Web-based Supplementary Materials for “A New Method for Peak Detection on Comprehensive Two-dimensional Gas Chromatography Mass Spectrometry Data”

Seongho Kim[1,*], Ming Ouyang[2], Jaesik Jeong[3], Changyu Shen[4], and Xiang Zhang[5,*]

[1] Biostatistics Core, Karmanos Cancer Institute, Wayne State University, Detroit, MI.
[2] Department of Computer Science, University of Massachusetts Boston, Boston, MA.
[3] Department of Statistics, Chonnam National University, Gwangju, Korea.
[4] Department of Biostatistics, Indiana University, Indianapolis, IN.
[5] Department of Chemistry, University of Louisville, Louisville, KY.

Co-corresponding authors’ email: kimse@karmanos.org and xiang.zhang@louisville.edu

Contents

A brief introduction to GC×G- TOF MS data

The comprehensive two dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOF MS) instrument consists of two GC and one TOF MS. As shown in the schematic representation of GC×GC-TOF MS instrument (Figure E1), the first column and the second column correspond to the first GC and the second GC instruments, respectively. Each GC is to separate the injected molecules (compounds) as the sample travels through the column, and the length of traveling time is called the retention time. The two columns separate various species of molecules such that the fraction coming out of the columns is dominated by one species of compound. This fraction is then injected into the TOF MS instrument and then broken into smaller ion fragments, whose abundance and mass to charge ratio (m/z) were captured through a “scan” by the mass spectrometer. The output of the experiment using GC×GC-TOF MS instrument looks like Figure E2 (a). The first two columns are the first dimension and the second

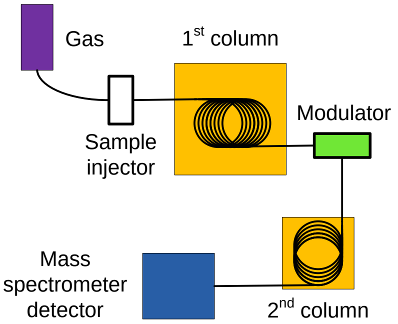
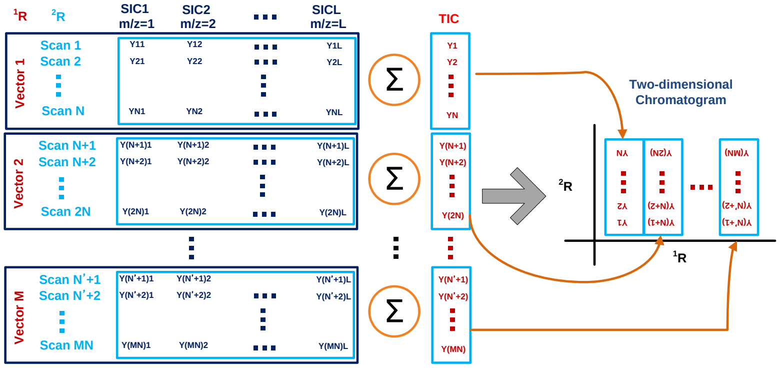


Figure E1. The schematic representation of GC×GC-TOF MS instrument.



(a) Output data from an experiment of GC×GC TOF MS

(b) TIC in a vector form

(c) TIC in a matrix form

Figure E2. The graphical representation of constructing two dimensional TIC chromatograms. (a) Output data from an experiment of GC×GC-TOF MS. (b) TIC in a vector form. (c) TIC in a matrix form.

dimension retention times, respectively. The rest of the columns are the observed intensities of the ionized fragments at each m/z value. Here the intensity measure is called SIC or XIC (single ion chromatogram), and, by marginalizing SICs with respect to m/z values, TIC (total ion chromatogram) will be obtained. In this stage, the TIC is formed into a vector as shown in Figure E2 (b). Then the vector TIC will be separated into several sub-vector TICs according to their different first dimension retention times. By putting these sub-vector TICs into the two dimensional plot, the two dimensional chromatogram will be created as depicted in Figure E2 (c).

A two dimensional chromatogram from a real GC×GC-TOF MS experiment is depicted in Figure E3. In an ideal case, each peak represents a pure molecule (compound), which will be identified by peak detection.

