Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Chromatography A, 1086 (2005) 165–170

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Peak pattern variations related to comprehensive two-dimensional gas chromatography acquisition

Mingtian Ni[a,*], Stephen E. Reichenbach[b], Arvind Visvanathan[b], Joel TerMaat[c], Edward B. Ledford Jr.[c]

[a] GC Image, LLC, 216 N 11th Street, Suite 302, Lincoln, NE 68508, USA
[b] Department of Computer Science and Engineering, University of Nebraska, Lincoln, NE 68588, USA
[c] Zoex Cooperation., 2611 West M ST, Suite D, Lincoln, NE 68522, USA

Available online 6 July 2005

Abstract

Identifying compounds of interest for peaks in data generated by comprehensive two-dimensional gas chromatography (GC × GC) is a critical analytical task. Manually identifying compounds is tedious and time-consuming. An alternative is to use pattern matching. Pattern matching identifies compounds by matching previously observed patterns with known peaks to newly observed patterns with unidentified peaks. The fundamental difficulty of pattern matching comes from peak pattern distortions that are caused by differences in data acquisition conditions. This paper investigates peak pattern variations related to varying oven temperature ramp rate and inlet gas pressure and evaluates two types of affine transformations for matching peak patterns. The experimental results suggest that, over the experimental ranges, the changes in temperature ramp rate generate non-linear pattern variations and changes in gas pressure generate nearly linear pattern variations. The results indicate the affine transformations can largely remove the pattern variations and can be used for applications such as pattern matching and normalizing retention times to retention indices.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Comprehensive two-dimensional gas chromatography (GC × GC); Compound identification; Peak pattern matching; Peak pattern variation; Transformation model

1. Introduction

Comprehensive two-dimensional gas chromatography (GC × GC) combines the resolving power of two columns interfaced by a thermal modulator, offering significantly greater separation capacity than traditional one-dimensional GC [1]. GC × GC can separate thousands of different compounds, whereas it is difficult to distinguish a few hundred peaks in data generated by traditional one-dimensional GC. The great performance of GC × GC holds promise for many important applications such as environmental monitoring [2], petrochemical processing [3], and chemical warfare agent detection [4].

Given a chemical sample, the GC × GC output data can be represented, visualized, and processed as an image. In the image, each resolved compound produces a small two-dimensional peak with values larger than background values. Identifying compounds for peaks of interest is a critical task in GC × GC analysis. GC × GC images contain potentially thousands of peaks in complex patterns, making compound identification a challenging problem. Manually identifying compounds is tedious and time-consuming.

Several approaches have been used to automate the compound identification process in GC × GC analysis, including library search, rule-based techniques, and pattern matching [5]. In library search, sample data are compared to reference data with associated compound information in a library. Library search has proven useful for compound identification with GC × GC–MS (mass spectrometry) [6]. Rule-

[*] Corresponding author. Tel.: +1 402 474 2942; fax: +1 402 472 7767.
E-mail addresses: mni@cse.unl.edu (M. Ni); reich@cse.unl.edu (S.E. Reichenbach).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.06.033

based techniques try to relate a set of rules to each compound of interest. Rules express criteria for compounds based on various features such as peak retention times and peak statistics. 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].

Pattern matching identifies compounds by matching previously observed patterns with known peaks to newly observed patterns with unidentified peaks [8]. Peak pattern matching involves two peak patterns: a peak template (or template peak pattern) and a target pattern (or target peak pattern). A peak template is a set of annotated peaks. Annotated peaks have both computed features and annotated information. Computed features, such as peak location and volume, are computed from GC × GC images directly. Annotated information, such as compound name, are provided externally and are used for identifying and characterizing the peaks. A target peak pattern is a set of unannotated peaks that have only computed features. Determining annotated information for the target peak pattern is the objective of the compound identification process.

Given a peak template and a target peak pattern, peak pattern matching tries to establish as many correspondences as possible from peaks in the template to peaks in the target peak pattern. After peak correspondences are established, the annotated information carried by the peaks in the template is copied into the corresponding peaks in the target peak pattern. Consequently, all the matched compounds in the target peak pattern are identified.

