IntelliXtract 2.0: Simplified Intelligent Component Extraction and Detection

Anne Marie Smith;[1] Richard Lee;[1] Vitaly Lashin;[2] Andrey Paramonov[2]

[1]Advanced Chemistry Development, Inc. (ACD/Labs), 8 King Street East, Toronto, ON. M5C 1B5. Canada; [2]ACD/Labs, Moscow, Russia

Overview

- Simplified detection and Component Extraction algorithm from LC-MS and GC-MS datasets
- New improved algorithm based on ion threads
- Reduced number of parameters to select for analysis
- Reduced false positives leading to reduced analysis time

Introduction

The analysis of real-world samples is becoming increasingly complex and time-consuming. Scientists frequently use techniques, such as chromatography to aid in the separation of differing compounds and components. Liquid Chromatography-Mass Spectrometry (LC-MS) has been the primary platform for determination of differing components when chromatographic co-elution was inevitable.

Software can aid in simplifying the data-mining process and increase the speed of discovery. An earlier algorithm for component detection, named IntelliXtract (IX), implemented a method of extracted ion chromatograms (XICs) that were automatically generated and assigned a component number, thereby simplifying the analysis process, but was computationally demanding. Here we describe a simplified and more optimized algorithm based on the use of ion threads vs. XICs.

Method

A second-generation feature-finding algorithm (IX 2.0) for identification of unknown components in LC-MS and now GC-MS data was developed to improve upon the previous version (IX 1.0). IX 1.0 involved a significant number of parameters which were highly sensitive to change and were significantly interdependent, hampering the data analysis process (Figure 1).

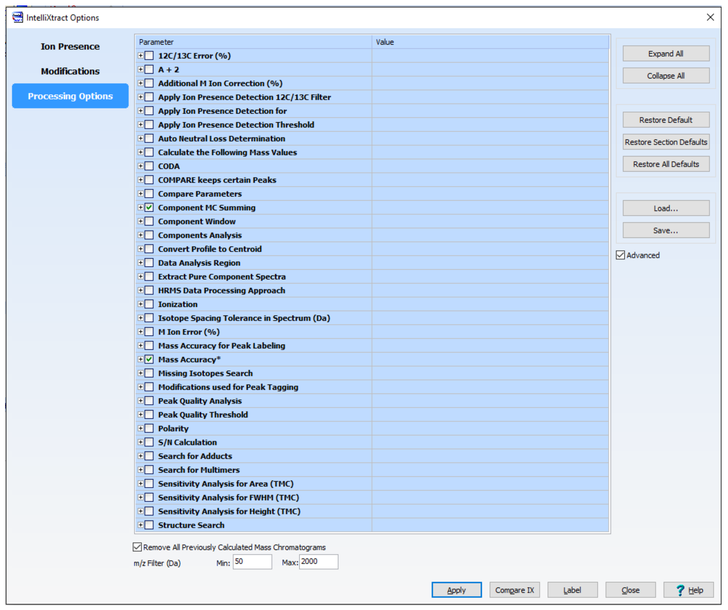

Figure 1: IntelliXtract 1.0 complete parameter options. Typically customers are overwhelmed by the number of parameters and instructed to focus on a few major ones (A+2, CODA, Mass Accuracy).

The identification of components within a LC-MS or GC-MS dataset was expedited through the use of a new component detection algorithm internally named “ion threads”.

Here, we show the effects of reducing the user input options for SNR for ^{12}C Peaks, Component Abundance Threshold (% of max), and the option to limit the number of components found. Data from various applications were used in the comparison of IX 1.0 to 2.0, where critical factors were assessed such as processing time and number of components found.

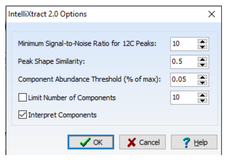

Figure 2: IntelliXtract 2.0 simplified parameter options using the new component detection algorithm.

In order to perform an accurate comparison of time required for processing and components found, both versions of the algorithm were installed on the same computer system and tested using the same data files. Nearly all settings remained as defaults except those which have a direct correlation: 1) Peak S/N Threshold (IX 1.0) to Maximum Signal-to-Noise Ratio for IX Peaks (IX 2.0); and 2) Component Window Value (IX 1.0) to Peak Shape Similarity (IX 2.0). These values were set to the defaults in IX 2.0 (10 and 0.5 respectively).