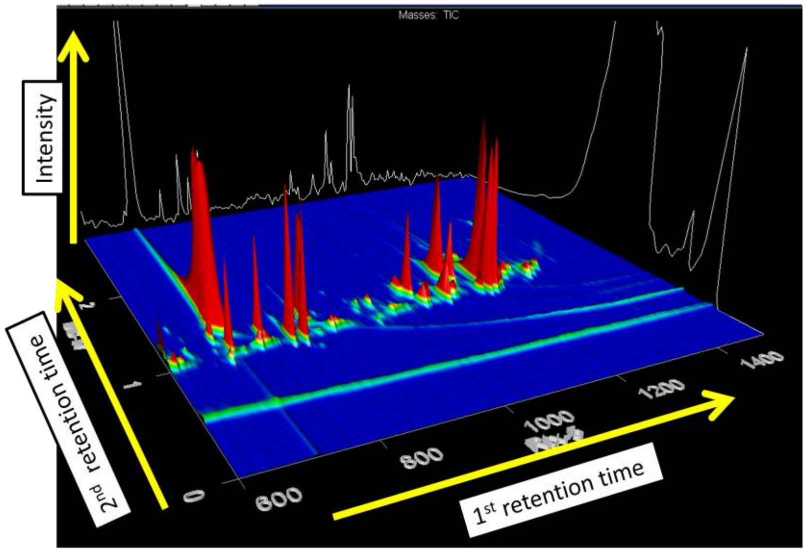


Figure E3. An example of the two dimensional chromatogram.

The current peak detection is performed based on the following three procedures: denoising, baseline removal, and peak picking. The denoising and the baseline removal will be performed in the stage of constructing either XIC (Figure E2 (a)) or TIC (Figure E2 (b)). And then the peak picking will be performed based on either the vector-formed TIC in Figure E2 (b) or the matrix-formed TIC in Figure E2 (c).

Derivation of Equations (3) and (4)

The true TICs of some proportion *r* are present (i.e., $\Theta_i \neq 0$), while the others remain at zero ($\Theta_i = 0$). For positions where the true TIC is present, we use the following model:

$$X_i \sim N(\Theta_i + \mu, \sigma^2) \text{ and } \Theta_i \sim Exp(\phi),$$

where X_i is an observed TIC at the *i*th position, Θ_i is the true TIC of X_i of the exponential distribution with ϕ, and μ is the mean background or baseline with variance σ^2. In case that no TIC is present, the background signal follows:

$$X_i \sim N(\mu, \sigma^2).$$

When $\Theta_i \neq 0$, the marginal density of x_{i} is driven by

$$p_1(x_i) = p(x_i | \Theta_i \neq 0)$$

$$= \int_0^\infty p(x_i, \theta_i) d\theta_i = \int_0^\infty p(x_i | \theta_i) f(\theta_i) d\theta_i$$

$$= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \theta_i - \mu)^2\right) \frac{1}{\phi} exp\left(-\frac{\theta_i}{\phi}\right) d\theta_i$$

$$= \frac{1}{\phi} exp\left(\frac{\sigma^2}{2\phi^2} - \frac{x_i - \mu}{\phi}\right) \cdot \Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right),$$

where

$$p(x_i|\theta_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \theta_i - \mu)^2\right) \text{ since } X_i | \Theta_i \neq 0 \sim N(\theta_i + \mu, \sigma^2);$$

$$p(\theta_i) = \frac{1}{\phi} exp\left(-\frac{\theta_i}{\phi}\right) \text{ since } \Theta_i \sim Exp(\phi).$$

When $\Theta_i = 0$, the marginal density of x_i is the probability density function (pdf) of a normal distribution with mean μ and variance σ^2 since $x_i | \Theta_i = 0 \sim N(\mu, \sigma^2)$:

$$p_0(x_i) = p(x_i | \Theta = 0) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right).$$

■

Derivation of Equation (12)

Once the significant TICs (true signals) are detected by the posterior odds, baseline correction and denoising are performed simultaneously based on the estimated parameters. To do this, we assume that the observed TIC is a true signal with $\Theta_\mathrm{i} \neq 0$. That is, the observed TIC follows the following model:

$$X_i \sim ND(\Theta_i + \mu, \sigma^2) \text{ and } \Theta_i \sim Exp(\phi),$$

where *ND* stands for a normal distribution, X_i is an observed TIC at the *i*th position, Θ_i is the true TIC of X_i of the exponential distribution with ϕ, and μ is the mean background or baseline with variance σ^2. Then the convoluted TIC $\hat{x}_i$ is predicted by the expected true TIC Θ_i given the observed TIC $\mathrm{x_i}$, i.e.,

$$\hat{x}_i = E(\Theta_i | x_i) = \int_0^\infty \theta_i p(\theta_i | x_i) d\theta_i,$$

where $\mathrm{p(\theta_i | x_i)}$ is the conditional density of Θ_i given the observed TIC $\mathrm{x_i}$. Moreover, by Bayes' law,

$$\mathrm{p(\theta_i | x_i)} = \frac{\mathrm{p(x_i | \theta_i) p(\theta_i)}}{\mathrm{p(x_i)}}.$$

