Optimized Automatic Noise Level Calculations for Broadband FT-ICR Mass Spectra of Petroleum Give More Reliable and Faster Peak Picking Results

Manhoi Hur, Han Bin Oh,† and Sunghwan Kim‡,*

BNF Technology Inc., Daejeon 305-500, Korea
†Department of Chemistry, Sogang University, Seoul 121-742, Korea (200811036)
*‡Kyungpook National University, Department of Chemistry, Daegu 702-701, Korea. *E-mail: sunghwank@knu.ac.kr*
Received July 4, 2009, Accepted September 17, 2009

A new algorithm for determining noise level is proposed for more reliability in interpreting spectral data for complex Fourier transform ion cyclotron resonance (FTICR) mass spectra of petroleum. In the new algorithm, a moving window with a fixed number of data points was adopted, instead of a fixed *m/z* width. In the analysis of petroleum, it was found that a moving window of 50,000 or more data points was optimal. This optimized automated peak picking performed well even with frequency-dependant noise in the mass spectrum. Additionally, this fast, automated peak picking algorithm was suitable for the analysis of a large set of samples.

Key Words: FT-MS, 1/f noise, Baseline shift, Automated peak picking, Noise level calculation

Introduction

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) is widely accepted as one of the most powerful tools available for studying organic mixtures at the molecular level. The ultra-high mass resolution of FT-ICR MS often makes it possible to identify organic molecules based solely on the measured mass to charge (*m/z*) values. FT-ICR MS has been successfully applied to organic mixtures such as metabolites,[1] vegetable oils,[2] wine,[3] explosives,[4] coal extracts,[5] humic materials,[6,7] and crude oils.[8,9] Broadband FT-ICR MS spectra of these types of organic mixtures are usually very complex, with peaks appearing over a wide dynamic range.

When interpreting the mass spectra of organic mixtures, particularly crude oils, the first step is usually peak picking. Conventionally, a peak picking procedure for petroleum FT-ICR MS spectra is based mainly on the following two steps. First, several *m/z* regions within the spectra are expanded and visually examined. Second, a noise level, which is often represented by a signal-to-noise ratio, S/N, is manually determined, and the peak threshold is set based on this S/N level (the threshold is usually set at three to five times the S/N level). Due to the complexities of the spectra and a wide dynamic range of peaks appearing in petroleum spectra, calculation of an appropriate noise level is a critical step. If this calculation was performed incorrectly, a significant amount of noise might be included in the peak list, or alternately, valuable peak information might be discarded. Furthermore, the manual determination of a baseline can be done arbitrarily, which would cause inconsistent interpretation of spectra.

A number of computer programs have been developed to determine the mono-isotopic masses of peptides and proteins from high-resolution mass spectra.[10,11-16] Currently, these programs are devoted mostly to the interpretation of biomolecular spectra. Typically in these programs, an algorithm for automatic noise level determination is included.[10-12] In principle, this algorithm could also be used to interpret petroleum spectra. However, the algorithm must be modified and optimized to account for complex differences between protein and petroleum spectra. A single crude oil spectrum generally contains 3,000 to 10,000 peaks, and 15 to 30 peaks may be found in a window of less than 0.5 *m/z* in width,[8,9] while peaks in protein or peptide spectra are relatively sparse.

In the present study, a new algorithm for automated calculation of noise levels that is particularly optimal for the interpretation of complex petroleum spectra will be presented, and its utility in understanding petroleum samples will be demonstrated.

Experimental

To profile the changes in baseline noise over a frequency range of 50 kHz to 1.2 MHz, mass spectra were obtained with 7 and 15 T FT-ICR mass spectrometers (Bruker Daltonics, Billerica, MA). The mass spectrometers were equipped with AQS data stations (Bruker Daltonics, Billerica, MA, USA). The crude oil samples used in this study were first diluted to 2 mg/mL in toluene, which was then diluted to 1 mg/mL in a 50:50 (v/v) toluene/methanol solution just before analysis. The analyses were performed with electrospray ionization (ESI). A turbo spray ionization source (Bruker Daltonics, Billerica, MA, USA) was used with a 150 μm OD/30 μm ID capillary inserted through the spray tip. The sample was directly injected using a syringe pump (Harvard, Holliston, MA) at a flow rate of 40 - 70 μL/h. For each spectrum, 2×10^6 data points were obtained.

The baseline noise from each spectrum was calculated with a computer code modified from the "THRASH" algorithm.[10] The program modifications are described in the following Results and Discussion section. The program was written in the C/C++ programming language and was tested using a 2.33-GHz Intel Xeon processor with 3 GB RAM.

