Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Chromatography A, 1071 (2005) 263–269

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry[☆]

Stephen E. Reichenbach[a,*], Visweswara Kottapalli[b], Mingtian Ni[a], Arvind Visvanathan[b]

[a] *Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln, NE 68588-0115, USA*
[b] *GC Image, LLC, P.O. Box 57403, Lincoln, NE 68505-7403, USA*

Abstract

This paper describes a language for expressing criteria for chemical identification with comprehensive two-dimensional gas chromatography paired with mass spectrometry (GC×GC–MS) and presents computer-based tools implementing the language. The Computer Language for Indentifying Chemicals (CLIC) allows expressions that describe rules (or constraints) for selecting chemical peaks or data points based on multi-dimensional chromatographic properties and mass spectral characteristics. CLIC offers chromatographic functions of retention times, functions of mass spectra, numbers for quantitative and relational evaluation, and logical and arithmetic operators. The language is demonstrated with the compound-class selection rules described by Welthagen et al. [W. Welthagen, J. Schnelle-Kreis, R. Zimmermann, J. Chromatogr. A 1019 (2003) 233–249]. A software implementation of CLIC provides a calculator-like graphical user-interface (GUI) for building and applying selection expressions. From the selection calculator, expressions can be used to select chromatographic peaks that meet the criteria or create selection chromatograms that mask data points inconsistent with the criteria. Selection expressions can be combined with graphical, geometric constraints in the retention-time plane as a powerful component for chemical identification with template matching or used to speed and improve mass spectrum library searches.

Keywords: Gas chromatography; Comprehensive two-dimensional gas chromatography; Mass spectrometry; GC–MS interpretation; Chemical analysis

1. Introduction

Comprehensive two-dimensional gas chromatography (GC×GC) paired with high-speed mass spectrometry (GC×GC–MS) produces data that provides rich information for identifying the chemical constituents of a sample. The two technologies are a highly compatible couple. Mass spectrometry (MS) provides “masses of the ionized molecule and its fragments” [1]. An experienced chemist can infer molecular sub-structures and chemical identity from a mass spectrum. However, one of the principal limitations of MS is that it does not separate chemical compounds, so the mass spectra of mixtures are difficult (or impossible) to interpret, particularly in the presence of noise. GC×GC is a powerful separation technology that achieves an order of magnitude increase in separation capacity and a significant increase in signal-to-noise ratio (SNR) over conventional gas chromatography (GC) [2]. GC×GC provides a two-dimensional chemical ordering (by retention times) that is useful for recognizing individual chemical compounds and chemical groups [3], but GC×GC does not provide structural information for chemical identification. One of the principal challenges presented by GC×GC is that it produces distinct separations of thousands of chemicals, so identifying the large number of constituents is difficult. GC×GC is important for MS because the highly effective separations significantly reduce the problems of mass spectral mixing; and, MS is important for GC×GC because it provides structural information for chemical identification.

[☆] This work was supported by the National Science Foundation, Award 0312173.

[*] Corresponding author. Tel.: +1 402 472 5007; fax: +1 402 472 7767.
E-mail address: reich@cse.unl.edu (S.E. Reichenbach).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.08.125

Several types of approaches have been used for chemical identification with GC×GC and MS, including: library search, pattern matching, and rule-based systems [4]. In library search, sample data are compared to reference data with associated structural information. Library search has proven useful for chemical identification with MS [5]. In pattern matching, data for the sample are matched with previously observed patterns of data (e.g., developed with a training set). Pattern recognition has proven useful for chemical identification with GC×GC retention-time templates [6]. Rule-based systems most closely resemble the approach that an experienced analytical chemist uses to deduce chemical identity. Rules express the reasons or criteria for chemical identification. Welthagen et al. used a rule-based approach based on GC×GC retention times and MS fragmentation patterns to produce preliminary classification of compound classes in the analysis of airborne particulate matter [7].

