Speed Improvement of an FTICR Mass Spectra Analysis Program by Simple Modifications

Sang Hyun Jeon, Hyeong Soo Chang, Manhoi Hur,†,a Kyung-Hoon Kwon,† Hyun Sik Kim,† Jong Shin Yoo,† Sunghwan Kim,‡ Soojin Park,§ and Han Bin Oh§,*

Department of Computer Science and Engineering (200811037), Sogang University, Seoul 121-742, Korea
†Korea Basic Science Institute, Daejeon 305-333, Korea
‡Department of Chemistry, Kyungpook National University, Daegu 702-701, Korea
*§Department of Chemistry and Interdisciplinary Program of Integrated Biotechnology, Sogang University, Seoul 121-742, Korea (200811036). *E-mail: hanbinoh@sogang.ac.kr*
Received July 14, 2009, Accepted August 4, 2009

Two simple algorithm modifications are made to the THRASH data retrieval program with the aim of improving analysis speed for complex Fourier transform ion cyclotron resonance (FTICR) mass spectra. Instead of calculating the least-squares fit for every charge state in the backup charge state determination algorithm, only some charge states are pre-selected based on the plausibility values obtained from the FT/Patterson analysis. Second, a modification is made to skip figure-of-merit (FOM) calculations in the central *m/z* region between two neighboring peaks in isotopic cluster distributions, in which signal intensities are negligible. These combined modifications result in a significant improvement in the analysis speed, which reduces analysis time as much as 50% for ubiquitin (8.6 kDa, 76 amino acids) FTICR MS and MS/MS spectra at the reliability (RL) value = 0.90 and five pre-selected charge states with minimal decreases in data analysis quality (Table 3).

Key Words: THRASH, FTICR mass spectrum, Code modifications, Spectral analysis

Introduction

An increasing use of Fourier-transform mass spectrometry (FTMS) has been witnessed in both proteomics research and in the analyses of complex mixtures, such as lipids and petroleum.[1-5] The wide use of FTMS is largely due to its extraordinary resolving power, its accompanying mass accuracy, and its versatility, which allows a variety of different tandem mass spectrometry methods to be combined. As a result, FTMS often produces very complex mass spectra, which makes the manual interpretation of mass spectra data very time-consuming.

Several data-reduction algorithms have been developed with the aim of facilitating the analysis of complex FTMS spectra.[6-9] These include the deconvolution of FTMS spectra,[6] the Z score algorithm,[7] THRASH (Thorough High Resolution Analysis of Spectra by Horn),[8] and MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction).[10] The deconvolution method is based on the principle that charge states have only integer values.[6] Thus, ion peaks of the same mass but of different charge states are combined to determine the unique mass value of the ions. However, this method often suffers from poor representation of peaks that are shown at only one charge state, and is even worse for peaks with a low signal-to-noise ratio. In the Z score algorithm, the charge-state deconvolution was improved by employing a charge scoring scheme that incorporated all above-threshold members of a family of charge states or isotopic components.[7] Later, Horn *et al.* developed an automated program, THRASH, which can operate with minimal human intervention.[8] This program combined a variety of functional modules, including subtractive peak finding, primary charge determination by Fourier transform/Patterson method,[9] least-squares fitting to a theoretically derived isotopic abundance distribution for *m/z* determination of the most abundant isotopic peak,[10,11] and the statistical reliability determination.[8] By combining the above-mentioned modules, a complicated mass spectrum could be reduced into a single peak mass list. Recently, an improved data reduction program, MasSPIKE, was introduced by Kaur and O'Connor.[12] In this program, a matched filter algorithm was adopted for charge-state determination and monoisotopic masses were determined by aligning the theoretical and experimental isotopic distributions.

Analysis speed is an important aspect to be considered in developing an automated mass spectra interpretation program, along with reliability of data interpretation. In the present study, we describe an improvement of the THRASH algorithm that reduces analysis time through simple code modifications. Some of the results described here were originally presented at the Korean Information Science Society Conference in 2005.[13] In the present paper, a more elaborate description of the code modifications and improvements in analysis speed will be given. We would also like to emphasize that the speed enhancement is achieved with a minimal decrease in the accuracy and reliability of spectral interpretation.