The fundamental difficulty of the matching process comes from peak pattern distortions, which cause the same compound peaks to appear at different locations in different images. Peak pattern matching algorithms seek a transformation in some transformation space to remove the distortions. Two categories of distortions are distinguished: peak pattern variations and uncorrected distortions. Peak pattern variations are caused by differences in controllable data acquisition conditions such as oven temperature ramp rate and inlet gas pressure. Uncorrected distortions are caused by differences in unpredictable acquisition conditions such as column deterioration over time and instrument-to-instrument variations in physical parameters. Uncorrected distortions typically can not be modeled by practical transformations and are left as noise in the matching process. This paper investigates peak pattern variations related to oven temperature ramp rate and inlet gas pressure and affine transformations models for the variations.

Section 2 of this paper introduces the concept of transformation spaces and the two affine transformation models used in the experiments. Section 3 describes the two GC × GC data sets. Section 4 presents the experimental results. Section 5.

2. Transformation spaces

A transformation is a one-to-one mapping from the Euclidean space R^2 to itself. A transformation model gives the type information of a set of transformations. Each transformation model typically corresponds to a specific parametric form. A transformation is an instantiation of a transformation model. For example, the affine transformation model $t(a, b, e, c, d, f)$ (denoted *Affine-6*) has the following parametric form:

$$t(a, b, e, c, d, f)(x, y) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}. \qquad (1)$$

So, for example, $t(1.0, 0.0, 1.0, 0.0, 1.0, 1.0)$ is an affine transformation (which performs a vertical and horizontal shift by 1 unit). A transformation space is a set of transformations under a specific transformation model. It encodes the transformation model and the parameter ranges. For example, $\{t(a, b, e, c, d, f)|a \in [a_l, a_r], b \in [b_l, b_r], e \in [e_l, e_r], c \in [c_l, c_r], d \in [d_l, d_r], f \in [f_l, f_r]\}$ is an affine transformation space. The size of a transformation space is then determined by the dimensionality of the transformation model (the number of variables in the model) and the parameter ranges.

The transformation model is designed or selected based on the assumptions about the variations present among the peak patterns. For example, if it is assumed that the variations are translational, then the transformation model should be translation. However, if after translation, the peak patterns still do not match well, a more powerful transformation model must be used. Designing or selecting the transformation model is a challenging task. If the model is under-constrained, i.e., it has too many variables, then many inferred peak correspondences may be incorrect and searching the transformation space is computationally expensive. If the model is over-constrained, it may not be able to remove the variations effectively and establish the desired correspondences. The effectiveness and efficiency of a peak pattern matching technique primarily depends on the transformation space. A larger transformation space typically is more powerful for removing distortions. On the other hand, searching a larger space is more computationally expensive. In practice, it is desirable to select a transformation model that is just powerful enough to remove the existing variations. Given a transformation model, its parameter ranges can be determined by statistical estimation on training data [9].

Affine transformation models are used widely for aligning geometric patterns (images) due to their simplicity. The experiments in Section 4 assess the effectiveness of two affine transformation models, *Affine-6* and *Affine-4*, for removing the variations generated by changes in oven temperature ramp rate and inlet gas pressure. *Affine-4* is:

$$t(a, e, d, f)(x, y) = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}. \qquad (2)$$

3. Data sets

Two calibration data sets, *Oven-temperature* and *Gas-pressure*, were acquired at Zoex Corporation in February 2004. The two data sets contain a variety of chemical compounds, among which 10 compounds are used for evaluation: *1,2,4,5–tetramethylbenzene*, *1,2-dibromobenzene*, *1-decanol*, *1-undecanol*, *2-methylnaphthalene*, *dodecane*, *hexadecane*, *hexamethylbenzene*, *naphthalene*, and *tetradecane*.

The two data sets were generated by the same GC × GC unit with similar column configurations.

