Fast and Quantitative Analysis of Data for Investigating the Heterogeneity of Intact Glycoproteins by ESI-MS

Kate Zhang [1], Robert Alecio [2], Stuart Ray [2], John Thomas [1] and Tony Ferrige [2].
[1] Genzyme Corp., Framingham, MA, [2] Positive Probability Ltd, Isleham, U.K.

Overview

Aim: To compare related probability-based methods for charge deconvolution including a novel methodology for the unambiguous analysis of highly complex ESI-MS data of large, heterogeneous glycoprotein samples.

Methods: The methods compared were Bayesian transformations, maximum entropy and the novel ***ReSpect***™ algorithm.

Results: The ***ReSpect***™algorithm provided much cleaner and more highly resolved results that allowed unambiguous analysis of the data.

Introduction

The investigation of the heterogeneity of intact proteins larger than 20 kDa has always been difficult by ESI-MS due to instrument resolution limits. With the latest generation of instruments (e.g. Qq-TOF), the resolution now exceeds 10,000 for MW less than 20 kDa, provides a possible means of characterizing intact proteins. However, the characterization of larger glycoproteins, ~65 kDa is still a challenge, since the traditional deconvolution algorithms often fail to resolve complex features in the data even when the ion intensity of the proteins is respectable. In the present study, data for human recombinant glycoproteins, which have MW of ~65 kDa, were acquired by QSTAR and the data then processed using several charge deconvolution methods in an attempt to obtain detailed information on their glycosylation.

The Algorithms

a) Bayesian Transformation

Advantage: Probability-based method. Superior to algebraic transformations.

Disadvantages: Slow. No resolution enhancement. No error assessments. Prone to produce artefacts.

b) Maximum Entropy

Advantages: Mathematically plausible results often free of artefacts. Substantial gains in resolution.

Disadvantages: Very slow. Works best with a section of the ESI-MS envelope because calibration errors reduce result quality. Extraneous signals - solvents and chemical noise - allow unwanted correlations and artifacts. The methods require a random number seed on which the quantified errors are dependent.

c) ReSpect™

Advantages: Very fast. Gives the most plausible results based on the physics of the experiment. Results are virtually free of artifacts. Accommodates varying noise levels. Provides highly resolved zero-charge spectra and a single set of robust error assessments.

Disadvantages: Slower than algebraic methods.

The work described here is aimed at comparing the performance of Bayesian transformations, maximum entropy and the new ***ReSpect***™ method for obtaining zero-charge spectra from complex ESI-MS data.

Experimental Conditions & Raw Data

The glycoprotein samples A and B were dissolved in methanol/water (50/50) with 0.5% acetic acid at a concentration of 10 pmol/µl and infused into QSTAR at a flow rate of 2 µl/min. The spectra were acquired using MCA mode. The data for the two samples were acquired under identical conditions. Glycoprotein A has a molecular weight around 65 kDa and four N-linked glycosylation sites. Glycoprotein B is the recombinant form of the human glycoprotein A. Probably due to different buffer conditions, the S/N for the acquired spectrum of Glycoprotein A was inferior to that of Glycoprotein B. The processing parameters determined for Glycoprotein B were therefore used for processing both samples.The enhanced resolution of the QSTAR instrument gave a clear separation for many of the different combinations of glycoforms on the proteins. However, a close examination of the spectra showed that some peaks were broader than others, suggesting the presence of severely overlapped peaks. These data therefore represent a severe test for the different charge deconvolution methods.

Data Processing

1. Data Preparation

The different methods were presented with the same baseline corrected data so that the comparison was as objective as possible. The method employed used the novel Nadir baseline correction algorithm provided by Positive Probability Limited. Therefore, any differences between the results could not be attributed to different baseline corrections. The full spectra and the corresponding baseline corrected spectra are shown in Figures 1 & 2. Figure 3 shows the quality of the computed baseline for a single charge state ($[M+20H]^{20+}$) for Glycoprotein B. The peaks at m/z 3001, 3011 and 3021 are clearly broader and are indicative of severe peak overlap.