This paper develops a language for expressing rules for chemical identification with GC×GC–MS, such as were used by Welthagen et al. [7], and presents computer-based, interactive tools implementing the language. The Computer Language for Identifying Chemicals (CLIC) allows expressions that describe rules or constraints based on multi-dimensional retention times and mass spectral characteristics. CLIC offers chromatographic functions of retention times, functions of mass spectra, numbers for quantitative and relational evaluation, and logical and arithmetic operators. A software implementation of CLIC provides a calculator-like graphical user-interface (GUI) for building and applying expressions. From the selection calculator, expressions can be specified and used to select chromatographic peaks that meet the criteria or to create selection chromatograms that mask data points inconsistent with the criteria. Selection expressions can be combined with graphical, geometric constraints in the retention-time plane as a powerful component for chemical identification with pattern matching or used to speed and improve mass spectrum library searches.

Section 2 of this paper describes GC×GC–MS data, introduces the semantics of the language, and presents the formal syntax. Section 3 demonstrates the language with examples, using the search criteria described by Welthagen et al. [7]. Section 4 presents a software implementation with a calculator-like GUI for building and applying expressions. Section 5 considers applications for the language.

2. Language definition

2.1. GC×GC–MS data

GC×GC separates chemical species with two capillary columns interfaced by a modulator [10,11]. A two-stage thermal modulator (e.g., the KT2004 loop modulator from Zoex Corporation, Lincoln, NE, USA) employs temperature changes to trap, compress, and inject successive portions of the first-column eluents into a second, shorter column. Compression of the first-column eluents allows fast chromatography in the second column, where as many as 10 or more peaks can be separated in a few seconds. Structure-retention relations in GC×GC cause compounds of the same class to elute in ordered patterns that are useful for identifying chemicals. Compression also increases the chromatographic signal-to-noise ratio by an order of magnitude, thus improving detection and quantification of minor components [12]. The eluents of the second column can be input to a high-speed mass spectrometer to produce a data stream rich with information for identifying chemical constituents of highly complex mixtures [13].

GC×GC–MS data is structured as a three-dimensional array. The first dimension is the retention time for the first-column separation. The size of the first dimension of the array is the number of thermal modulation cycles during the data acquisition period. For example, with a modulation cycle of 5 s and a data acquisition period of 30 min, the first dimension size is 360 cycles and the index scale is 5 s/cycle. The second dimension is the retention time for the second-column separation. The size of the second dimension of the array is the length of the thermal modulation period multiplied by the sampling rate. For example, with a modulation cycle of 5 s and a sampling rate of 100 Hz, the second dimension size is 500 data points and the index scale is 0.01 s. The third dimension is the mass-to-charge ratio (m/z) of the mass spectrum. The size of the third dimension is the number of discrete m/z intervals present in the mass spectra and the index scale is the size of the m/z interval (typically 1). The third-dimension of the array can be stored in a sparse format if only significant intensity values of each mass spectrum are retained. Each value in the three-dimensional array is a measure of intensity for the indicated retention times and $m/$z.

GC×GC–MS data can be treated as a multi-channel digital image, with the mass spectrum at each chromatographic data point treated as a multi-channel picture element or *pixel*. The pixels can be arranged so that the abscissa (X-axis, left-to-right) is the elapsed time for the first-column separation and the ordinate (Y-axis, bottom-to-top) is the elapsed time for the second-column separation. Fig. 1 illustrates a total intensity chromatogram (TIC) of weathered gasoline [8], formed by totaling all intensity values at each pixel, viewed in three-dimensional perspective with log scaling. Each resolved chemical substance in a sample produces a small *blob* or cluster of pixels with larger intensity values than the surrounding background. Each blob has a *peak pixel*—the blob pixel with the largest total intensity. In Fig. 1, the smaller values of the background are colorized blue and the larger values of the blobs are green and red showing increasing total intensity values. Ion chromatograms can be constructed with intensities for individual m/z channels and ion-range chromatograms can be constructed with summed intensities for m/z channel range(s). Clearly, determining the chemical identity or chemical group for each of the many chromato-