Experimental

The original THRASH codes were written in PV-Wave 6.10 language, which was kindly provided by the McLafferty group at Cornell University. We first translated THRASH code into C++ (with the aid of the C++ version provided by Prof. Siu Kwan Sze at Nanyang Technological University, Singapore), allowing the THRASH algorithm to operate in the Microsoft

[a]Current address: BNF technology Inc., Daejeon, Korea

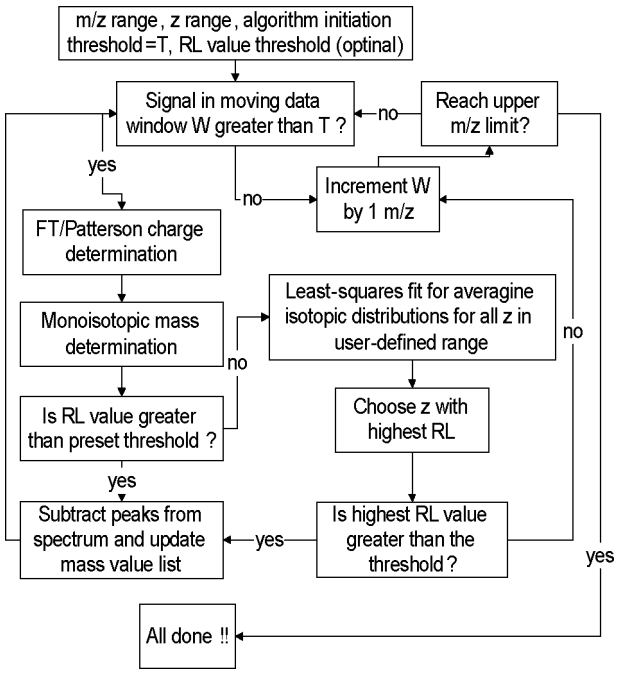

Figure 1. Flowchart of the THRASH program. Reprinted with permission from: *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320-332. Copyright 2000 American Society for Mass Spectrometry.

Windows® environment. The validity of the C++ translation was confirmed by cross-checking peak lists generated with the original and translated THRASH algorithms. Code modifications were made to the translated version of THRASH, and the details will be described in the following sections. The modified THRASH algorithm was used to analyze 292 FTMS spectra that were previously obtained on a 6 Tesla FTICR mass spectrometer at Cornell University. The analyzed spectra include 6 SORI-CAD (sustained off-resonance irradiation-collisionally activated dissociation),[14] 39 IRMPD (infrared multiphoton dissociation),[15] 207 ECD (electron capture dissociation),[16-18] and 40 simple ESI spectra. For the batch-mode analyses, only the mass spectra of bovine ubiquitin (8.6 kDa) were analyzed, and identical user-defined threshold parameters were employed. The given user-defined parameters are as follows: *m/z* range to be examined: 400-1,800; maximum mass of the molecules of interest: 10,000 Da; charge range: 1-20; positive ion mode; threshold reliability (RL) value (the probability of a successful fit; its definition will be given in detail in the following section): 0.90; signal-to-noise ratio: 3. The analyzed mass spectra were pre-processed with one zero-fill and no apodization. All analyses were performed on a personal computer with an Intel Pentium 43.2 G CPU processor, and 512 MB of RAM.

Results and Discussion

Brief description of the original THRASH program. Figure 1 shows the summarized flowchart of the original THRASH program.[8] Detailed descriptions of the program and component algorithms can be found elsewhere.[8] Only a brief overview of the algorithm components will be described here.

For a very complex FT mass spectrum, peak identifications are made for isotopic peak clusters that satisfy user-defined input parameters, such as the signal-to-noise (S/N) threshold, the range of expected charge states, the desired *m/z* range, and the minimum reliability value. The minimum reliability value is used to determine whether or not peak identification with a least-squares fitting procedure was successful (see below). For isotopic peak clusters with an abundance higher than a given S/N threshold, a charge state is assigned using a combination of Fourier transform (FT) and the Patterson charge determination algorithm.[9] The determined charge state value is then used for calculating an approximate mass value of the most abundant isotopic peak. Next, for this mass value, the corresponding virtual model peptide, the so-called *averagine*, is matched.[8,10,11] For each averagine, a theoretical isotopic abundance distribution (TID) is derived, and this distribution is used for determining the most abundant *m/z* value and the monoisotopic mass. The calculated isotopic abundance distribution is then compared to the experimentally obtained isotopic peak cluster to find the best least-squares fit. This calculation is repeated at a 1.00234 Da spacing. The least-squares fit result is returned quantitatively in the form of a figure-of-merit (FOM) value:

$$\text{Figure-of-merit (FOM)} = \frac{Number-of-comparisons}{\sum[(A_n - NI_n)^2 + (NI')^2]} \quad (1)$$