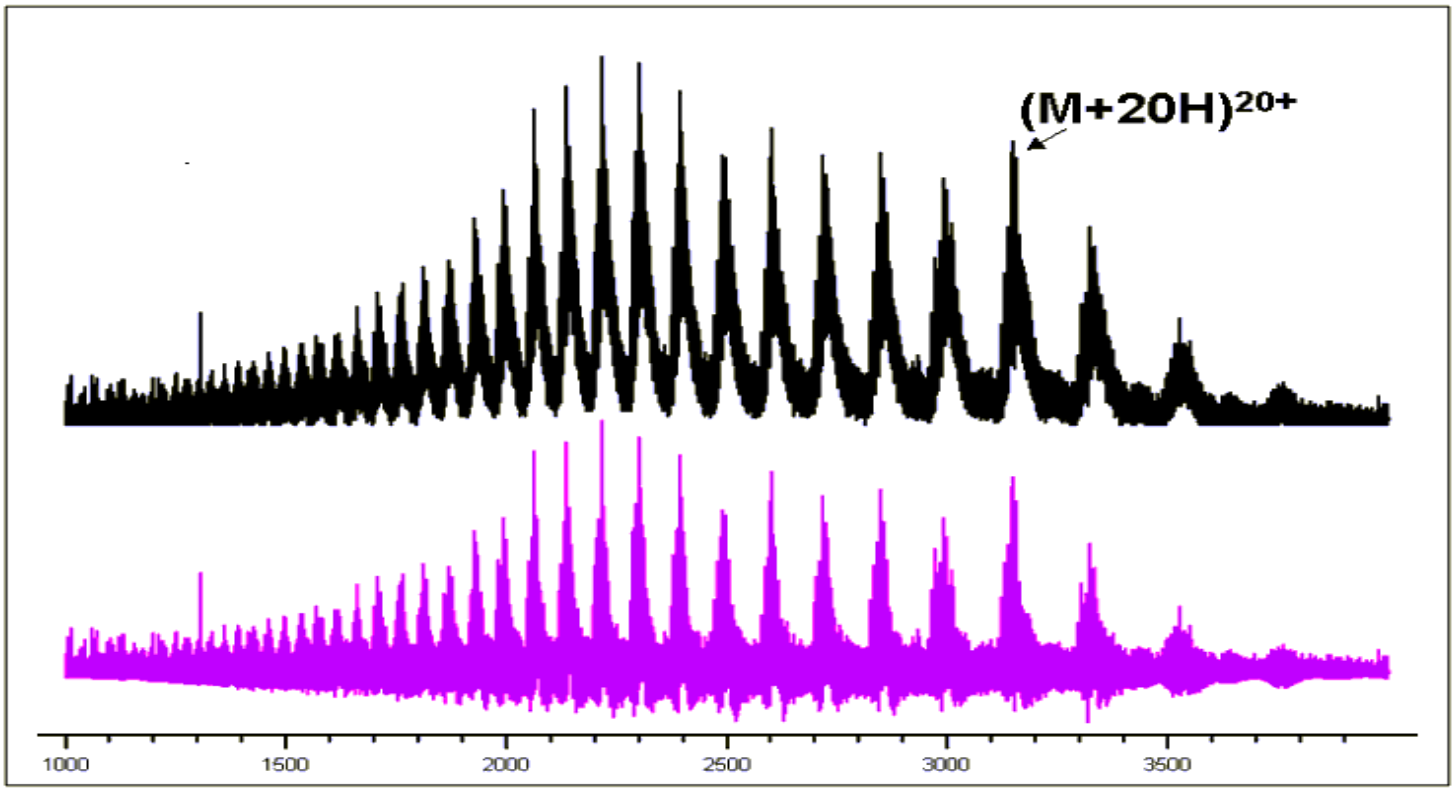

Figure 1: *Top*: Glycoprotein A raw data; *Bottom*: Baseline corrected with ***Nadir***™.

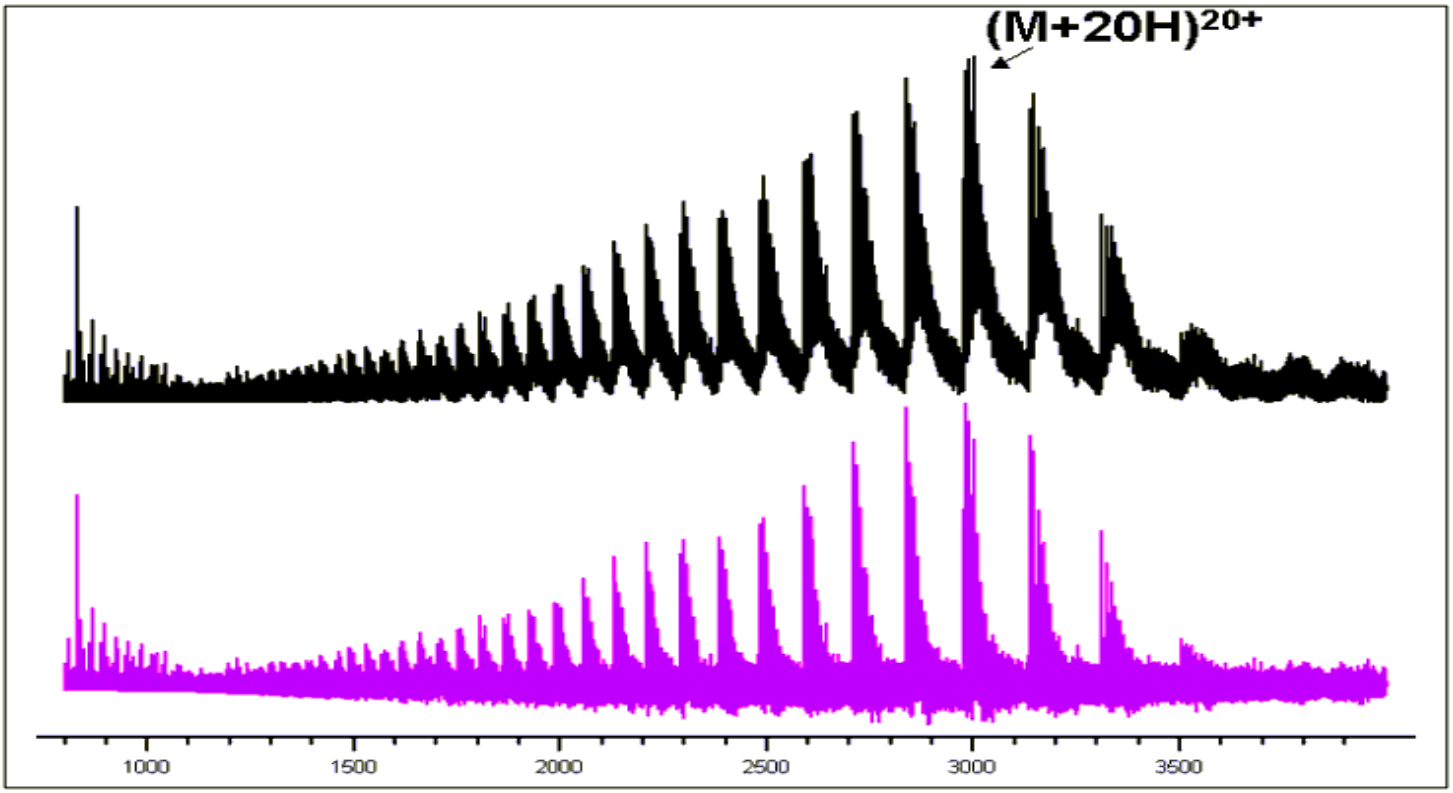

Figure 2: *Top*: Glycoprotein B raw data; *Bottom*: Baseline corrected with ***Nadir***™.

