INFORMATION-THEORETIC MASS SPECTRAL LIBRARY SEARCH FOR COMPREHENSIVE TWO-DIMENSIONAL GAS CHROMATOGRAPHY WITH MASS SPECTROMETRY

by

Arvind Visvanathan

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Stephen E. Reichenbach

Lincoln, Nebraska

August, 2008

INFORMATION-THEORETIC MASS SPECTRAL LIBRARY SEARCH FOR COMPREHENSIVE TWO-DIMENSIONAL GAS CHROMATOGRAPHY WITH MASS SPECTROMETRY

Arvind Visvanathan, Ph. D.

University of Nebraska, 2008

Adviser: Stephen E. Reichenbach

Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry (GCxGC-MS) combines two techniques providing increased separation capacity and enhanced capability for chemical identification. One of the most important methods for chemical identification is library search, which searches for an unknown mass spectrum in a library of known mass spectra to produce a list of potential matches ordered by match quality. Applications of compound identification include environmental monitoring, forensics, security, food and medicine.

This dissertation presents a new information-theoretic mass spectral library search technique for compound identification in GCxGC-MS and other MS applications. The method is based on a similarity measure between an unknown spectrum and a library spectrum involving the probability distribution functions of the intensities in the library and the noise in the data. The new method characterizes the library with an array of probability distribution functions of intensities as a function of mass-to-charge ratio. Each probability in the distribution function characterizes the fraction of spectra in the library having that intensity value at the given mass-to-charge ratio. The instrument noise is modelled with parameters estimated by statistically analyzing within individual GCxGC-MS peaks the intensity variations at each mass-to-charge ratio.

Experimental results demonstrate the effectiveness and robustness of the new information-theoretic mass spectral library search technique. In simulation experiments, random spectra from

the NIST/EPA/NIH Mass Spectral Library were corrupted with synthetic noise to generate random test spectra. Then, the corrupted spectra were submitted as unknowns for the library search using different search techniques. Experiments evaluated search performance with additive signal-independent noise, signal-dependent noise, (Johnson) colored noise, and spectral noise (from another spectrum selected randomly from the library). Other experiments evaluated search performance for real GCxGC-MS data. Search techniques were evaluated for many trials under each experimental condition by the Average Rank of the correct match in the ordered list of potential matches returned by the respective search techniques.

The new information-theoretic mass spectral library search technique performs better than NIST MS Search and Probability Based Matching (PBM) for all noise models; that is, the new search technique ranked the correct spectrum higher in the ordered list of potential matches than NIST MS Search and PBM for all noise models. In experiments with real data from GCxGC-Time-of-Flight-MS instruments and GCxGC-Quadrupole-MS instruments, the noise parameters were estimated by statistical analysis of mass spectral variations in multiple spectra of GCxGC-MS peaks and the weighted mean spectra of the peaks (added to the library as the correct match). In the experiments with real data, the information-theoretic mass spectral library search technique worked better than NIST MS Search and PBM in most cases.

Keywords: information theory, library search, compound identification, mass spectrum, noise model, similarity measure.

ACKNOWLEDGMENTS

I sincerely thank my advisor Dr. Stephen E. Reichenbach for continuing to provide both encouragement and being my mentor in the difficult journey over the duration of my doctoral program. The work presented in this dissertation could not be done without his encouragement, analysis, inspirational ideas, foresight, and enthusiasm. I also thank him for his patience which has improved my writing ability.

I also thank Dr. Myra B. Cohen, Dr. Ashok K. Samal, Dr. Stephen D. Scott, and Dr. Hendrik J. Viljoen for their service in my doctoral program committee and for their valuable comments on my dissertation.

A special thanks to Dr. Alex Henderson from the Surface Analysis Research Center at The University of Manchester who helped resolve many of my chemistry related queries. I also thank Dr. Qingping Tao from GC Image LLC and fellow students in the GCxGC group at the University of Nebraska - Lincoln for their valuable support.

I would also like to thank my fellow students (Andre, Jiazheng, Min, Shilpa, Victor, and Xue) whom I shared my office (122C Avery Hall), and all my friends without whom my doctoral program would not have been so much fun.