where A_n is the relative abundance of the nth peak in the theoretical abundance distribution; I_n is the spectral intensity at the *m/z* value of the nth peak; I' is the maximum spectral intensity in the adjacent valley; and N is a normalization factor.[8] The probability of a successful fit can be obtained by converting the best fit figure-of-merit value into a RL value, in which a conversion factor is calculated by comparing the results from the fully automated program and from the manually interpreted spectrum. When the calculated RL value is greater than a specified threshold value, the assigned charge state and the most abundant *m/z* value are assumed to be correct. A 90% threshold RL value was used as a default value in the original THRASH program, and therefore this value was adopted as the default value in our revised program. When the calculated RL value is less than the designated threshold value, *e.g.*, 0.90, it indicates that there is a good chance of incorrect assignment by the automated charge state determination program. In this case, a least-squares fit is recalculated for each charge state in the user-defined range, using a backup charge state determination routine. If the recalculated maximum RL value is greater than the specified threshold value, *e.g.*, 0.90, the charge state and the most abundant *m/z* are determined to be correct. When the threshold requirement is not met, the program recognizes that there is no isotopic cluster within this specific moving data window.

Run-time analysis of the translated THRASH algorithm. To improve the operation speed of the THRASH program, it is necessary to perform a run-time analysis for each component in the automated program. In the present study, run-time analyses were performed for 292 FTMS spectra of ubiquitin, and the results are shown in Table 1. The analyzed mass spectra included four spectra with over 100 identified isotopic clusters (ICs), 28 spectra with $80 \leq ICs < 100$, 38 spectra with $60 \leq ICs < 80$, 61

Table 1. Run-time analysis for THRASH algorithm components at the threshold RL value 0.90. A total of 292 FTICR spectra were analyzed.

Algorithm component[a]	Average run-time, s	% run-time
FT/Patterson charge state determination	0.0667	1.56
Most abundant and monoisotopic mass determination[b]	0.575	13.43
Peak subtraction	0.00112	0.026
Backup charge determination routine	3.187	74.49
Etc	0.449	10.50
Total	4.279	100

[a]The following parameters are used; *m/z* range, 400-1,800; maximum charge state, 20; threshold signal-to-noise ratio (peak level), 3.0; maximum mass, 10,000 Da. [b]The run-time for '*initial RL value calculation*' is included in the run-time for 'most abundant and monoisotopic mass determination' step.

Table 2. The number of ion cluster peaks identified by the translated C++ THRASH and its counterpart program with no backup charge state determination routine, at five different RL threshold values. Analyses were made for 292 FTMS spectra.

RL Threshold value	C++ THRASH	No backup routine	Difference
0.95	11,555	10,386	1,169 (10%)
0.90	12,776	11,261	1,515 (12%)
0.85	13,579	11,826	1,753 (13%)
0.80	14,235	12,258	1,977 (14%)
0.75	14,917	12,642	2,275 (15%)

spectra with $40 \leq ICs < 60$, and 161 spectra with less than 40 ICs.

As shown in Table 1, 74% of the total run-time, *i.e.*, 3.187 s per spectrum, is used in the backup charge state determination algorithm. As described above, the backup charge state determination routine becomes activated only when the initial charge determination turns out to be unreliable, or in other words, when the RL value was less than 0.90. Once this algorithm is called upon, a least-squares fit is recalculated for the each charge state in the user-defined range. In the present study, charge states were in the range of 1 ~ 20.

For other components, run-time allotments were negligible. For example, the algorithm used for finding the initial best least-squares fit and the most abundant/monoisotopic masses occupied only ~ 13% of the total run-time. For the FT/Patterson charge determination, only 1.6% of the total run-time was used. From this run-time analysis, it is clear that the backup charge state determination routine should be improved for analysis speed enhancement.

It was also found that the absolute analysis time improved dramatically from the original PV-Wave version, taking only 4.28 s per spectrum. The improved run-time was largely due to the translation of the original PV-Wave code into C++ language. It is also noteworthy that despite its high run-time usage, the backup charge state determination algorithm did not significantly increase the number of identified peaks compared with the mass spectral analyses performed without this backup charge state determination routine. Use of the backup charge state routine led to only a 12% increase in the number of identified ion peaks, at the threshold RL value of 0.90 (see Table 2).

Description of Modifications. From the above results, it is clear that the most efficient way to speed up analysis is to improve the backup charge state determination algorithm and its related routines. In the present study, code modifications are made exclusively to the following two routines: the backup charge state determination routine, and the figure-of-merits calculation routine. The details of the modifications and the accompanying results are described below.