Results and Discussion

Optimizing the noise level calculation algorithm for petroleum data-windows with a fixed mass width versus those with a fixed number of data points. The noise level calculation algo-

rithm included in the THRASH[10] program has been widely used in interpreting peptide/protein mass spectra. However, the direct application of the algorithm to the interpretation of petroleum spectra can pose a problem because of the complexities in these spectra, as described earlier. Therefore, it is necessary to modify and optimize the noise level calculation algorithm for the analysis of petroleum mass spectra.

In the original THRASH algorithm,[10,11] a noise level is calculated based on the assumption that the baseline noise should contain the highest data point density.

Based on this assumption, the number of data points observed in the mass spectrum that are equal to or less than a specific intensity value is plotted against that same intensity value. The specific intensity value where the second derivative of the graph becomes zero is determined to be the baseline noise. Since the noise level determination is related closely to the number of data points, the total number of points in any defined window can be an important variable in the noise level determination. However, in the original THRASH algorithm, a mass spectrum was divided simply into equally spaced mass-to-charge windows, with windows in different *m/z* regions having different numbers of data points.[10] For example, a window of 1 *m/z* width at 300 *m/z* in a 15 T FT-ICR spectra obtained at a sampling rate of ~ 1.5 MHz has about 7,000 data points, whereas the same 1 *m/z* window at 900 *m/z* has less than 1,000 data points. This means that the noise level at 300 *m/z* is determined using a much larger number of data points than that at 900 *m/z*. Since the number of data points included in the defined window is an important variable for the signal-to-noise level determination algorithm, noise level calculations may not be consistent, particularly when a very complex mass spectrum is involved.

To eliminate the drawbacks described above, windows with equal numbers of data points should be used for noise level calculations. Note that a window with the same number of data points also means an equally spaced frequency window in FT-ICR MS spectra. In the present study, in order to make a comparison, peak finding was carried out using both approaches, *i.e.*, equally spaced *m/z* and frequency windows, for the same mass spectrum presented in Figure 1a, and two peak lists were obtained (refer to Supplementary data). The peak list obtained with a 1 *m/z* width equally spaced window included about 50 peaks identified in the region of 900 ~ 1,000 *m/z*, whereas the peak picking with a window of a constant number of data points (5,000 data points in this case) resulted in eight peaks over the same mass region. Manual inspection of the same mass region showed that there were only a few peaks existing at three times the noise level threshold (refer to Figure 1b). Most of the 50 peaks identified by the equally spaced *m/z* window were found to be noise.

To further examine the effect of a window with a constant number of data points in the peak picking algorithm, the same approach was also applied to the 300 ~ 400 *m/z* region. However, in this case, both methods identified an almost identical number of peaks, *e.g.*, 613 peaks by the equally spaced *m/z* window method, and 611 peaks using the fixed data point window. This discrepancy in the pick-finding results appears to arise from the fact that the data point density in a low *m/z* region

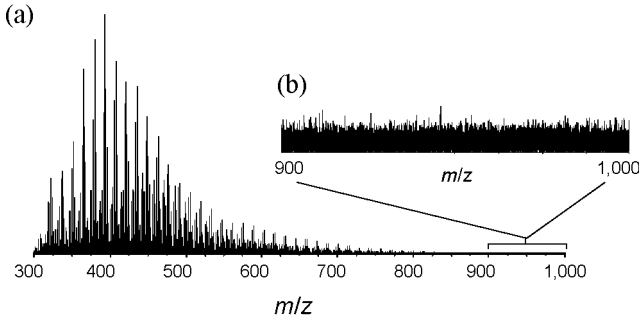

Figure 1. (a) A broadband spectrum of crude oil, and (b) the expanded region where no significant peaks exist.

is much higher than that in a high *m/z* region. Therefore, the influence of a window with a constant number of data points on the determination of a noise level is minimized in a low *m/z* region. In summary, in this study, it was found that noise level calculation and the associated peak-finding is more reliable in a fixed number of data points mode than in a fixed *m/z* window mode.

Optimal number of data points for the noise level determination. When the analysis is carried out in a fixed number of data points mode, *i.e.*, in a fixed frequency window mode, an entire spectrum is divided into many smaller windows. As stated in the previous section, the number of data points per window is an important parameter. Therefore, optimizing the number of data points for the window was attempted, in order to reliably interpret complex petroleum mass spectra.