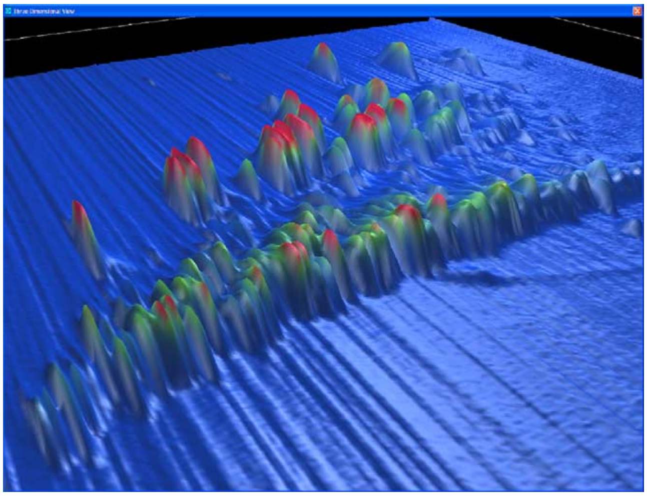


Fig. 1. A colorized, three-dimensional image view of a total intensity chromatogram from GC×GC–MS data [8]. This image was produced by GC Image™ software [9] and is best viewed in full color.

graphic peaks is a difficult task for which MS data is highly useful.

2.2. Selection language semantics

Functions characterizing chromatographic properties and mass spectral characteristics are key features of the language. Currently, CLIC has five functions, listed in Table 1, which can be applied in one of three expression modes, listed in Table 2. The context of the expression modes and semantics of the functions are relatively straightforward. For example, in pixel expression mode, Retention(1) returns the retention time on the first column for a pixel in the chromatogram. In blob-peak expression mode, Intensity(14) returns the intensity value for $m/z = 14$ in the mass spectrum of a blob's peak pixel (i.e., the pixel that has the largest total intensity of all pixels in the blob).

The functions are used in expressions with comparative operators to express the selection criteria. The relational operators are: less than, less than or equal to, greater than, greater than or equal to, equal to, and not equal to. For example, in blob-peak expression mode, Retention(2) $<$ 1.0 selects all blobs for which the second-column retention time of the blob pixel with peak total intensity is less than 1.0 s. In GC×GC data with a 5 s modulation period, this would select all blobs for which the peak is in the first 20% of the secondary chromatograms. In pixel expression mode, Ordinal(14) $=$ 1 would select all pixels for which the intensity value of the pixel's mass spectrum at $m/z = 14$ is the largest intensity value in that mass spectrum.

CLIC provides for addition (+), subtraction (–), and arithmetic negation (–) of values and parentheses for grouping arithmetic terms. For example, in blob-peak expression mode, (Percent(14) $+$ Percent(28) $+$ Percent(42)) $<$ 50 selects all blobs for which the sum of the intensities at $m/z = 14$, 28, and 42 is less than 50% of the total intensity for the blob peak pixel.

CLIC provides logical-and (&), logical-or (|), and logical-negation (!) operations and parentheses for grouping logical elements. For example, in pixel expression mode, (Relative(57) $>$ 20) & (Retention(2) $>$ 2.0) selects all pixels for which the intensity at $m/z = 57$ is greater than 20% of the largest intensity value in the pixel mass spectrum and the second-column retention time is greater than 2 s. As in the C programming language [14], non-zero operands are treated

Table 1
Selection language functions

Selection mode	Description
Retention (*dimension*)	Returns the retention time of the current object (either pixel or blob) with respect to the chromatographic column indicated by the dimension parameter (either 1 or 2 for GC×GC). For both blob-peak and blob-integration modes, the function returns the retention time of the peak pixel. Retention time for $dimension = 1$ is expressed in minutes and retention time for $dimension = 2$ is expressed in seconds.
Intensity (*channel*)	Returns the intensity value of the indicated channel (m/z in a mass spectrum) in the multi-channel intensity array of the current object (either pixel or blob). If the indicated $channel = 0$ (or null), the function returns the total intensity.
Ordinal (*channel*)	Returns the ordinal position of the indicated channel (m/z in a mass spectrum) in the intensity-ordered multi-channel array of the current object (either pixel or blob).
Percent (*channel*)	Returns the intensity value of the indicated channel (m/z in a mass spectrum) in the multi-channel intensity array of the current object (either pixel or blob) as a percentage of the total intensity of the array.
Relative (*channel*)	Returns the intensity value of the indicated channel (m/z in a mass spectrum) in the multi-channel intensity array of the current object (either pixel or blob) as a relative percentage of the largest intensity value of the array.