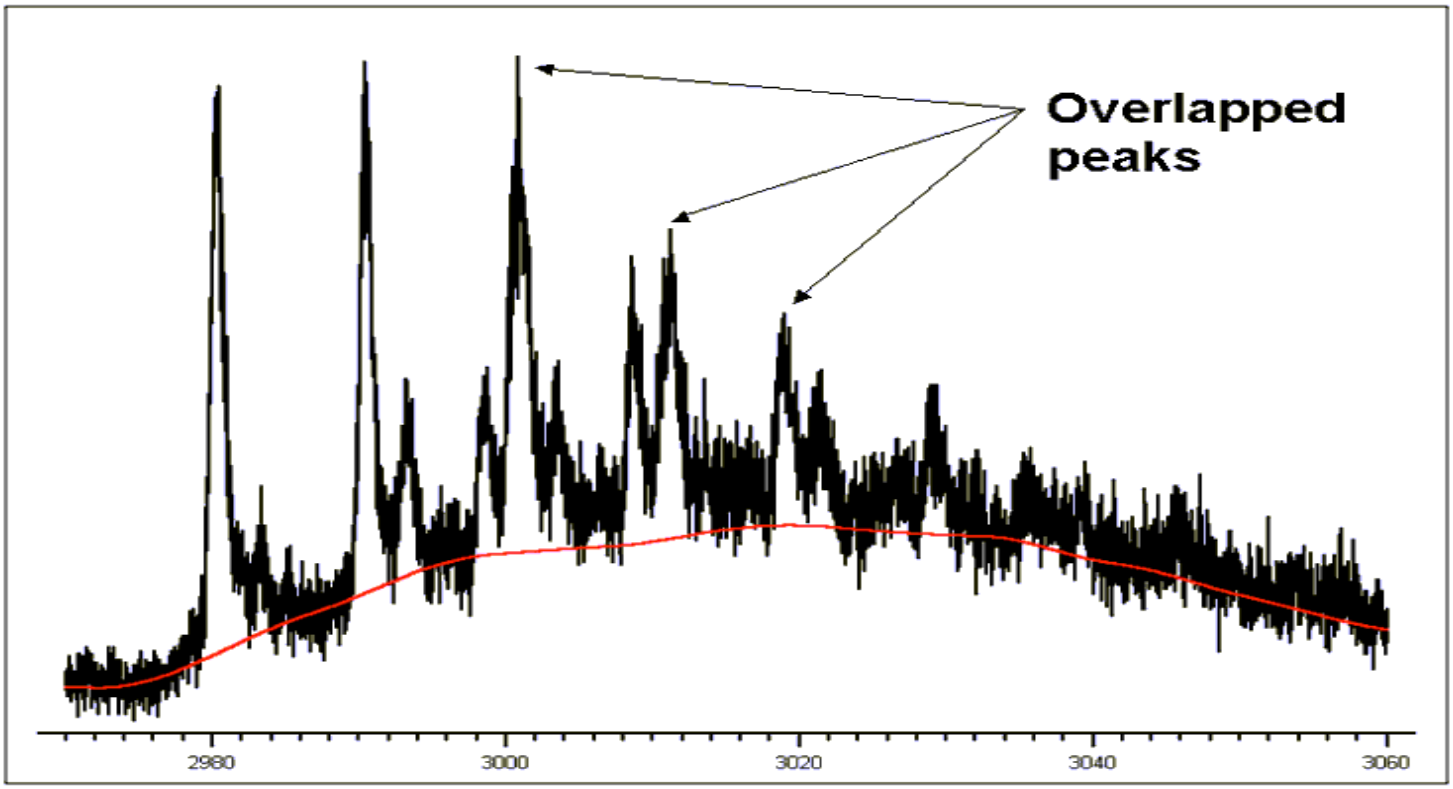

Figure 3: Black: Glycoprotein B $(M+20H)^{20+}$; Red: Baseline computed by ***Nadir***™.
Note: The peaks at m/z 3001, 3011 & 3021 are broad, indicating overlapped peaks.

2. Charge Deconvolution with Peak Model

The charge deconvolutions were performed on both the whole data (m/z ~800-4000) using an output mass range of 50-70 kDa and on a reduced range from m/z 2025-3450 with an output range of 58-62 kDa. These experiments were performed to determine how the input parameters affected the results.

Bayesian transformation: Does not require a user-set model.

Maximum entropy: The peak model width was measured from the first peak of the most intense clusters. Output resolution was set to 2 Da/point so that the computation would not be excessively long.

ReSpect: This used the same model width as for maximum entropy but with the shape parameters measured from the data included. The method first performs a spectral deconvolution that is then analyzed to compute the most probable zero-charge spectrum, along with the quantified mass errors. Output resolution limits do not apply.

3. Zero-charge Results

Figure 4 compares the zero-charge results for Glycoprotein A for the different methods using all the data and an output mass range of 50-70 kDa. Expansions of the results for Glycoproteins A & B are compared in Figures 5 & 6 respectively. Also shown are the peak clusters of $[M+20H]^{20+}$ for each sample, to demonstrate the evidence for the found masses in the raw data.