Finally, I would like to thank my parents for their support, guidance and their belief throughout my life. I dedicate my doctoral dissertation to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Comprehensive Two-Dimensional Gas Chromatography

Comprehensive two-dimensional gas chromatography (GCxGC) is an emerging instrumental technology for chemical separation that provides an order-of-magnitude increase in separation capacity over traditional gas chromatography (with a single column) and is capable of resolving several thousands of chemical compounds from a complex sample [1, 2]. Examples of sample mixtures include diesel fuel, gasoline, perfume, etc. Diesel fuel, for example, contains thousands of individual chemical compounds such as benzene, toluene, ethylbenzene, etc.

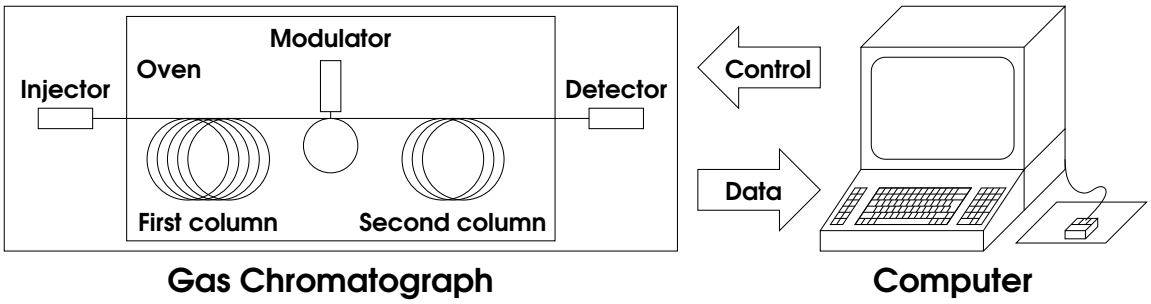

Figure 1.1: GCxGC instrumentation.

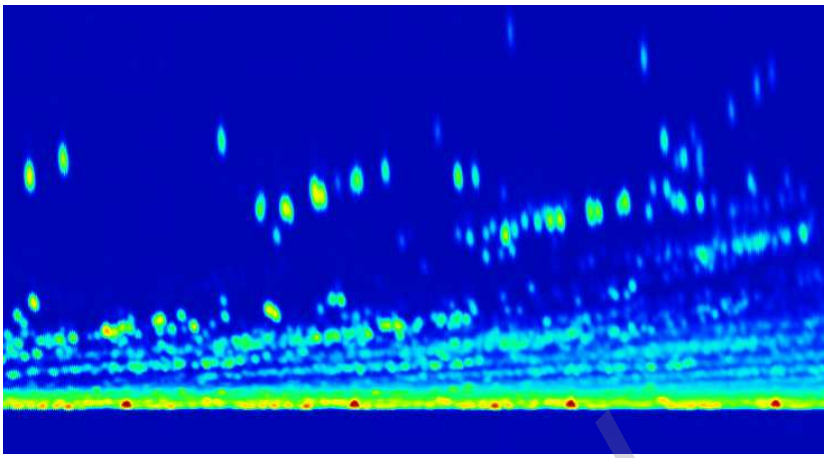

Figure 1.2: A portion of an example GCxGC image.

GCxGC separates chemical samples with two independent capillary columns interfaced with a modulator [3, 4], as illustrated in Figure 1.1. Typically, in a gas chromatograph, the sample is introduced by an injector, then heated as it passes through two independent separating columns. Partially resolved compounds from one column (carrying out the initial separation) are introduced into a second column of different selectivity [2]. For example, if one column is temperature selective, the compounds with a lower boiling point (lighter) exit the column earlier and the those with higher boiling point (heavier) exit later. The modulator between the two columns applies temperature changes to trap, focus, and inject successive portions of the first column eluent into a second, usually shorter column, allowing further separation of the sample [4]. Focusing increases the chromatographic signal-to-noise ratio by an order of magnitude, thus improving detection and quantification of trace components [5]. The eluent 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.