(i) First column: SPB-1, 15 m × 0.25 mm I.D. × 1.0 μm d.f.
(ii) Modulator tube: non-polar fused silica, 1.8 m × 0.1 mm I.D.
(iii) Second column: Supelcowax-10, 0.1 mm I.D., 0.1 μm d.f.

The length of the second column was 50 cm for *Oven-temperature* and 100 cm for *Gas-pressure*. For all runs, the oven temperature was programmed from 100 to 260 °C and the sampling rate was 200 Hz. For *Oven-temperature*, the inlet gas pressure was fixed to be 20 psi, the modulation period was 3 s, and the oven temperature ramp rate was varied from 2 to 11 °C/min in increments of 1 °C/min, generating 10 images. For *Gas-pressure*, the oven temperature ramp rate was fixed to 4 °C/min, the modulation period was 4 s, and the inlet gas pressure was varied from 17 to 24 psi in increments of 1 psi, generating eight images.

4. Experimental results

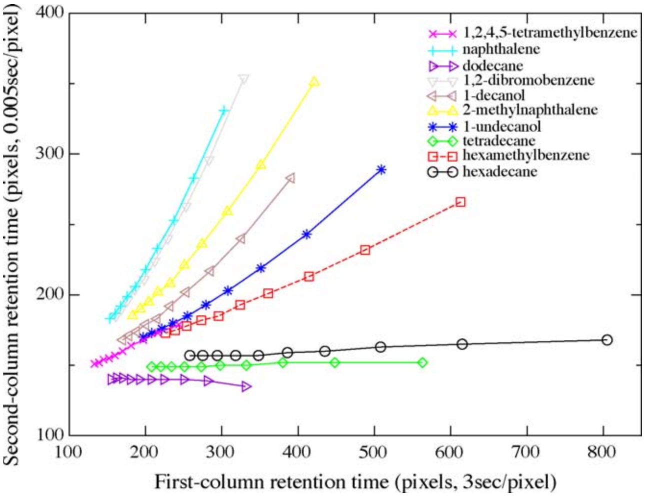

Fig. 1. Retention times vary with oven temperature ramp rate in *Oven-temperature*.

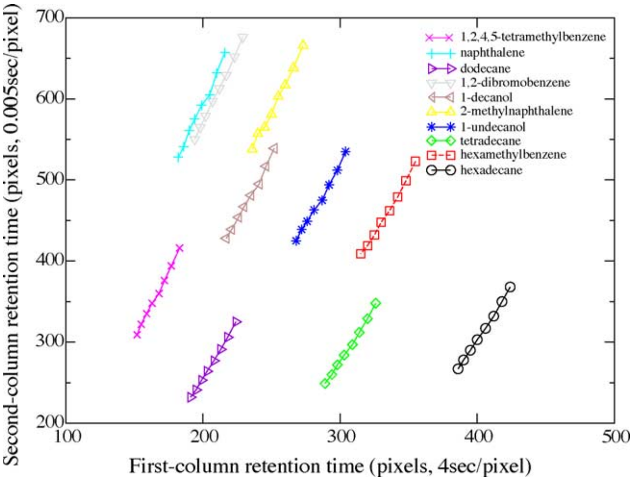

Fig. 2. Retention times vary with inlet gas pressure in *Gas-pressure*.

The peak patterns of *Oven-temperature* and *Gas-pressure* are illustrated in Figs. 1 and 2, respectively. In the figures, each line corresponds to a chemical compound. Each point shows the location (retention times) of a compound peak. Each sequence of points connected by a line shows the variation of the peak locations of a specific compound with oven temperature ramp rate (or inlet gas pressure). For example, in Fig. 1, the retention times for hexadecane vary for temperature ramp rate from (805,168) pixels or (40.25 min,0.84 s) at 2 °C/min to (258,157) pixels or (12.90 min,0.785 s) at 11 °C/min (as faster ramp rates cause shorter retention times). And, in Fig. 2, the retention times for hexadecane vary for inlet gas pressure from (424,368) pixels or (28.27 min,1.84 s) at 17 psi to (386,267) pixels or (25.73 min,1.335 s) at 24 psi (as higher gas pressures cause shorter retention times).