Table 2
Selection evaluation modes

Selection mode	Description
Pixel	Evaluate the function with respect to each pixel or data point
Blob peak	Evaluate the function with respect to each blob, considering the peak pixel (i.e., with the largest total intensity) in that blob
Blob integration	Evaluate the function with respect to each blob, considering the retention times of the peak pixel in that blob and the multi-channel intensity array integrated from all pixels in that blob

Table 3
Expression language grammar

```
<Expression> ::= <AndExpression>
      |   <Expression> | <AndExpression>
<AndExpression> ::= <EqualityExpression>
      |   <AndExpression> & <EqualityExpression>
<EqualityExpression> ::= <RelationalExpression>
      |   <EqualityExpression><EqualityOperator><RelationalExpression>
<EqualityOperator> ::= = | !=
<RelationalExpression> ::= <AdditiveExpression>
      |   <RelationalExpression><RelationalOperator><AdditiveExpression>
<RelationalOperator> ::= < | <= | > | >=
<AdditiveExpression> ::= <MultiplicativeExpression>
      |   <AdditiveExpression><AdditiveOperator><MultiplicativeExpression>
<AdditiveOperator> ::= + | −
<MultiplicativeExpression> ::= <ValueExpression>
      |   <MultiplicativeExpression><MultiplicativeOperator><ValueExpression>
<MultiplicativeOperator> ::= * | /
<ValueExpression> ::= ( <Expression> )
      |   <UnaryOperator><ValueExpression>
      |   <Function>
      |   <Number>
<UnaryOperator> ::= + | − | !
<Function> ::= <FunctionName>(<Integer>)
<FunctionName> ::= Retention | Intensity | Ordinal | Percent | Relative
<Number> ::= <Integer>
      |   <Integer>.
      |   <Integer>.<Integer>
      |  . <Integer>
<Integer> :: = <Digit>
      |   <Integer><Digit>
<Digit> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```

as true, zero operands are treated as false, true outputs evaluate to 1, and false outputs evaluate to 0. This allows counting of true conditions by arithmetic addition, as will be illustrated in one of the examples of Section 3.

Section 2.3 presents the formal grammar for CLIC. More extensive examples are provided in Section 3.

2.3. *Formal grammar*

CLIC is a context-free language. Table 3 presents the generative grammar that defines the set of possible expressions of the language. The grammar is presented in Backus–Naur Form (BNF). Each production rule consists of a non-terminal source, followed by “::=” (which can be read as “is defined as”), followed by a right-hand-side consisting of one or more alternative specifications separated by “|” (which can be read as “or”) [14]. Non-terminals are enclosed with “<” and “>” and terminal symbols are in bold.

3. Example selection expressions

This section defines selection expressions for the descriptive criteria for compound class identification used by Welthagen et al. [7] and for chemical identifications in a Grob mix. These expressions are presented for illustrative purposes. Different rules would be required for different chemical samples and different GC×GC conditions. For each of seven compound classes, the selection rules for MS and GC×GC are described and then written as selection expressions. For all of these rules, MS channels m/z <50 are ignored.