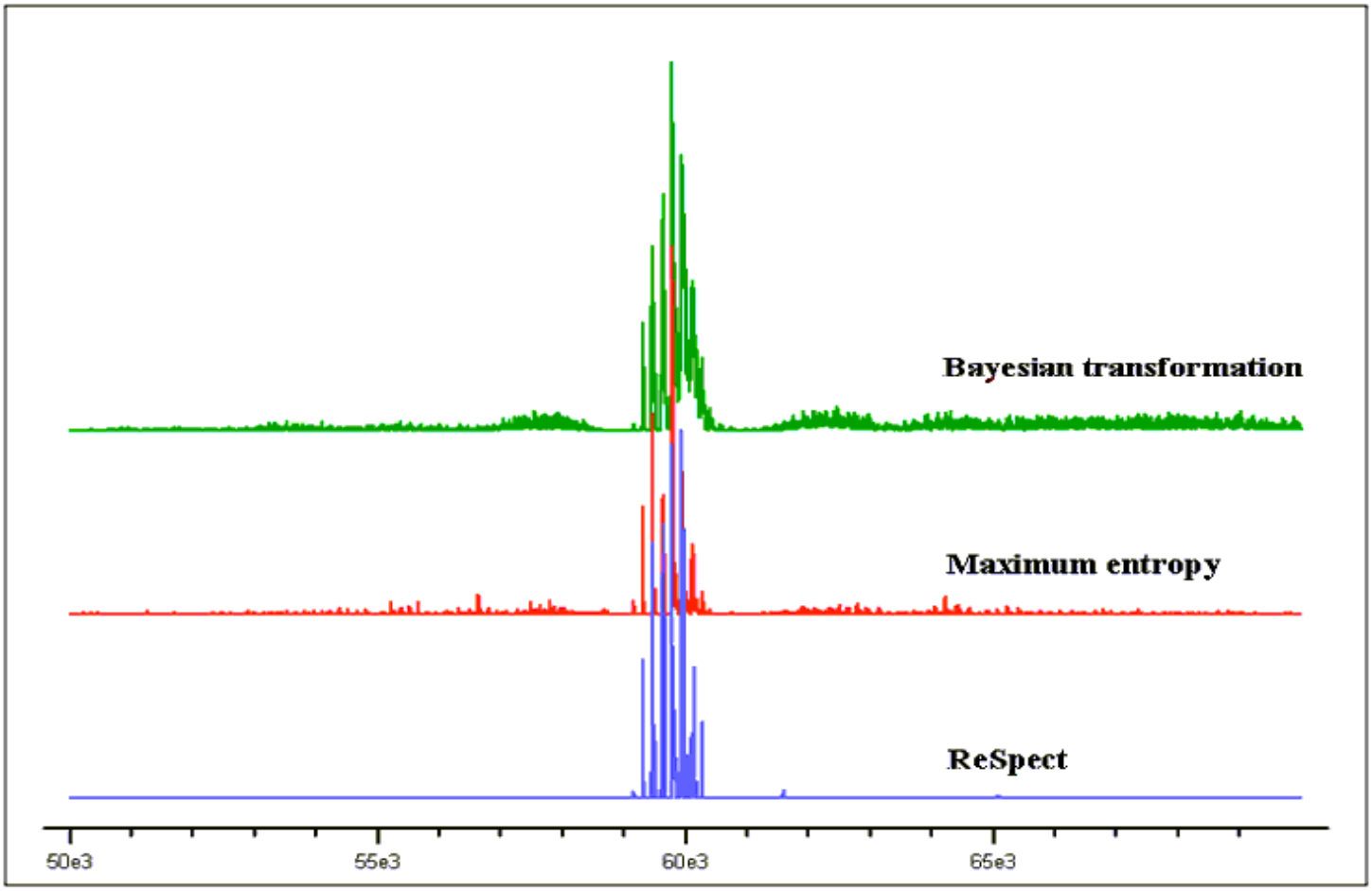

Figure 4: Zero-charge results for glycoprotein A. Data range m/z ~400-4000; Output mass range 50-70 kDa. **Top**: Bayesian transformation; **Centre**: Maximum entropy; **Bottom**: ***ReSpect***™.