The GCxGC output can be displayed as an image with pixels 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 secondary column separation. Figure 1.2 illustrates an example GCxGC image for a diesel sample. Each individual chemical compound forms a cluster of pixels or a *blob* in GCxGC images as shown in Figure 1.2. The example image is generated by a temperature selective first column and a polarity selective second column. The compound structure (spatial arrangement of atoms in a compound molecule) is made of one or more chemical bonds between different atoms. The polarity (intermolecular force) of each bond (atom to atom link) within the compound determines the overall polarity of the compound. As the first column is temperature selective, the larger the abscissa of a blob the higher the boiling point of the compound forming the blob. The example image is generated by a polarity selective second column causing the compounds to exit the column according to their polarity (lowest to highest). As the second column is polarity selective, the larger the ordinate of a blob the higher the polarity of the compound forming the blob. GCxGC provides a two-dimensional chemical ordering (by retention times) that is useful for recognizing individual chemical compounds and chemical groups [6], but GCxGC does not provide structural information for chemical identification.

1.2 Mass Spectrometry

Mass spectrometry ("mass-spec" or MS) is an analytical technique used to measure the mass-to-charge ratio of ions [7]. The importance of mass-to-charge ratio, according to classical electrodynamics, is that two particles with the same mass-to-charge ratio move in the same path in a vacuum when subjected to the same electric and magnetic fields [7]. Mass spectrometry is used to evaluate the composition of a sample by generating a mass spectrum representing the masses of sample components. The mass spectrum is measured by a mass spectrometer.

The fact that different chemicals have different masses is exploited in a mass spectrometer to determine the chemicals in a sample mixture. For example, common salt (NaCl) may be vaporized

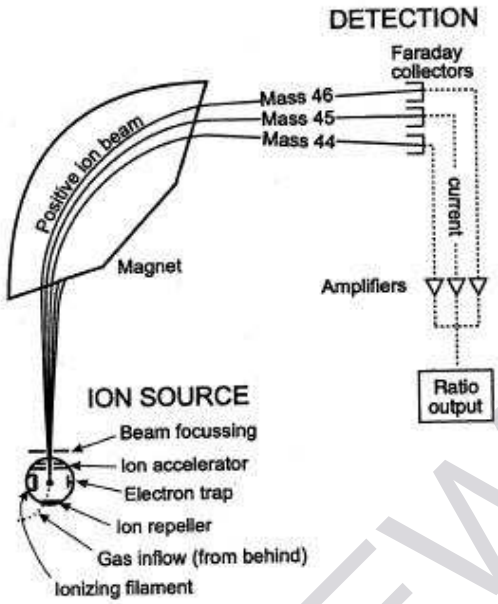

Figure 1.3: Mass spectrometry [8].

(into gas) and ionized (broken down) in the first phase of a mass spectrometer into electrically charged particles (Na^+ and Cl^-), called ions. Sodium ions have mass 23u (where u is the average molecular mass unit) and chloride ions have a mass of 35u or 37u (two different isotopes). The ions also have a charge, which allows their direction and speed to be changed with an electric or magnetic field. The ions are accelerated (usually in a vacuum) to a high speed by an electric field, after which they are directed into a magnetic field. The magnetic field applies force which deflects the ions to curve to differing degrees depending on their mass-to-charge ratio as illustrated in Figure 1.3 [8].

According to Newton's Second Law of Motion, the acceleration of a particle is inversely proportional to its mass. Thus, the magnetic field deflects the lighter ions more than the heavier ions. The detector measures the deflection of each resulting ion beam. From this measurement, the mass-to-charge ratios of all the ions produced in the source can be determined. From this informa-

tion it is possible to resolve the chemical composition of the original sample (*i.e.*, that both sodium and chlorine are present in the sample).

1.2.1 Types of Mass Spectrometers

The previous section described the working of one type of mass spectrometer (sector), but there are many types of mass spectrometers. All mass spectrometers have the following three components:

1. **Ion Source** - produces ions.
2. **Analyzer** - sorts the ions based on their masses.
3. **Detector** - measures relative intensities of different masses.