3.1. *Alkanes*

Criteria: Base peak 57 or 71 and second largest peak 71 or 57. No time rule is needed for this group, but a retention window 1.0–1.5 s can be used.

Selection expression: (Ordinal(57) <= 2) & (Ordinal (71) <= 2) & (Retention(2) >= 1) & (Retention(2) <= 1.5).

3.2. *Alkenes and cycloalkanes*

Criteria: Base peak 55 or 69 with both present and with three of peaks from set {56, 57, 70, 83, 97} greater than 15% relative intensity. Must be between 1 and 2 s on second dimension.

Selection expression: ((Ordinal(55) = 1) | (Ordinal(69) = 1)) & (Intensity(55) > 0) & (Intensity(69) > 0) & (((Relative(56) > 15) + (Relative(57) > 15) + (Relative(70) > 15) + (Relative(83) > 15) + (Relative(97) > 15)) >= 3) & (Retention(2) >= 1) & (Retention(2) <= 2).

3.3. *n-Alkane acids*

Criteria: Base peak 60 and second largest peak 73. No time rule.

Selection expression: (Ordinal(60) = 1) & (Ordinal (73) = 2).

3.4. *Alkyl-substituted benzenes*

Criteria: (1) Peak 91 greater than 15% relative intensity and greater than peak 77 and Peak 77 greater than 5% relative intensity. No time rule is needed, but generally greater than 2 s on second dimension with exceptions less than 1700 s on first dimension. (2) Compounds with mass 77 greater than 25% relative intensity. Less than 2 s on second dimension and less than 1700 s on first dimension.

Selection expression: ((Relative(91) > 15) & (Intensity(91) > Intensity(77)) & (Relative(77) > 5) & ((Retention(2) > 2) | (Retention(1) < 28.33))) | ((Relative(77) > 25) & (Retention(2) < 2) & (Retention(1) < 28.33)).

3.5. *Polar benzenes with or without alkyl groups*

Criteria: Peak 77 greater than 25% relative intensity. Greater than 2 s on second dimension.

Selection expression: (Relative(77) > 25) & (Retention (2) > 2).

3.6. *Partly hydrated naphthalenes and alkanyl-substituted benzenes*

Criteria: Peak 91 greater than 15% relative intensity, peak 77 greater than 5% relative intensity, and peak 128 greater than 10% relative intensity. No time rule needed, generally greater than 2 s on second dimension.

Selection expression: (Relative(91) > 15) & (Relative (77) > 5) & (Relative(128) > 10) & (Retention(2) > 2).

3.7. *Naphthalene and alkyl-substituted naphthalenes*

Criteria: (1) Peak 128 greater than 15% relative intensity and peak 77 greater than 5% relative intensity. (2) Peaks in set {141, 155, 169} relative intensity greater than 50%. Both: No time rule needed, generally greater than 2 s on second dimension.

Table 4
Selection expressions for a Grob Mix

Chemical name	Selection expression
2,3-Butanediol	Ordinal(45) < 3
Decane	Ordinal(57) < 3 & Retention(1) < 2.3
Undecane	Ordinal(57) < 3 & Retention(1) > 2.3 & Retention(2) < 1.8
1-Octanol	Ordinal(56) < 3 & Retention(1) < 3.7
Nonanal	Ordinal(57) < 3 & Retention(2) > 1.8
2-Ethylhexanoic acid	Ordinal(73) < 3
2,3-Dimethylphenol	Ordinal(122) < 3
2,6-Dimethylaniline	Ordinal(121) < 3
Methyl decanoate	Ordinal(74) < 3 & Retention(1) < 4.6
Methyl undecanoate	Ordinal(74) < 3 & Retention(1) > 4.6 & Retention(1) < 5.2
Methyl dodecanoate	Ordinal(74) < 3 & Retention(1) > 5.2