Over the experimental ranges, the peak retention times vary nearly linearly with inlet gas pressure in Fig. 2, but the retention times vary non-linearly with oven temperature ramp rate in Fig. 1. However, note that the *Oven-temperature* dataset has much larger retention time ranges. For example, for *Oven-temperature*, the ratios of the longest retention times (for the slowest temperature ramp rate) to the shortest retention times (for fastest temperature ramp rate) for 1-undecanol are 2.6 for the first column and 1.7 for the second column. For *Gas-pressure*, the retention time ratios of the longest to shortest retention times for 1-undecanol are just 1.1 and 1.3. Over a smaller range, such as might be expected from small run-to-run changes in experimental conditions, the variations for both are *Oven-temperature* and *Gas-pressure* are relatively linear.

The two affine transformation models, *Affine-6* and *Affine-4*, are evaluated on *Oven-temperature* and *Gas-pressure* for removing the peak pattern variations generated by varying oven temperature ramp rate and inlet gas pressure. The peaks of the 10 selected compounds form a peak pattern for each image in the data sets. For each pair of peak patterns within each data set, a least-squares optimal transformation is computed based on the known peak correspondences.

Assume that $\Gamma = \{P^i\}_{i=1}^{m}$ is one of the two data sets and each P^i is a peak pattern in Γ. The least-squares op-

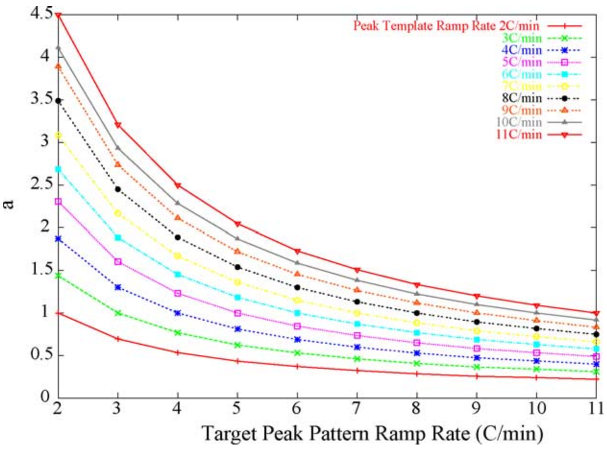

Fig. 3. Transformation parameter distribution of *a* for *Affine-4* (Eq. (2)) and *Oven-temperature*.

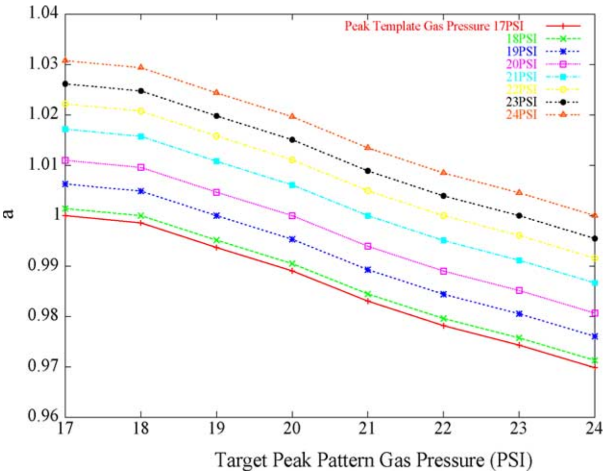

Fig. 4. Transformation parameter distribution of *a* for *Affine-4* (Eq. (2)) and *Gas-pressure*.