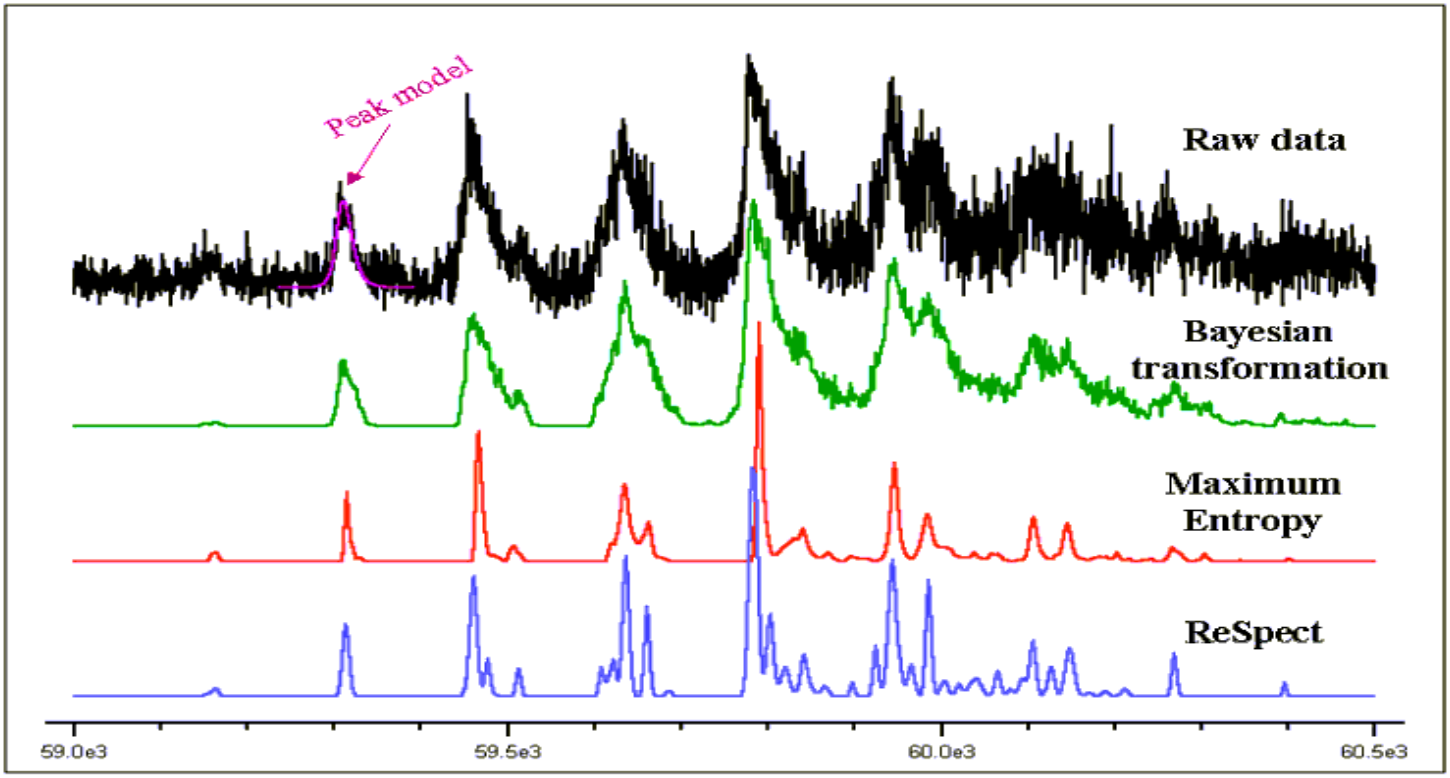

Figure 5: Glycoprotein A results. Input raw data m/z ~400-4000; Output mass range 50-70 kDa. The raw data cluster $(M+20H)^{20+}$ is scaled to overlay the output mass range of the zero-charge spectra.

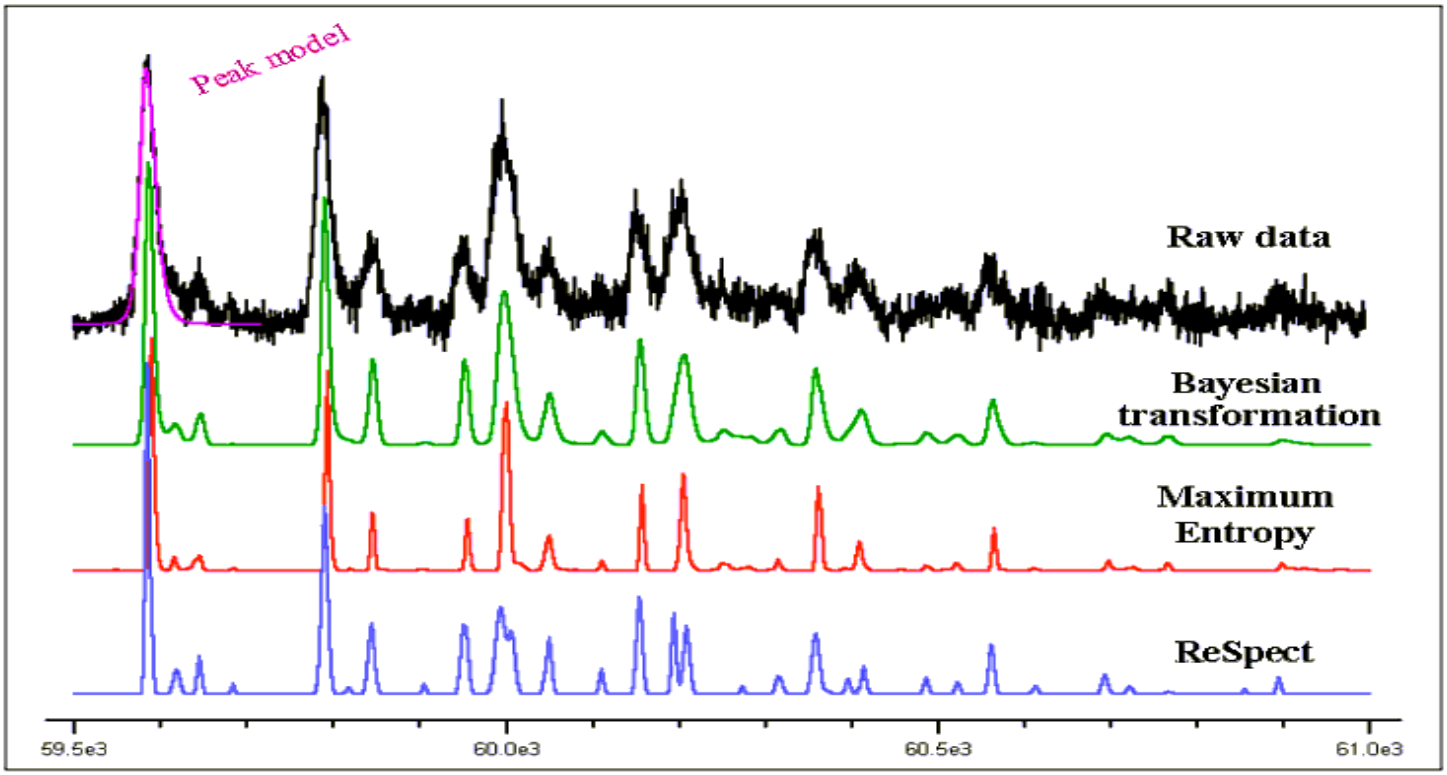

Figure 6: Glycoprotein B results. Input raw data m/z ~400-4000; Output mass range 50-70 kDa. The raw data cluster $(M+20H)^{20+}$ is scaled to overlay the output mass range of the zero-charge spectra.

Figures 7 & 8 compare the results using the reduced data range from m/z 2025-3450.