Modifications in the Backup Charge State Determination Module: In the original backup charge state determination module, a least-squares fit is performed for every charge state that a user defines at the start of the THRASH program. If fewer charge states are used in the backup charge state determination routine, analysis time can be reduced significantly. For this purpose, the user-defined charge states were ranked based on the plausibility value obtained by the FT/Patterson algorithm,[9] and the backup charge state determination routine was run for only charge states that ranked high in the FT/Patterson algorithm. Having a smaller number of candidate backup charge states would be expected to be advantageous in reducing the analysis time, however, it could also adversely affect the peak identification quality. Therefore, a compromise is made between interpretation quality and analysis time. In addition, +1 and +2 charge states are always included because +1 and +2 charge states are very likely to be underestimated or overlooked in the FT/Patterson calculation due to the fact that a moving window of 1 *m/z* width is employed.

Reduction of the Fitting Range in Figure-of-Merits Calculations: As briefly described above, figure-of-merit (FOM) values are used for charge state determination in both the initial and backup charge state determination routines. The calculation of FOM values demands substantial run-time. We made some changes to this routine in order to reduce analysis time.

In the original THRASH algorithm, FOM values are repeatedly calculated by shifting theoretical isotopic cluster distributions (TID) in 0.001 *m/z* steps across a 1 *m/z* width window and aligning the TID against the experimental isotopic cluster distribution (EID).[8] This process requires substantial run-time. However, it is not necessary to calculate FOM values for all *m/z* regions between two neighboring EID peaks, since the abundances of TID and EID are negligible around the center of two neighboring EID peaks. Because of this, in the present study, FOM calculations are skipped for some regions between the neighboring peaks in the EID. In particular, a central *m/z* region between two neighboring EID peaks is skipped in FOM calculations.

To improve speed, the determination of an *m/z* width of the central skipped region is important. If the width is half of the spacing between two neighboring EID peaks, then the total run-time will be reduced by about half. However, since the width is also closely related to interpretation reliability, a wider width is not always recommended. For these reasons, a width of 3/5 spacing between two neighboring peaks is skipped in the present study in order to balance reduction of the total run-time and spectral interpretation quality.

Table 3. Average analysis time per spectrum, and the total number of peaks (in 292 spectra) obtained with four different RL threshold values.

RL value	Original THRASH		Modified backup routine[a]		Reduction in FOM calculation range[b]		Combination[a]	
	run-time	peaks	run-time	Peaks	run-time	Peaks	run-time	peaks
0.95	4.91 s	11,555	3.59 s	11,553	2.82 s	11,468	2.35 s	11,466
0.90	4.81 s	12,776	3.69 s	12,767	2.89 s	12,655	2.40 s	12,651
0.85	4.78 s	13,579	3.72 s	13,424	2.91 s	13,430	2.42 s	13,284
0.80	4.80 s	14,235	3.74 s	13,979	2.92 s	14,043	2.42 s	13,799

[a]5 pre-selected charge states were used, including −1 and +2. [b]FOM calculations were skipped for 3/5 of the region between the two neighboring peaks.

Table 4. The number of missed ion clusters and average run-times obtained using the original THRASH program, and the modified version with a different number of pre-selected charge states in FT/Patterson analysis.

Original THRASH		Missed ion clusters[b]	Average run-time[b], s
		0 (0%)[c]	4.814
The number of pre-selected charge states[a]	**3**	27 (0.21%)	2.881
	4	17 (0.13%)	3.200
	5	9 (0.07%)	3.703
	6	5 (0.04%)	4.031
	8	1 (0.008%)	4.226
	10	0 (0%)	4.483

[a]−1 and +2 charge states are always included. [b]At the threshold RL value of 0.90. [c]In total, 12,776 ion cluster peaks are identified.

Speed Enhancement. The modified THRASH program showed a significant improvement in analysis speed. Table 3 shows average analysis times per spectrum obtained for four different RL threshold values. The analysis was made for a total of 292 spectra. As shown in Table 3, the modifications in the backup charge state determination routine, the reduction in FOM calculation range, and the combined code modifications showed an improvement of analysis speed by 28 ~ 37, 64 ~ 74, and 97 ~ 109%, respectively, where % analysis speed improvement is defined by the improved analysis speed (1/run-time) divided by the original analysis speed, *e.g.*, for the combined code modifications at RL = 0.95, [(1/2.35)/(1/4.91) − 1] × 100 = 109%. With a RL value of 0.90, which is the default value in the program, 2.40 s was the average run-time. When other user-defined parameters were given, the improvement becomes more obvious. For example, at the maximum mass of 20,000 Da, the run-time decreased from 14.30 s to 3.75 s at RL = 0.90, corresponding to approximately four-fold time improvement (data not shown). It is also quite noticeable that the total number of identified peaks decreased by only less than 1% at the RL value of 0.90, even after code modifications. This clearly indicates that our code modifications improved analysis time with minimal sacrifice in peak identification capability.