Dicyclohexylamine	Ordinal(138) < 3

Selection expression: (((Relative(128) > 15) & (Relative(77) > 5)) | ((Relative(141) > 50) | (Relative(155) > 50) | (Relative(169) > 50))) & (Retention(2) > 2).

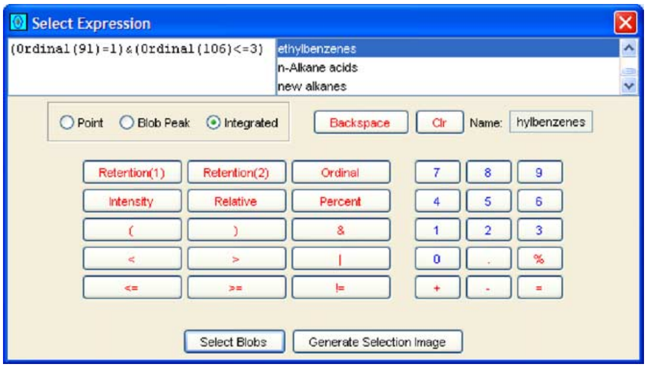
3.8. *Grob mix*

CLIC also can be used to identify individual compounds (as well as compound groups). For example, given a GC×GC–MS separation of a Grob mix of 12 chemicals listed in Table 4, several of the chemicals can be identified uniquely by a single constraint involving one of the two largest values in the mass spectrum and the others can be identified uniquely with the addition of simple retention time constraints. The process used to generate such identifying rules from a training set can be automated.

4. Implementation

CLIC has been implemented in software with a graphical-user interface (GUI) that resembles a calculator. The selection tool, pictured in Fig. 2, has a text box for the current expression and buttons for numbers, functions, arithmetic operators, logical operators, relational operators, and parentheses for grouping. The user builds the expression by typing directly in the text box and/or clicking buttons. Buttons for ‘Backspace’ and ‘Clear’ also are available to construct the expression.

Once constructed, the expression is applied in pixel mode, blob-peak mode, or blob-integration mode, depending on the setting of the radio buttons. In pixel selection mode, the expression is applied by clicking the ‘Generate Selection Image’ button, which produces an image in which pixels that evaluate to true are unchanged and pixels that evaluate to false are set to zero (i.e., masked). In blob-peak and blob-integration modes, the expression is applied by clicking the ‘Select Blobs’ button, which causes the blobs that evaluate to true to be selected and the blobs that evaluate to false to be deselected, or by clicking the ‘Generate Selection Image’ button, which produces an image in which pixels in


Select Expression

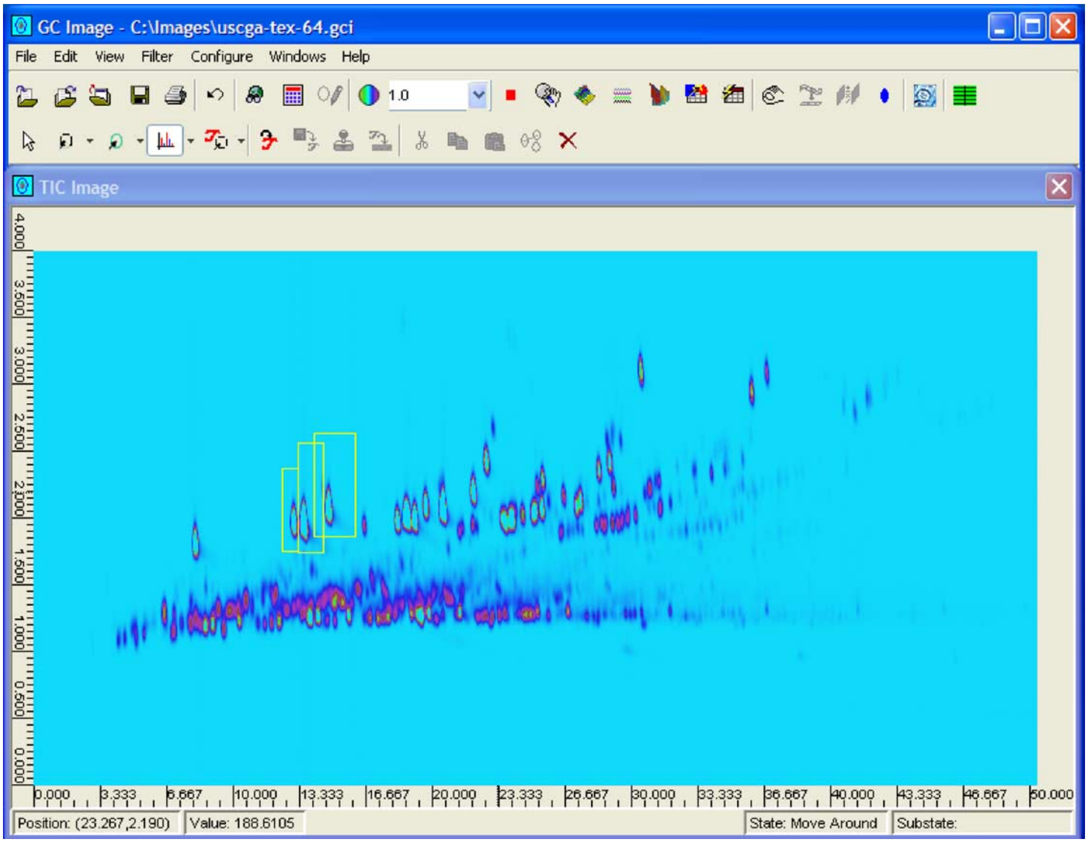
(Ordinal(91)=1)&(Ordinal(106)<=3)

ethylbenzenes
n-Alkane acids
new alkanes

○ Point ○ Blob Peak ● Integrated

Name: hylbenzenes

Fig. 2. A graphical user-interface (GUI) for selection expressions [15].



Position: (23.267,2.190) Value: 188.6105 State: Move Around Substate:

Fig. 3. A total ion chromatogram [15] with highlighted blobs selected by an expression for ethylbenzes [8].

blobs that evaluate to true are unchanged and pixels outside blobs that evaluate to true are set to zero (i.e., masked). Fig. 3 illustrates a total ion chromatogram with blobs selected by an expression for ethylbenzenes, “(Ordinal(91) = 1) & (Ordinal(106) <= 3))” applied in blob-integration mode.

Expressions can be stored and recalled with the selection calculator. To save an expression, type a name for the expression in the ‘Name’ box before evaluating the expression. To recall an expression, select the expression name from the list of named expressions. The user is prompted before overwriting a previously saved expression or reading to replace the current expression in the expression calculator text box.

5. Utilization