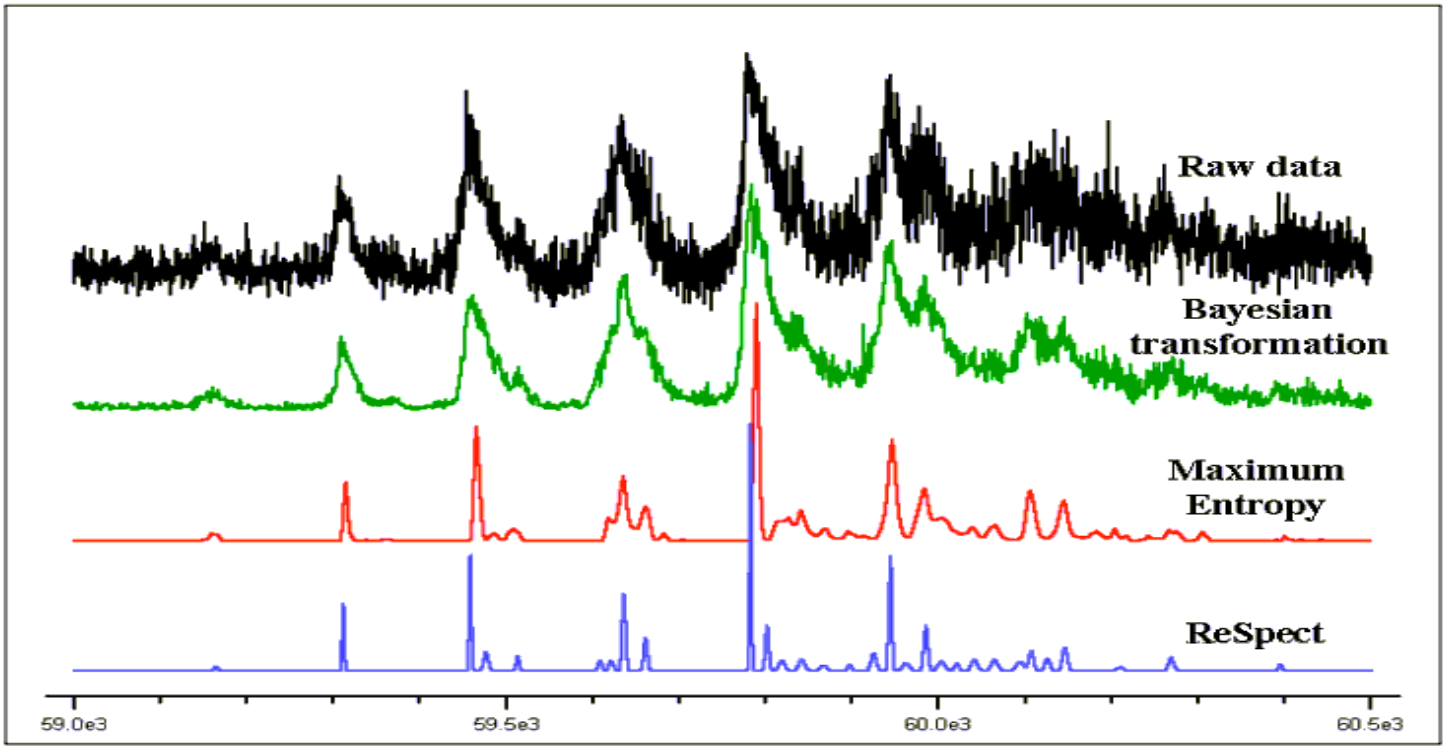

Figure 7: Glycoprotein A results. Input raw data m/z 2025-3450; Output mass range 58-62 kDa. The raw data cluster $(M+20H)^{20+}$ - scaled to overlay the output mass range of the zero-charge spectra.

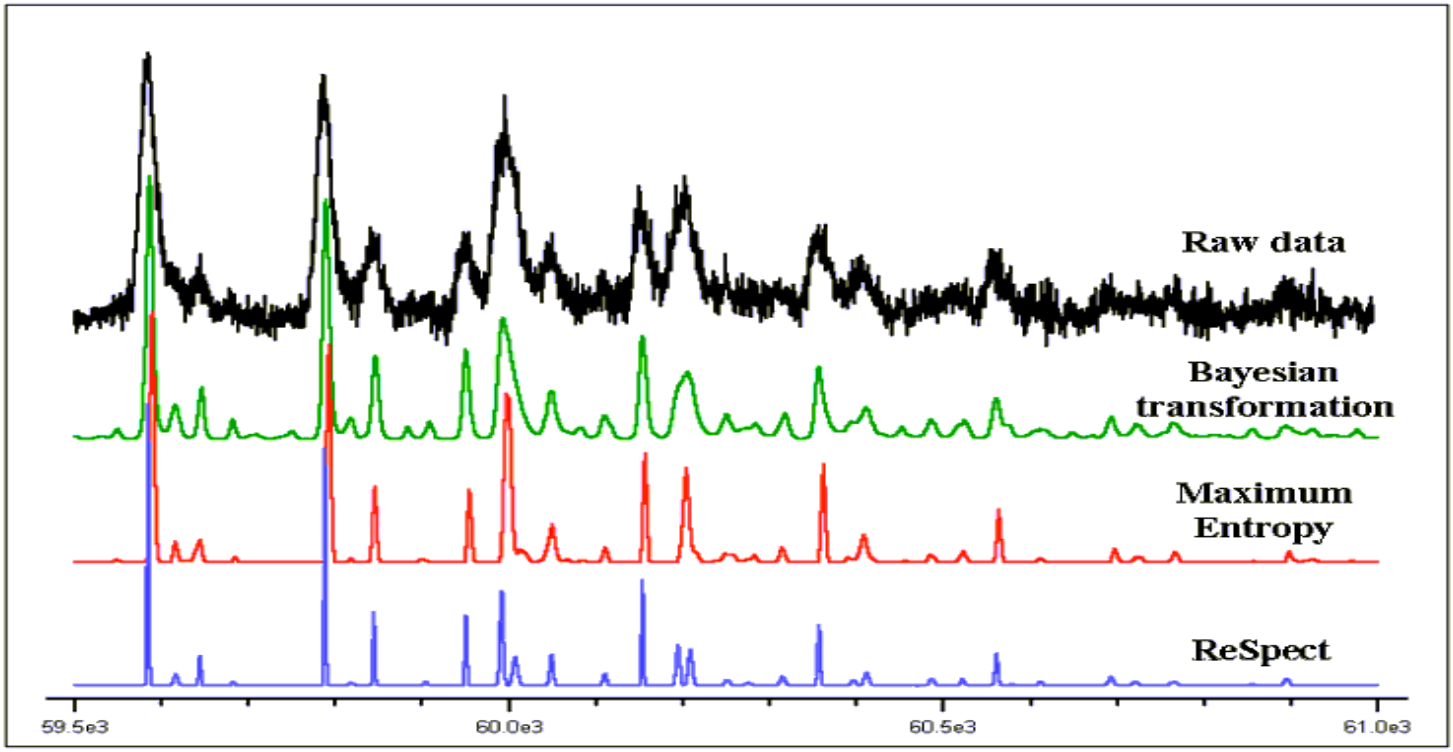

Figure 8: Glycoprotein B results. Input raw data m/z 2025-3450; Output mass range 58-62 kDa. The raw data cluster $(M+20H)^{20+}$ - scaled to overlay the output mass range of the zero-charge spectra.

