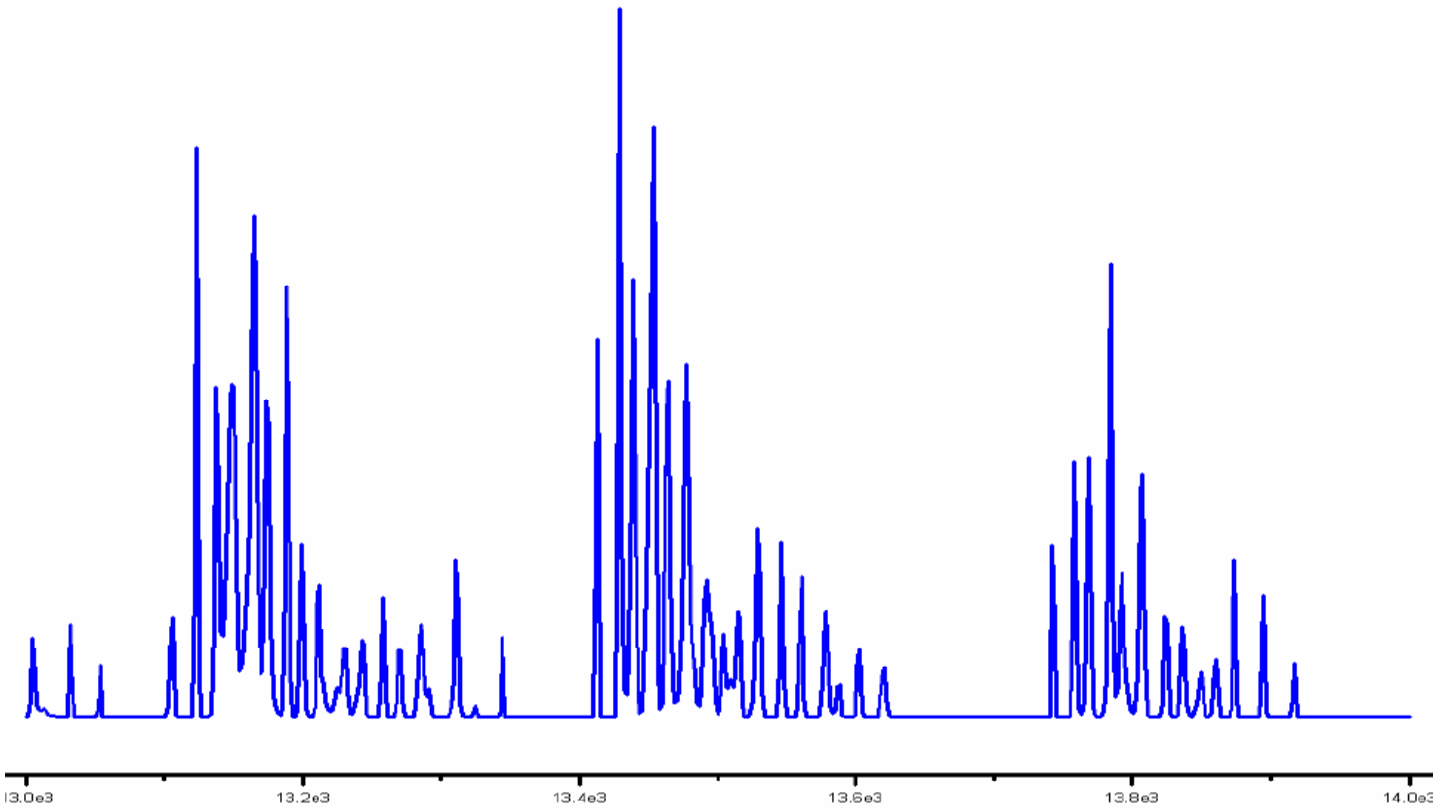

Figure 3. Deconvolved result.

The table below shows the found m/z values for the result. Since the mass difference between T & G is 25.01 and C & T is 15.01, it is possible to predict the mass differences from the first peak in each cluster.

The green highlight shows the relevant peaks. Yellow highlight shows the theoretical separations of each peak from the first peak of each cluster. The found differences are within the number of standard deviations shown and conform to statistical expectation – e.g. for 15 measurements the expectation is that 1 in 3 will be outside 1 SD but within 2 SD.

Peak	m/z	Err	Intensity	Err	Th dif	Found	Err	SD	Ave dif
133	**13123.21**	0.44	2972	281					
134	**13138.47**	0.54	3075	330	**15.01**	**15.25**	**0.69**	**1**	**15.96**
135	**13148.73**	0.38	6678	422	**25.01**	**25.51**	**0.58**	**1**	**25.67**
136	**13163.72**	0.33	7556	383	**40.02**	**40.51**	**0.55**	**1**	**40.59**
137	**13174.70**	0.51	3826	374	**50.02**	**51.49**	**0.67**	**2**	**50.91**
138	**13188.82**	0.50	3679	381	**65.04**	**65.60**	**0.66**	**1**	**64.70**
147	**13412.90**	0.80	2175	454					
148	**13429.64**	0.51	4272	446		**16.74**	**0.95**	**2**	
149	**13438.77**	0.62	3885	488		**25.87**	**1.01**	**1**	
150	**13452.64**	0.41	7171	522		**39.73**	**0.90**	**1**	
151	**13464.08**	0.55	3718	393		**51.18**	**0.97**	**2**	
152	**13477.70**	0.41	4775	353		**64.80**	**0.90**	**1**	
160	**13742.56**	0.89	1047	269					
161	**13758.45**	0.74	1843	350		**15.90**	**1.16**	**1**	
162	**13768.18**	0.79	2144	398		**25.62**	**1.19**	**1**	
163	**13784.09**	0.79	3242	510		**41.53**	**1.19**	**2**	
164	**13792.61**	1.12	1683	543		**50.06**	**1.44**	**1**	
165	**13806.27**	0.56	2273	275		**63.71**	**1.06**	**2**	

Conclusions

A ***Sleuth***™ deconvolution provides a dramatic improvement in both S/N and resolving power for severely overlapped peaks. In addition, fully quantified results may be obtained.