Table 1
Average residual errors (in pixels distance) over all template-target pairs

Compound	*Oven-temperature*		*Gas-pressure*	
	Affine-6	*Affine-4*	*Affine-6*	*Affine-4*
1,2,4,5–Tetramethylbenzene	3.69	6.16	1.35	1.44
1,2-Dibromobenzene	1.71	3.11	1.19	1.28
1-Decanol	3.65	4.66	1.88	1.98
1-Undecanol	4.40	3.88	2.88	2.84
2-Methylnaphthalene	2.32	5.31	2.63	2.81
Dodecane	1.56	5.40	1.18	1.36
Hexadecane	1.77	5.49	1.33	1.62
Hexamethylbenzene	5.09	7.16	1.54	1.88
Naphthalene	3.78	2.29	2.65	2.77
Tetradecane	2.80	1.80	0.72	0.77
Average	3.08	4.52	1.74	1.88

timal transformation $t^{i,j}$ from peak pattern $P^i = \{p_k^i\}_{k=1}^n$ to $P^j = \{p_k^j\}_{k=1}^n$ is given by $\operatorname{argmin}\{d_E(t(P^i), P^j)\}$, where $t(P^i)$ denotes the transformed peak pattern of P^i by transformation t. The Euclidean distance $d_E(P^i, P^j)$ between P^i and P^j is defined as $(1/n)\sum_{k=1}^n \|p_k^i - p_k^j\|$, where $\|p_k^i - p_k^j\|$ is the Euclidean distance between point p_k^i and p_k^j.

Oven-temperature contains 10 peak patterns (images). In calculating least-squares optimal transformations, each pattern is used as a peak template and all the 10 patterns are used as target peak patterns, generating a total of $10 \times 10 = 100$ transformations for each of the two affine transformation models. Similarly, *Gas-pressure* generates 64 transformations for each transformation model.

The optimal transformation parameters for *Oven-temperature* have larger ranges and more evident nonlinearities than for *Gas-pressure*. For example, the optimal values of parameter *a* in Eq. (2) for *Affine-4* for *Oven-temperature* and *Gas-pressure* are shown in Figs. 3 and 4. In Fig. 3, *a* varies from about 0.4 to 4.5, whereas *a* only varies from about 0.97 to 1.03 in Fig. 4. The specific values for the other parameters of Eq. (1) and for the parameters of Eq. (2) are applicable for this data, but not more generally, and so are not presented.

Table 1 reports the average residual errors for the 10 compounds after applying the optimal transformations. Assume that $\{p_k^i\}_{i=1}^m$ are the peaks generated by a compound in the sequence of images in Γ. Then the average residual error for this compound is $\binom{m}{2}^{-1}\sum_{i\neq j}\|t^{i,j}(p_k^i) - p_k^j\|$.

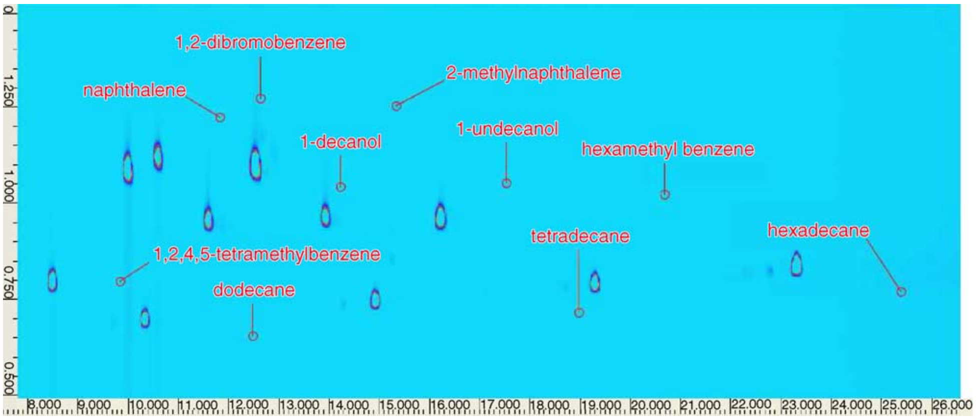

Fig. 5. Peak template from *Oven-temperature* at 4 °C/min overlaid on an image from *Oven-temperature* at 6 °C/min.