Some of the widely used mass analyzers in mass spectrometers include:

1. **Sector** - A sector field mass analyzer uses an electric and/or magnetic field to affect the path and/or velocity of the charged particles in some way [9].
2. **Time-of-flight** - Perhaps the easiest to understand is the time-of-flight (TOF) analyzer. It uses an electric field to accelerate the ions through the same potential and then measures the time they take to reach the detector. If the particles all have the same charge, the kinetic energies will be identical and their velocities will depend only on their masses. Lighter ions will reach the detector first [10].
3. **Quadrupole** - Quadrupole mass analyzers use oscillating electrical fields to selectively stabilize or destabilize ions passing through a radio frequency (RF) quadrupole field. A quadrupole mass analyzer acts as a mass selective filter [7].

1.2.2 Data Representation

A mass spectrum is an array of pairs of the form *(mass-to-charge ratio,intensity)*. Each mass-to-charge ratio is the mass of the particular ion divided by the unified atomic mass unit (unit of mass equal to the atomic mass constant, defined as one twelfth of the mass of a carbon-12 atom used to express masses of atomic particles [1 atomic mass unit = $1.6605402 \times 10^{-27} kg$] [11]) and its charge number (positive absolute value). Each mass-to-charge ratio has a corresponding intensity value which is the detector output. The detector output, for example, could be the number of ions at a particular mass-to-charge ratio.

$$m = \frac{mass}{|charge| \times 1.6605402 \times 10^{-27}} \tag{1.1}$$

where

m is the mass-to-charge ratio of an ion,

$mass$ is the mass of the ion, and

$charge$ is the charge of the ion, for example +1 for H_20^+.

The mass-to-charge ratios are measured relative to 1 atomic mass unit, for example, the mass-to-charge ratio of the ion H_20^+ is 18, where the mass of the ion is 18 and the charge is +1. Both the mass-to-charge ratios and intensity values are non-negative integers. The mass-to-charge ratios of a mass spectrum are unique and listed in an ascending order from smallest to largest value. If the mass-to-charge ratios are floating point values (high resolution mass spectrometer), the mass-to-charge ratios are binned (intensity values are integrated) to the nearest integer value. Every mass spectrum also has a *base peak*, which is the largest intensity valued mass-to-charge ratio. Intensity values of a mass spectrum are usually *normalized* to the base peak such that the base peak intensity is 999, *i.e.*, each intensity in the mass spectrum is scaled by the base peak intensity (*999/base peak intensity*). After normalizing, the intensity values are rounded to the closest integer value.

Intensity values at many mass-to-charge ratios are zero and may be omitted in the mass spectrum representation.

A mass spectrum can be represented in any of the following ways.

1. **Ordered Set**: If the mass spectrum has *n* pairs of the form *(mass-to-charge ratio,intensity)*:

$$S = \{(m_1, a(m_1)), (m_2, a(m_2)), ..., (m_n, a(m_n))\}, \tag{1.2}$$

$$m_i \in \mathcal{I}, \qquad m_i > 0 \qquad \forall 1 \leq i \leq n, \tag{1.3}$$

$$m_i < m_{i+1}, \qquad \forall 1 \leq i < n-1, \tag{1.4}$$

$$a(m_i) \in \mathcal{I}, \qquad a(m_i) \geq 0 \qquad \forall m_i, \tag{1.5}$$

where

S is the mass spectrum,

m_i is the i^{th} mass-to-charge ratio,

$a(m_i)$ is the intensity value at mass-to-charge ratio m_i, and

$\mathcal{I}$ is the set of integers.

For example, a mass spectrum of water can be represented as an ordered set as:

$$S(water) = \{(16, 9), (17, 212), (18, 999), (19, 5), (20, 3)\}. \tag{1.6}$$

For the mass spectrum of water, the *base peak* is 18, and the intensity of the spectrum at all mass-to-charge ratios other than in the range 16-20 are zero.