LC-MS Results

Data files from a variety of different sources were used in this algorithm comparison including pharmaceutical, environmental, and metabolic examples. Different types of instrumental resolution were also assessed. In the first comparison, time and number of components found were evaluated in LC-MS datasets (Table 1). In general, the time for IX 2.0 to complete analysis was less. In cases where the run time was greater in IX 2.0 than 1.0, a significant decrease in overall user analysis time was still seen, as there were usually fewer components (up to four times fewer) generated; reducing the time the user would spend analyzing the components.

	IX 1.0		IX 2.0		
Sample Type	# Components	Time	# Components	Time	Resolution
Pharmaceutical	193	8 s	194	31 s	High
Pharamceutical	31	10 s	15	8 s	Low
Metabolic	308	12 s	104	11 s	Low
Metabolic	355	8 s	81	26 s	High
Enviromental	1069	87 s	806	772 s	High
Enviromental	660	9 s	335	34 s	High

Table 1: Comparison of IX 1.0 and 2.0 by number of components and time for analysis for different sample types and resolutions using LC-MS data.

Both IX algorithms create extracted ion chromatograms (XICs) for each component. While the table above gave a quantitative assessment of the components generated, a visual display is also important in order to assess if any critical peaks were missed in the analysis. A subset of the samples are shown below (Figures 3 and 4) to illustrate the variance seen between IX 1.0 and 2.0 for different sample types and MS resolutions.

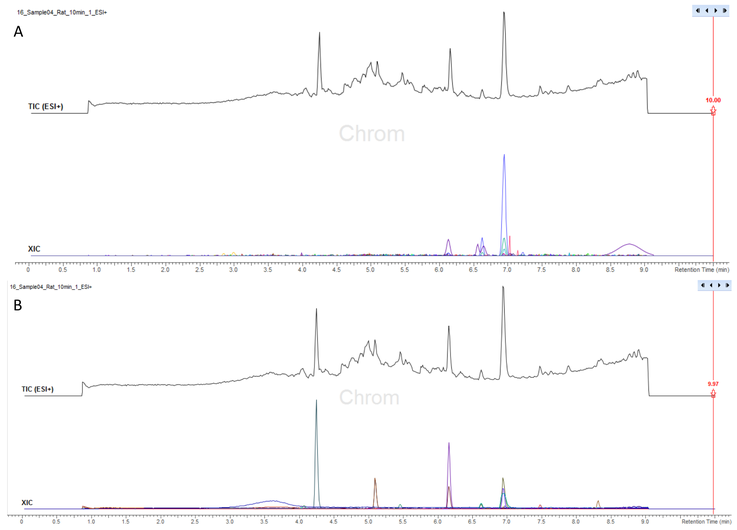

Figure 3: Metabolic data collected on a high resolution MS instrument was run through A) IX 1.0 and B) IX 2.0. IX 2.0 was found to generate approximately one-third fewer components and identify many of the components missed in IX 1.0.

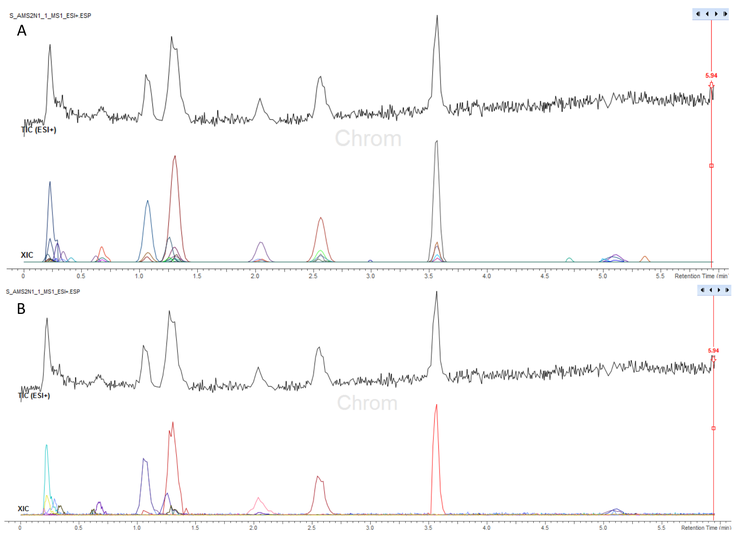

Figure 4: Low resolution, pharmaceutical MS data was also assessed in both A) IX 1.0 and B) IX 2.0. In the case of this low resolution data, nearly half the number of components were generated in IX 2.0 in approximately one-third of the time. The components generated from IX 2.0 contain all components that were in the sample run, illustrating the generation of false-positives in IX 1.0.