To examine the effect of window size on noise level calculations, noise levels were determined with different numbers of data points, for the petroleum spectrum in Figure 1, and the results are given in Figure 2. In Figure 2, the noise levels were determined using (a) 25,000, (b) 50,000, (c) 100,000, and (d) 150,000 data points, respectively. With a 25,000 data point window, noise levels were observed to fluctuate 15% or more between adjacent windows (*e.g.*, the circled area in Fig. 2a). However, manual inspection of the original spectra did not exhibit such a large fluctuation. The noise level calculation with 50,000 or more data points resulted in much less noise level fluctuation (refer to Figures 2b and c). The observed fluctuation in Figure 2a likely came from the limited number of data points for the window. Windows containing 50,000 or more data points appear to be suitable for the analysis of petroleum data obtained in the FT-ICR mass spectrometer. On the other hand, calculating a noise level over a rather broad frequency range, *i.e.*, with too many data points, such as the 150,000 data points in Figure 2(d), may risk overlooking local frequency-dependent noise. A more detailed description of the local frequency-dependent noise will be given in the following section. Based on the observations described above, the window of 50,000 data points appeared to be optimal and was used for further noise level calculations that will be described in the following section.

Validation of the improved automated noise level determination method. Conventional data analysis programs other than THRASH calculate noise levels from the entire mass (or frequency) range of the spectrum and use this single noise

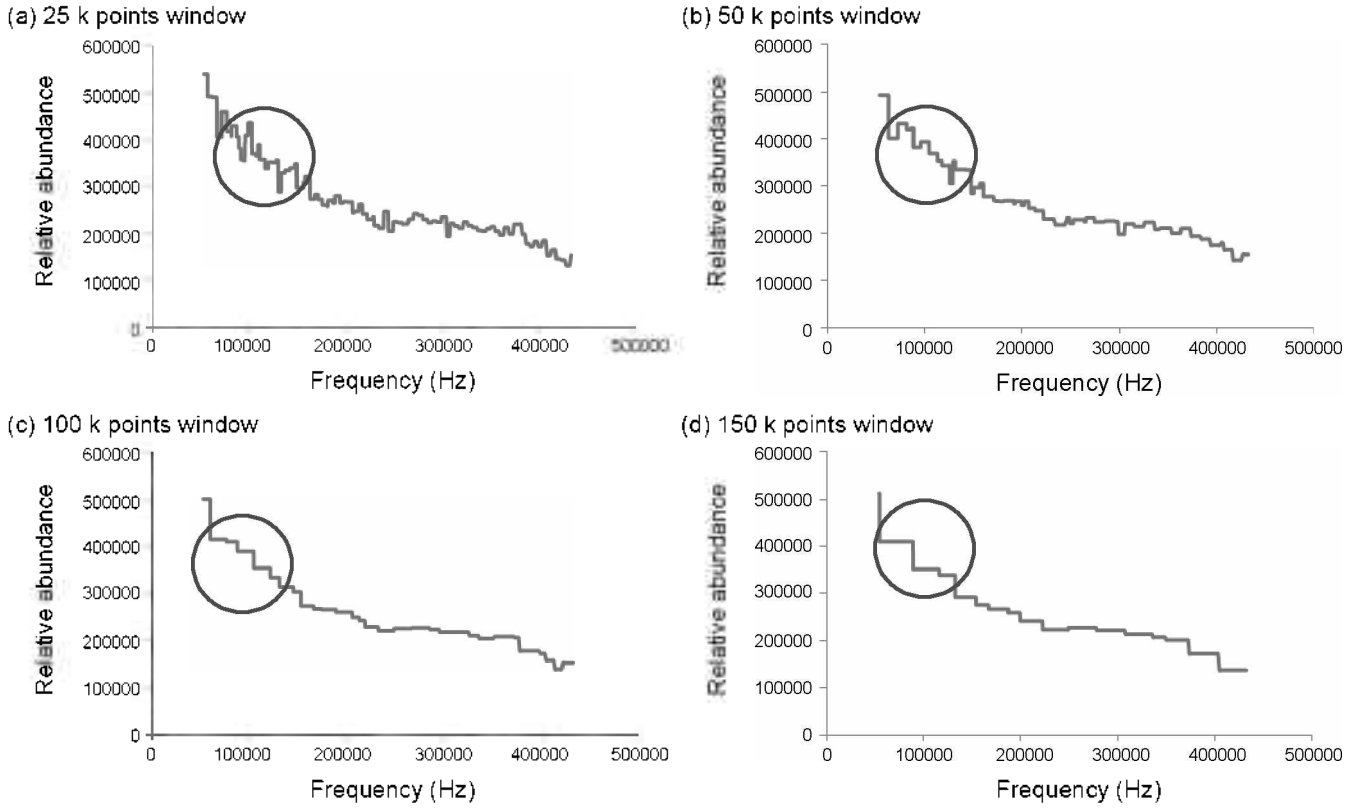

Figure 2. Profiled noise levels calculated with (a) 25 k, (b) 50 k, (c) 100 k, and (d) 150 k point windows.