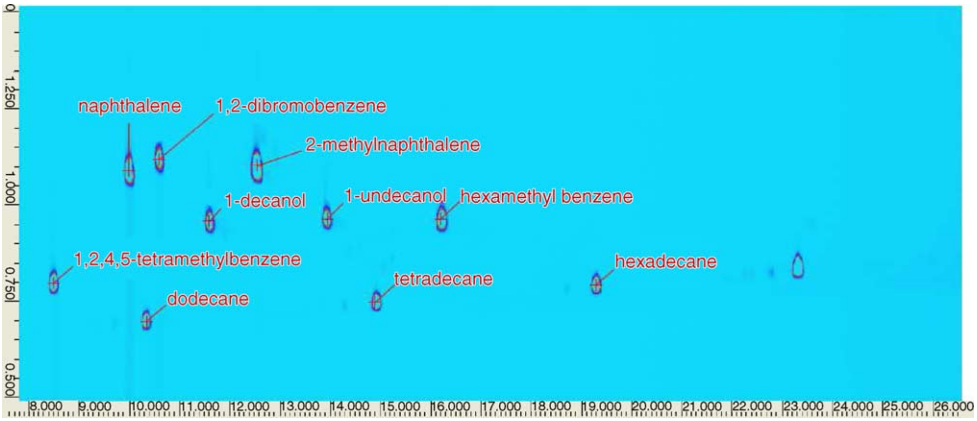

Fig. 6. Peak template from *Oven-temperature* at 4 °C/min with optimal *Affine-4* transformation overlaid on an image from *Oven-temperature* at 6 °C/min.

Several observations can be made based on the results shown in Table 1:

(i) The average residual errors are relatively small compared to the peak pattern variations. Roughly speaking, both *Affine-6* and *Affine-4* effectively removed the peak pattern variations in the two data sets. This is the most important conclusion from these experiments. As illustrated in Figs. 5 and 6, even the simpler *Affine-4* transformation provides an excellent matching between a peak template and target peak pattern for datasets acquired with quite different conditions. Fig. 5 shows the peak template extracted from the *Oven-temperature* run at 4 °C/min overlaid on an image from *Oven-temperature* at 6 °C/min. The retention times of the peaks in the template are quite different than the retention times of the peaks in the image. Fig. 6 shows the template peak points after the least-squares optimal *Affine-4* transformation. Template matching with the *Affine-4* transformation is effective even between peak patterns acquired under very different conditions.

(ii) From *Affine-6* to *Affine-4*, the average residual error only decreases by 1.44 pixels for *Oven-temperature* and by 0.14 pixels for *Gas-pressure*. So, for applications in which computational time is an important issue, *Affine-4* may be a better choice for peak pattern matching to avoid the computation related to the two additional parameters in *Affine-6*.

(iii) The residual errors for *Oven-temperature* are larger than those for *Gas-pressure*, which suggests that affine transformations are less effective in removing the non-linear pattern variations over the larger ranges related to oven temperate ramp rate changes. Fig. 7 plots the residual errors after the least-squares optimal transformations of the template from *Gas-pressure* 21 psi for each of the target peak sets. All of the errors are small. Fig. 8 plots the residual errors after the least-squares optimal transformations of the template from *Oven-temperature* 6 °C/min for each of the target peak sets. The errors are relatively small for peak patterns acquired under similar conditions and for peak patterns acquired at a more

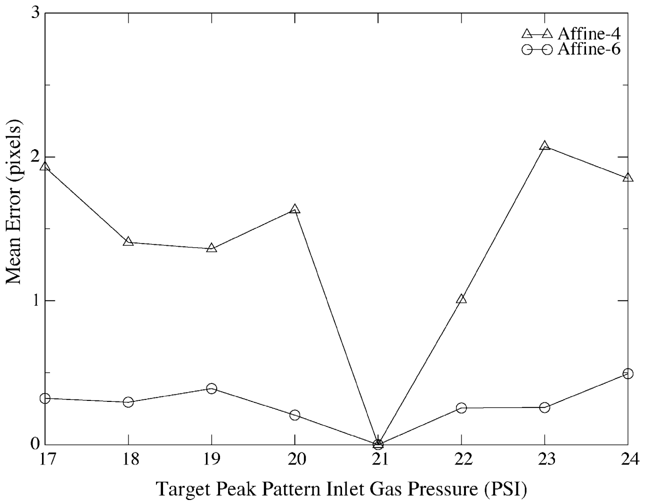

Fig. 7. Residual errors after least-squares optimal transformations of the template from *Gas-pressure* 21 psi for each of the target peak sets.