We also analyzed the relationship between the number of charge states pre-selected through FT/Patterson analysis and the number of missed ion clusters, and the results are summarized in Table 4. The number of missed ion clusters and the average run-times were compared for the original THRASH program and the speed enhanced version, in which a modification was made only for the backup charge state determination step without the reduction in an FOM calculation region. The threshold RL value was set at 0.90. As shown in Table 4, even when three charge-states were pre-selected, the number of unidentified (or missed) ion clusters was negligible, i.e., only 0.21%. This suggests that most of the ion cluster peaks that had an initial RL value less than 0.90 are singly- or double-charged ions, since the pre-selected charge states included only +1, +2, and the most probable charge state from FT/Patterson analysis. With those three pre-selected charge states, the average run-time was reduced significantly, from 4.81 to 2.88 s. As expected, when a larger number of charge states was considered, fewer ion clusters were found to be missing, but with some sacrifice in the analysis speed. For 8.6 kDa ubiquitin, five pre-selected charge states appears to be the most reasonable choice, considering the missed ion clusters and the analysis speed improvement (see Table 3). For larger proteins, the speed improvement is expected to be more significant since in the original program, least-squares fit calculations in the backup chare state determination routine would be performed for a much wider range of possible charge states of the larger protein ions, while in the improved program, a fixed number of charge states would be considered in the backup charge state determination step. In the modified THRASH program, the number of candidate charge-states is given as an option, in order to provide users with more flexibility.

Conclusions

A significant improvement in the analysis speed of the THRASH program is achieved through simple modifications in the backup charge state determination routine, and reduction in the FOM calculation range. When operated on a computer system with fast, multiple CPUs, real-time analysis of simple peptide MS/MS spectra should be possible.

Acknowledgments. The authors are very grateful to the generosity of Prof. Fred W. McLafferty (Cornell University, NY, USA) and Prof. Siu Kwan Sze (Nanyang Technological University, Signapore) for providing us with the original PV-Wave and the C++ THRASH codes, respectively. This work was financially supported from Korea Basic Science Institute. HBO is thankful to the support from National Research Foundation of Korea (NRF-2009-0075245).

References

1. Jones, J. J.; Stump, M. J.; Fleming, R. C.; Lay, J. O., Jr.; Wilkins,

C. L. *J. Am. Soc. Mass Spectrom.* **2004**, *17*, 1665.
2. Yu, S. H.; Cho, K.; Kim, Y. H.; Park, S. J.; Kim, J. D.; Oh, H. B. *Bull. Korean Chem. Soc.* **2006**, *27*, 793.
3. Han, X.; Jin, M.; Breuker, K.; McLafferty, F. W. *Science* **2006**, *314*, 109.
4. Mann, M.; Kelleher, N. L. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18132.
5. Kim, S. H.; Rodgers, R. P.; Blakney, G. T.; Hendrikson, C.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 263.
6. Mann, M.; Meng, C. K.; Fenn, J. B. *Anal. Chem.* **1989**, *61*, 1702.
7. Zhang, Z. Q.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 225.
8. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320.
9. Senko, M. W.; Beu, S. C.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 52.
10. Senko, M. W.; Beu, S. C.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229.
11. Rockwood, A. L.; Van Orden, S. L.; Smith, R. D. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 54.
12. Kaur, P.; O'Connor, P. B. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 459.
13. Jeon, S. H.; Chang, H. S.; Oh, H. B. *Proceedings of the Korean Information Science Society Conference(B)* **2005**, *32*, 241.
14. Gauthier, J. W.; Trautman, T. R. *Anal. Chim. Acta* **1991**, *246*, 211.
15. Little, D. P.; Speir, J. P.; Senko, M. W.; O'Connor, P. B.; McLafferty, F. W. *Anal. Chem.* **1994**, *66*, 2809.
16. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265.
17. Oh, H. B.; Breuker, K.; Sze, S. K.; Ying, G.; Carpenter, B. K.; McLafferty, F. W. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 15863.
18. Lee, S. Y.; Park, S. J.; Lee, Y. W.; Oh, H. B.; Kang, H.; Cho, K. H.; Ahn, W. K.; Rhee, B. K. *Bull. Korean Chem. Soc.* **2008**, *29*, 1673.