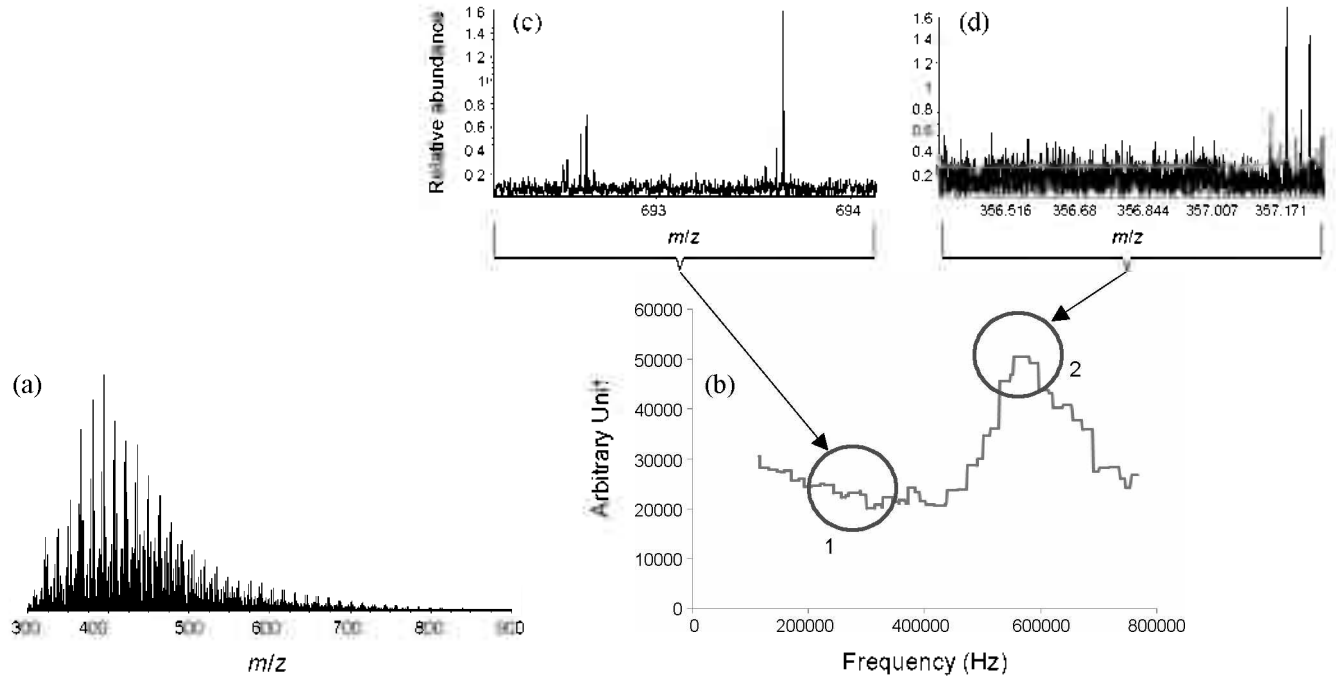

Figure 3. Displays of (a) a broadband spectrum of petroleum with frequency-dependent noise, (b) profiled noise level of the spectrum, and (c) & (d) expanded regions.

threshold for peak selections. However, this approach can lead to significant errors, since noise levels in mass spectra can vary significantly from one mass spectral region to another. This is exemplified in Figure 3, which shows a broadband mass spectrum of petroleum obtained with a 15 T FT-ICR MS, in which the profiled noise level of the spectrum is given in arbitrary units (see Figure 3(b)). In Figure 3(b), the noise level in the circled region 2 is almost two times larger than that in the circled area

Table 1. Numbers of peaks found by conventional and automated peak picking

	# of peaks found	# of unassigned peaks
Conventional peak picking, with a relative abundance threshold at 0.5% of the base peak	3086	597
Automated peak picking, with a S/N ratio cutoff of 4.5	2937	181

1. From these two specific cases of expanded spectra, taken from circled areas 1 and 2 (Figures 3(c) and (d)), it is clear that peak selections based on a fixed noise threshold over an entire mass could lead to missed valid peaks or mistakenly included noise peaks.