2. **String**: A mass spectrum can be represented as a string with individual mass-to-charge ratios separated from the intensity values by a comma and each pair of mass-to-charge ratios and

intensity values separated by a semicolon. If the mass spectrum has *n* pairs:

$$S = \text{“}m_1, a(m_1); m_2, a(m_2); ...; m_n, a(m_n)\text{”}. \quad (1.7)$$

For example, a mass spectrum of water can be represented in string form as:

$$S(water) = \text{“}16, 9; 17, 212; 18, 999; 19, 5; 20, 3\text{”}. \quad (1.8)$$

3. **Sparse Array**: A mass spectrum can be represented as two corresponding arrays of the same length. One array represents the mass-to-charge ratios and the other represents the corresponding intensity values:

$$S = M, A \quad (1.9)$$
$$M = [m_1, m_2, ..., m_n] \quad (1.10)$$
$$A = [a(m_1), a(m_2), ..., a(m_n)]. \quad (1.11)$$

For example, a mass spectrum of water can be represented as:

$$S(water) = M, A \quad (1.12)$$
$$M = [16, 17, 18, 19, 20] \quad (1.13)$$
$$A = [9, 212, 999, 5, 3]. \quad (1.14)$$

4. **Table**: A mass spectrum can be represented by a table with two columns. The first column represents the mass-to-charge ratios and the second column represents the corresponding intensity values. Table 1.1 illustrates a mass spectrum of water in tabular format.

5. **Graph**: A mass spectrum can be represented as a bar graph with mass-to-charge ratios along

Table 1.1: Tabular representation of a mass spectrum (water).

mass-to-charge ratio	intensity
16	9
17	212
18	999
19	5
20	3

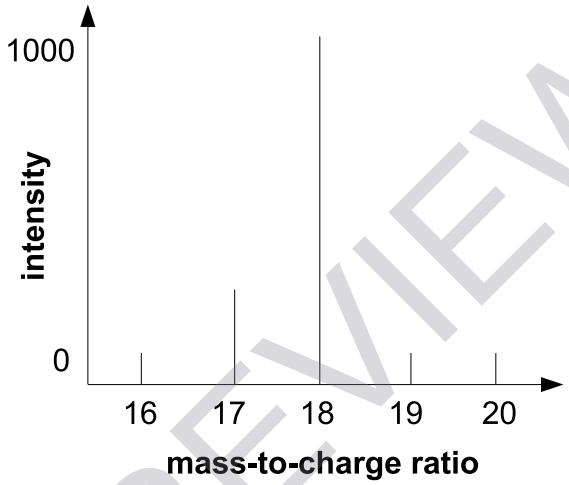

Figure 1.4: Graphical representation of the mass spectrum of water.

the X-axis and the intensity values on the Y-axis. Figure 1.4 illustrates a mass spectrum of water in graphical format.

1.3 Comprehensive Two-Dimensional Gas Chromatography - Mass Spectrometry

Comprehensive two-dimensional gas chromatography with mass spectrometry (GCxGC-MS) combines two techniques providing increased separation capacity and enhanced capability for chemical

identification. Similar to GCxGC, the pixels in a GCxGC-MS image 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. Each pixel in a GCxGC-MS image has an associated mass spectrum.

Figure 1.5(a) illustrates an example GCxGC-MS total ion chromatogram. Each pixel value in the total ion chromatogram (TIC) is the sum of all the intensity values of the mass spectrum associated with the pixel. Each individual compound in the sample produces a *blob* or a cluster of pixels with larger pixel values than the surrounding pixels in the TIC. Figure 1.5(a) also illustrates the blob apex mass spectrum of two different individual compounds. The blob apex mass spectrum is the mass spectrum of the pixel that has the highest TIC value for the blob or individual chemical compound. Clearly, the two mass spectra are different, *i.e.*, they have different intensity levels at corresponding mass-to-charge ratios.

Figure 1.5(b) illustrates mass spectra of two pixels from the same chemical compound (spectra from the same cluster of pixels) are similar, *i.e.*, the two spectra have similar relative intensity levels at corresponding mass-to-charge ratios, but differ in scale, with slight variations in relative magnitude because of instrument induced differences.

1.4 Compound Identification

1.4.1 Problem Statement

One of the most important tasks in mass spectrometry is compound identification. Figure 1.6 pictures library search for the compound identification problem. The input to compound identification is an unknown mass spectrum, for example, one of the blob mass spectra from GCxGC-MS data.