Data Analysis

The well-resolved zero-charge spectra allow unambiguous identification of combinations of the different glycosylation on the four N-glycosylation sites of the proteins (Figs 9 & 10). It also illustrates the clear glycosylation difference for the two proteins. Table 1 shows the results interpretation for glycoprotein A.

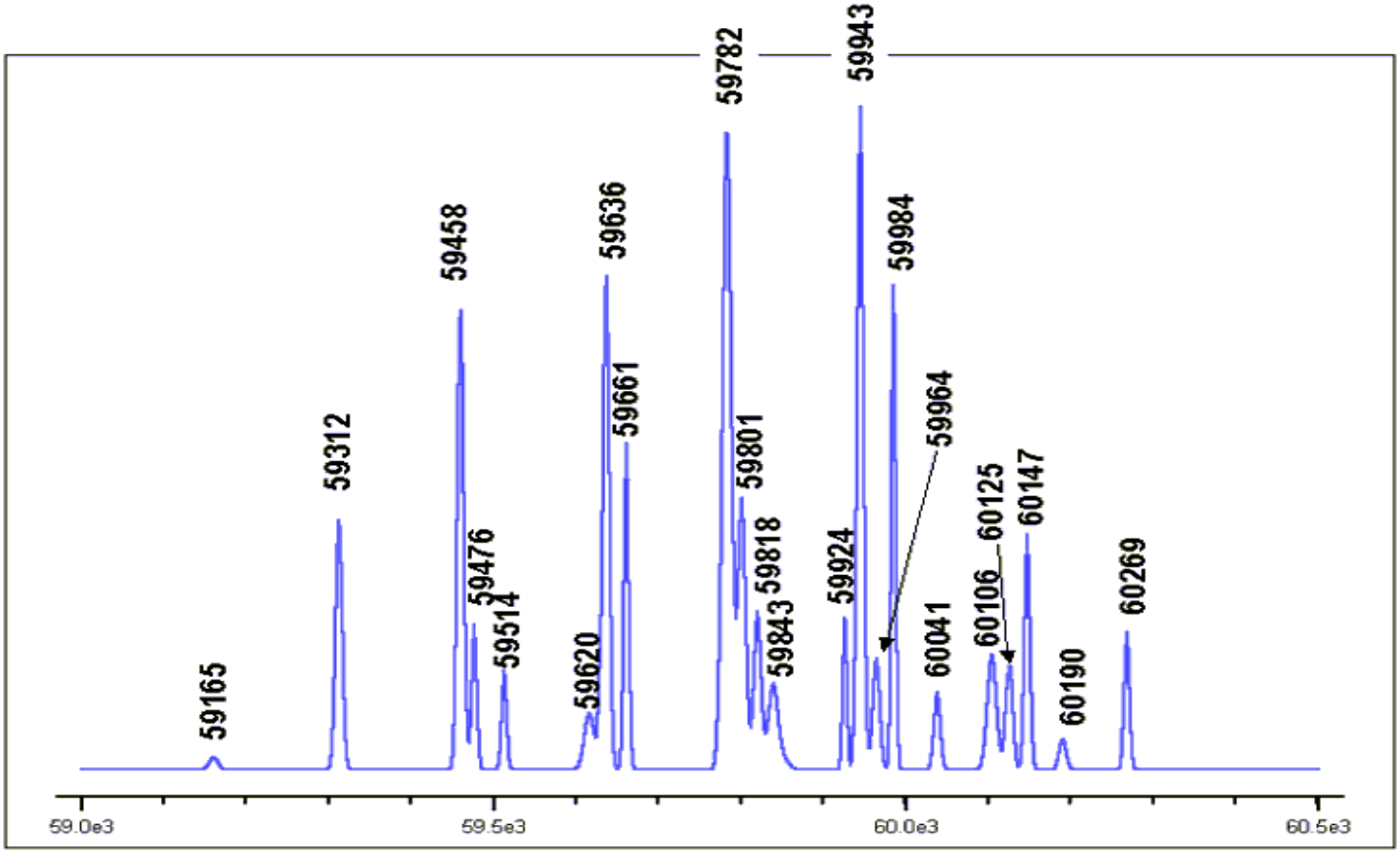

Figure 9: The zero-charge spectrum of glycoprotein A using all the data.

Table 1: Glycoprotein A Results

Found M*	Predicted M	Dif	Proposed CHO Structures
59167.9	59168.9	1.0	4Core**
59314.1	59315.0	0.9	4Core+1Fuc
59460.9	59461.2	0.3	4Core+2Fuc
59478.9	59477.2	-1.8	4Core+1Fuc+1Hex
59516.9	59518.2	1.3	4Core+1Fuc+1HexNAc
59623.3	59623.3	0.0	4Core+2Fuc+1Hex
59639.3	59639.3	0.0	4Core+1Fuc+2Hex
59663.6	59664.4	0.7	4Core+2Fuc+1HexNAc
59784.9	59785.4	0.5	4Core+2Fuc+2Hex
59804.4	59801.4	-2.9	4Core+1Fuc+3Hex
59845.7	59842.5	-3.2	4Core+1Fuc+2Hex+1HexNAc
59926.7	59931.6	4.9	4Core+3Fuc+2Hex
59946.5	59947.6	1.1	4Core+2Fuc+3Hex
59967.5	59963.6	-3.9	4Core+1Fuc+4Hex
59987.3	59988.6	1.3	4Core+2Fuc+2Hex+1HexNAc
60006.5	60004.6	-1.9	4Core+1Fuc+3Hex+1HexNAc
60044.1	60045.7	1.6	4Core+1Fuc+2Hex+2HexNAc