GC-MS Results

The previous IX 1.0 algorithm was not capable of generating reliable componentization for GC-MS data wherein the algorithm was unable to determine the monoisotopic peak. Another benefit of the new IX 2.0 algorithm is the capability of accurately analyzing GC-MS data. This further increases the analysis proficiency of the algorithm to increase utility while reducing the time needed for analysis. In order to illustrate the difference between IX 1.0 and 2.0 for GC-MS samples, a comparison was also run on a variety of different sample types from different industries showing the varying number of components generated and time for analysis as before (Table 2). In all but one case the number of components generated and time for analysis were less in IX 2.0 when compared with IX 1.0. Considering this, along with the improved algorithm to make the analysis more accurate, the benefit of IX 2.0 for GC-MS can be seen.

	IX 1.0		IX 2.0	
Sample Type	# Components	Time	# Components	Time
Petroleum	1343	50 s	573	32 s
Petroleum	527	20 s	215	33 s
Environmental	3283	511 s	783	100 s
Environmental	489	50 s	124	47 s
Fragrances	1048	235 s	186	93 s
Other	165	40 s	241	75 s
Other	380	244 s	27	71 s

Table 2: Comparison of IX 1.0 and 2.0 by number of components and time for analysis for different sample types and resolutions using GC-MS data.

Protein and Peptide Samples

Unlike IX 1.0, the second generation algorithm has the ability to analyze both peptide and protein data. For the purpose of illustrating this feature, one example of a protein sample has been chosen

The Marcotte Lab of the University of Texas at Austin has allowed for a selection of high resolution MS data to be downloaded by the public. The group largely studies the organization of proteins to better understand gene functions. The data obtained here involves wild-type yeast cells that were grown in yeast peptone dextrose to the log-phase. The data shown below is one of eight optimization experiments involving varying amounts of ammonium chloride separated on a strong cation exchange column followed by reverse-phase chromatography. The comparison below shows that IX 2.0 is able to identify far fewer components to better cover the chromatographic space occupied by the samples when compared to IX 1.0.

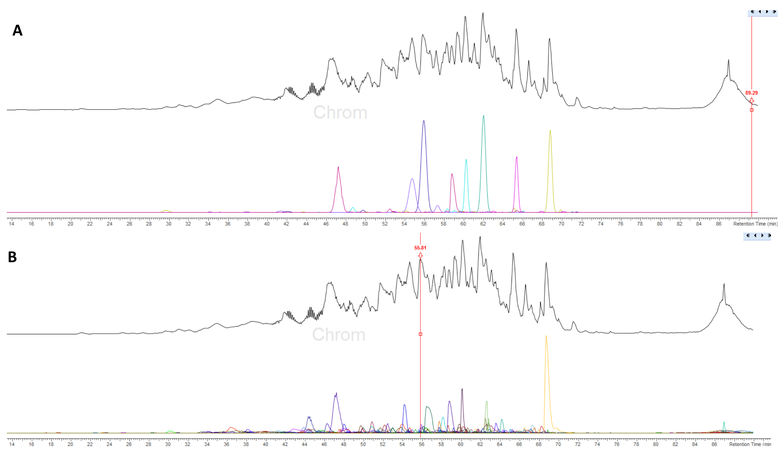

Figure 5: High resolution mass spectrometry data of wild-type yeast cells. A) Analyzed with IX 1.0 to identify 1264 components. B)Analyzed with IX 2.0 to 375 components found in the protein sample.

Conclusion

In this poster IX 1.0 was compared to IX 2.0. IX 2.0 saw a great reduction in processing options; making it far easier for users to operate. The new algorithm incorporating ion threads also lead to a reduced number of components generated, which greatly reduced user analysis/review time. With the reduction in number of components generated, we also saw an increase in the accuracy of the identified components. Often, analysis run time in IX 2.0 was less than that of IX 1.0. Coupled with the increased accuracy of component identification and ease of use, IX 2.0 greatly improves the user experience. When assessing LC-MS datasets, IX 2.0 was able to find peaks/components that had been missed completely by the previous algorithm. In addition, the algorithm was tested and found to be useful in the analysis of GC-MS datasets as well as protein and peptide datasets.