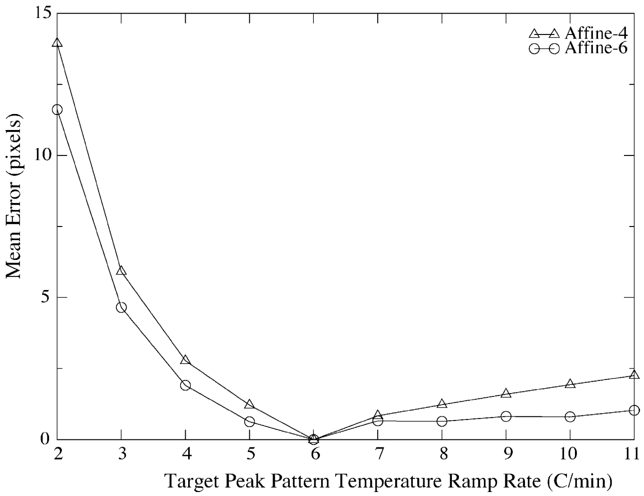

Fig. 8. Residual errors after least-squares optimal transformations of the template from *Oven-temperature* 6 °C/min for each of the target peak sets.

rapid oven temperature ramp rate, but the errors are large for peak patterns acquired at a slower oven ramp rate. These (and the other) results suggest that it is better to have templates that are acquired with similar conditions as the target pattern and that it is better to match a large template to small target (as in Figs. 5 and 6) rather than matching a small template (which has less precision) to a large target. To better remove the non-linear variations generated by large differences in oven temperature ramp rate, a more sophisticated transformation model should be investigated.

5. Conclusion

These results indicate that affine transformations can largely remove peak pattern variations related to acquisition conditions for comprehensive two-dimensional gas chromatography (GC × GC). Therefore, affine transformations can be used as the transformation search space for pattern matching to identify chemical compounds in GC × GC. Although standard retention indices have not yet emerged for GC × GC, these results suggest that when such standards are developed, then local affine transformations may be useful for transforming GC × GC data to generate two-dimensional retention indices, just as piecewise linear transformations are used for generating one-dimensional retention indices.

Acknowledgement

This work was partially supported by the National Science Foundation, Award 0231746.

References

[1] W. Bertsch, J. High Resolut. Chromatogr. 23 (3) (2000) 167.
[2] P.M. Lemieux, J.V. Ryan, Development of a Hazardous Waste Incinerator Target Analyte List of Products of Incomplete Combustion, US Environmental Protection Agency, Washington, DC, 1998.
[3] G.S. Frysinger, R.B. Gaines, E.B. Ledford Jr., J. High Resolut. Chromatogr. 22 (4) (1999) 195.
[4] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, E.B. Ledford Jr., J. Oostijk, H. Trap, Proc. SPIE 5085 (2003) 28.
[5] M.E. Monk, J. Chem. Inf. Comput. Sci. 38 (6) (1998) 997.
[6] US National Inst. of Standards and Technology, MS Search Program, 2002.
[7] W. Welthagen, J. Schnelle-Kreis, R. Zimmermann, J. Chromatogr. A 1019 (2003) 233.
[8] M. Ni, Ph.D. thesis, University of Nebraska-Lincoln, 2004.
[9] M. Ni, Q. Tao, S.E. Reichenbach, in: Proceedings of the IEEE Workshop on Statistical Signal Processing, 2003, pp. 497.