**Calibrated masses; **Core = $(HexNAc)_2Hex_3$*

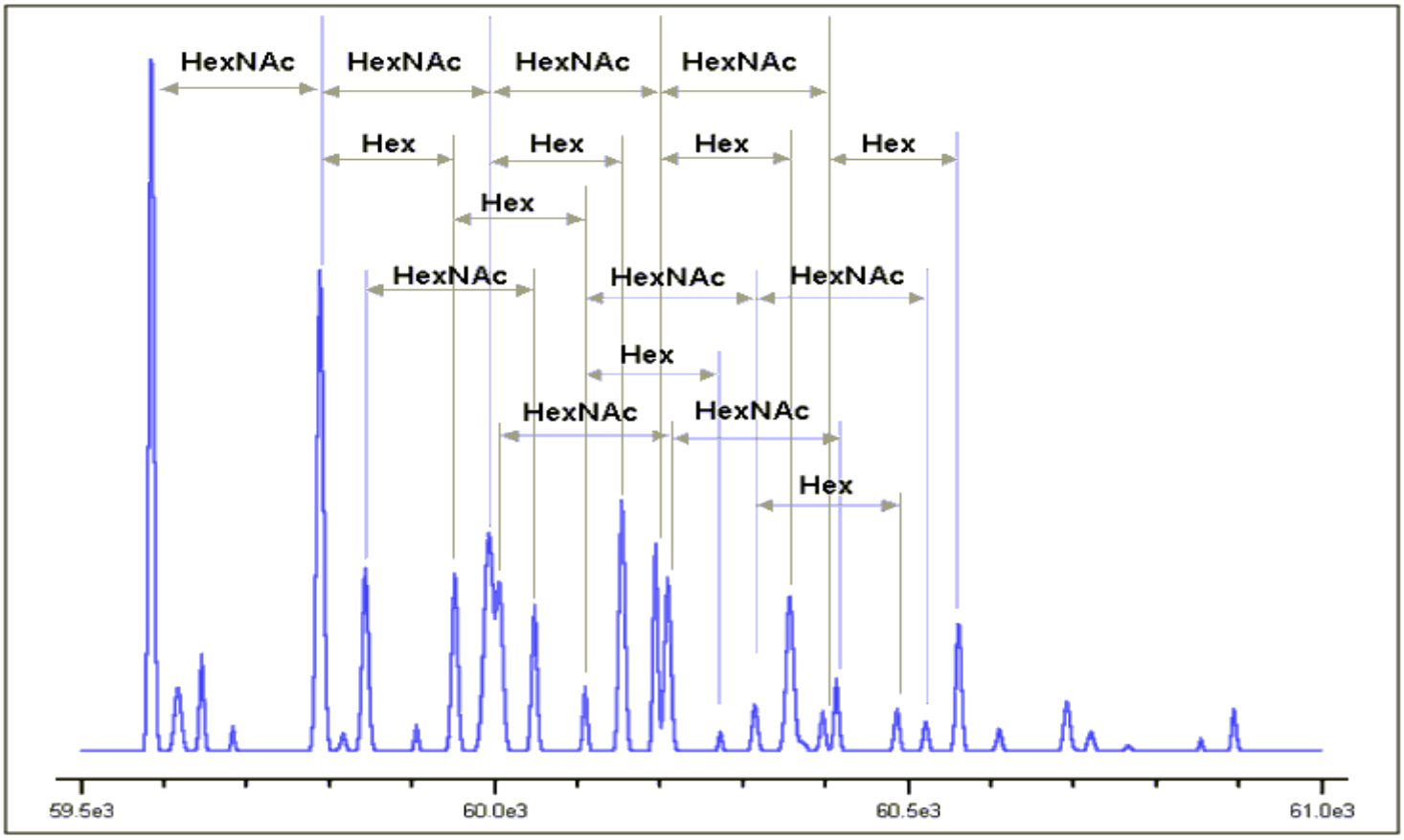

Figure 10: The mass differences in the zero-charge spectrum of glycoprotein B.
Note: Spectrum using all the data (see Fig. 6).

Discussion

Bayesian transformation generates numerous harmonics and artefacts (Fig. 4) and there is no improvement in resolution (Figs 5 & 6). It is also clear that the method works poorly on low S/N data since there is substantial residual noise in the result compared with the data (Fig. 5). Although using a reduced data range would normally be beneficial, the smaller output window forces all the harmonics and artefacts into the result causing further degradation to the result (Figs 5 & 7).

Maximum entropy provides substantial gains in resolution and fewer artefacts (Figs 5 & 6) but overlapped peaks are unresolved due to program model limitations and a small calibration error. The results are also compromised by using a global noise estimate and over- and under-deconvolution occurs. The smaller data and output ranges improves the resolution but overlapped peaks are still unresolved. A serious deficiency is that some peaks in the results are clearly displaced from their expected masses!

ReSpect™ produces almost no artefacts, if any (Fig. 4). There are substantial gains in resolution and overlapped peaks are well resolved (Figs 5 & 6). The improvement is due to several program features that include the ability to accommodate a calibration error and the incorporation of the "rules of electrospray" so that only masses consistent with the physics of the experiment are reported - mathematically plausible masses inconsistent with the "rules" are not computed. The smaller output range cannot have any effect but the smaller data range improves the quality and resolution of the results (Figs 5-8).

Although not reported here, the mass errors assigned by maximum entropy were confusing as they were smaller than the observed peak displacements. Therefore, only a partial analysis was possible. Of the three methods, ***ReSpect***™ produced the cleanest and most highly resolved results, allowing unambiguous data interpretation.

Conclusions

A comparison with existing methods has shown that the novel *ReSpect* algorithm provides a much more useful reconstructed zero-charge spectrum with very few (if any) artefacts and powerful resolution that allows unambiguous data analysis on complex ESI-MS data of heterogeneous glycoproteins.