To make a direct comparison between the conventional and our improved data analysis algorithms, peak picking was performed for the petroleum mass spectrum of Figure 3(a) using both methods; note that the conventional method involves the determination of a single noise level over an entire mass spectrum. The total numbers of peaks found by each method were tabulated in Table 1. After peak picking, elemental formulae were calculated and assigned, based on the *m/z* values. Normal conditions for petroleum data ($C_cH_hN_nO_oS_s$, c unlimited, h unlimited, $0 \le n \le 5$, $0 \ o \le 10$, $0 \le s \le 2$)[17] were used for the verification. The number of peaks that could not be assigned with elemental formulae were also tabulated (Table 1). Despite the fact that similar numbers of peaks (~3,000 peaks) were found by both methods, the number of unassigned peaks was about three times larger when the conventional peak finding algorithm was used. This large difference in the number of unassigned peaks may be attributed to misassignment of noise as real peaks in the conventional approach. This demonstrated that the improved automated method performed better than the conventional one.

Furthermore, the automated peak picking was faster and more convenient than the manual inspection method, particularly when multiple spectra had to be processed. The manual peak picking includes expanding several spectral regions in a spectrum, examining each expanded region, and then determining the noise level. These processes have to be repeated for each individual spectrum, since each spectrum can have different noise level. However, with the automated peak picking, a user can avoid such tedious procedures, which ensures reliable analyses of complex FT-ICR mass spectra of petroleum in less time.

Conclusions

A new, automated peak picking algorithm with a moving frequency window was proposed and optimized for high-resolution FT-ICR mass spectra of petroleum. A constant number of data points (or frequencies) window with at least 50,000 data points was found to be optimal for petroleum data analysis. The new algorithm operates well even in the presence of frequency-dependent noise. Compared with the manual analysis, the automated program gives more consistent data analysis results. With the improved reliability in peak-peaking, more efficient data analysis is possible, even for a large number of data sets. We expect that this new algorithm for mass spectral analysis can also be applied to the semi-quantitative or quantitative interpretation of mass spectra of other organic species.

Acknowledgments. The authors thank Professor Fred McLaferty for kindly providing the "THRASH" program and also Mr. Michael Easterling at Bruker Daltonics and Korea Basic Science Institute, ochang campus, for providing data from 7, 9.4, and 15 T FT-ICR MS instruments for comparison. We also thank Dr. Gregory T. Blankney for helpful discussions about peak picking algorithms. This research was supported by a Kyungpook National University Research Fund, 2009.

Reference

1. Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spec. Rev.* **2007**, *26*, 51-78.
2. Wu, Z.; Rodgers, R. P.; Marshall, A. G. *J. Agric. Food Chem.* **2004**, *52*, 5322-5328.
3. Cooper, H. J.; Marshall, A. G. *J. Agric. Food Chem.* **2001**, *49*, 5710-5718.
4. Wu, Z.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2002**, *74*, 1879-1883.
5. Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Energy & Fuels* **2004**, *18*, 1424-1428.
6. Kim, S.; Kaplan, L. A.; Hatcher, P. G. *Limnol. Oceanogr.* **2006**, *51*, 1054-1063.
7. Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336-5344.
8. Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53-59.
9. Marshall, A. G.; Rodgers, R. P. *PNAS* **2008**, *105*, 18090-18095.
10. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320-332.
11. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *PNAS* **2000**, *97*, 10313-10317.
12. Kaur, P.; O'Connor, P. B. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 459-468.
13. Chen, L.; Yap, Y. L. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 46-54.
14. Johnson, K. L.; Mason, C. J.; Muddiman, D. C.; Eckel, J. E. *Anal. Chem.* **2004**, *76*, 5097-5103.
15. McIlwain, S.; Page, D.; Huttlin, E. L.; Sussman, M. R. *Bioinformatics* **2007**, *23*, I328-I336.
16. Park, K.; Yoon, J. Y.; Lee, S.; Paek, E.; Park, H.; Jung, H. J.; Lee, S. W. *Anal. Chem.* **2008**, *80*, 7294-7303.
17. Kim, S.; Rodgers, R. P.; Marshall, A. G. *Int. J. Mass Spectrom.* **2006**, *251*, 260-265.