The implementation and GUI described in Section 4 support interactive selection of blobs and masking of chromatograms in GC×GC–MS images. Expressions in CLIC can be utilized for computer-automated identifications as well as interactive identification. For example, previous approaches to chemical identification by pattern matching have relied on retention time windows for one-dimensional GC data or templates in GC×GC [6]. Selection expressions could be used to prune or reduce the search space in computerized pattern matching, thereby reducing computational requirements, and to constrain the possible matches, thereby reducing erroneous identifications. Similarly, selection expressions could be used to subset libraries for faster library search and to improve library identifications.

Data from other multi-channel detectors, such as an atomic-emission detector (AED), can be structured similarly as multi-channel images, so the selection language can be adapted to other multi-channel analytical instruments. Different detector types require different selection expressions (as any differences in data acquisition may require different selection expressions), but the core language can support selection expressions for a wide range of multi-channel GC×GC data (and wide ranging acquisition conditions).

Additional features for CLIC are being designed and implemented. For example, it would be desireable to support chromatographic features other than retention time (e.g., blob volume or fractional response) and to allow m/z ranges in mass spectrum functions. Automating the process of generating expressions from a training set would be very useful.

References

[1] F.W. McLafferty, Interpretation of Mass Spectra, fourth ed., University Science Books, Herndon VA, 1996.
[2] W. Bertsch, J. High Resolut. Chromatogr. 23 (3) (2000) 167.
[3] G.S. Frysinger, R.B. Gaines, J. Sep. Sci. 24 (2) (2001) 87.

[4] M.E. Monk, J. Chem. Inf. Comput. Sci. 38 (6) (1998) 997.

[5] National Institute of Standards and Technology, MS Search Program, 2002.

[6] M. Ni, S.E. Reichenbach, Using edge pattern matching for automatic chemical identification in GC×GC, in: Automatic Target Recognition XIV, Proc. SPIE, vol. 5426, 2004.

[7] W. Welthagen, J. Schnelle-Kreis, R. Zimmermann, J. Chromatogr. A 1019 (2003) 233.

[8] G. S. Frysinger, Personal communication, texaco regular gasoline (obtained from local service station) 75% evaporated (evaporatively weathered in air until 25% of the original mass, simulating evaporation caused by a fire), 2003.

[9] B.W. Kernighan, D.M. Ritchie, The C Programming Language, second ed., Prentice-Hall, Englewood Cliffs NJ, 1988.

[10] E.B. Ledford Jr., C.A. Billesbach, J. High Resolut. Chromatogr. 23 (3) (2000) 202.

[11] J. Beens, M. Adahchour, R.J. Vreuls, J. Chromatogr. A 919 (1) (2001) 127.

[12] J. Phillips, et al., J. High Resolut. Chromatogr. 22 (1) (1999) 3.

[13] G.S. Frysinger, R.B. Gaines, J. High Resolut. Chromatogr. 22 (5) (1999) 251.

[14] F.G. Pagan, Formal Specification of Programming Languages, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[15] GC Image, LLC, GC Image™ software, Version 1.4, 2004.