Since $X_i | \mathit{\Theta}_i \neq 0 \sim ND(\theta_i + \mu, \sigma^2)$ and $\mathit{\Theta}_i \sim Exp(\phi)$,

$$p(x_i | \theta_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \theta_i - \mu)^2\right);$$

$$p(\theta_i) = \frac{1}{\phi} exp\left(-\frac{\theta_i}{\phi}\right).$$

Consequently,

$$p(\theta_i | x_i) = \frac{\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \theta_i - \mu)^2\right) \frac{1}{\phi} exp\left(-\frac{\theta_i}{\phi}\right)}{\frac{1}{\phi} exp\left(\frac{\sigma^2}{2\phi^2} - \frac{x_i - \mu}{\phi}\right) \cdot \Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right)}$$

$$= \frac{\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(x_i - \theta_i - \mu)^2 - \frac{\theta_i}{\phi} - \frac{\sigma^2}{2\phi^2} + \frac{x_i - \mu}{\phi}\right)}{\Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right)}$$

$$= \frac{\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}\left(\theta_i - x_i + \mu + \frac{\sigma^2}{\phi}\right)^2\right)}{\Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right)}$$

$$= \frac{\frac{1}{\sigma}\varphi\left(\frac{\theta_i - x_i + \mu + \frac{\sigma^2}{\phi}}{\sigma}\right)}{\Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right)}, \tag{S1}$$

where φ and Φ are the probability density and cumulative distribution functions of the standard normal distribution $ND(0,1)$, respectively

Therefore, by inserting Equation (S1),

$$E(\Theta_i|x_i) = \int_0^\infty \theta_i p(\theta_i|x_i) d\theta_i$$

$$= \int_0^\infty \theta_i \frac{\frac{1}{\sigma}\varphi\left(\frac{\theta_i - x_i + \mu + \frac{\sigma^2}{\phi}}{\sigma}\right)}{\Phi\left(\frac{x_i - \mu - \frac{\sigma^2}{\phi}}{\sigma}\right)} d\theta_i. \tag{S2}$$

In Equation (S2), we can observe that Θ_i follows a truncated normal distribution $TN\left(x_i - \mu - \frac{\sigma^2}{\phi}, \sigma^2\right)$ within the interval $\Theta_i \in (0, \infty)$. In addition, if a random variable X follows a truncated normal distribution within the interval $X \in (0, \infty)$:

$$X \sim TN(m, v^2), X \in (0, \infty),$$

its expectation is $E(X) = m - \frac{\varphi\left(\frac{m}{v}\right)}{\Phi\left(\frac{m}{v}\right)} v$. Therefore, the expected true TIC given the observed TIC, $\mathrm{E}(\Theta_\mathrm{i}|\mathrm{x}_\mathrm{i})$, becomes

$$E(\Theta_i|x_i) = x_i - \left(\mu + \frac{\sigma^2}{\phi}\right) + \sigma \cdot \frac{\varphi\left(\frac{x_i - \left(\mu + \frac{\sigma^2}{\phi}\right)}{\sigma}\right)}{\Phi\left(\frac{x_i - \left(\mu + \frac{\sigma^2}{\phi}\right)}{\sigma}\right)}.$$

In turn, the convoluted TIC, $\hat{x}_i$, is predicted based on the estimated parameters by the following equation:

$$\hat{x}_i = x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right) + \hat{\sigma} \cdot \frac{\varphi\left(\frac{x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)}{\hat{\sigma}}\right)}{\Phi\left(\frac{x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)}{\hat{\sigma}}\right)},$$

where φ and Φ are the probability density and cumulative distribution functions of the standard normal distribution $ND(0,1)$, respectively.

■

PDFs of five probability models

The pdf and its parameter for Poisson, truncated Gaussian, Gaussian, Gamma, and exponentially modified Gaussian are, respectively, as follows*:*

$$f(t|\xi = \lambda) = \frac{\lambda^t}{t!} \cdot e^{-\lambda}, \lambda > 0; \qquad (S3)$$

$$f\left(t\middle|\xi = (\mu, \sigma, a, b)\right) = \frac{\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \sigma > 0; \qquad (\mathrm{S4})$$

$$f\left(t\middle|\xi = (\mu, \sigma)\right) = \frac{1}{\sqrt{2\pi}\sigma} \cdot Exp\left(-\frac{1}{2\sigma^2} \cdot (t-\mu)^2\right), \sigma > 0; \qquad (\mathrm{S5})$$

$$f\left(t\middle|\xi = (\alpha, \beta)\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, \alpha > 0, \beta > 0; \qquad (\mathrm{S6})$$

$$f\big(t\big|\xi = (\mu,\sigma,\lambda)\big) = \frac{\lambda}{2}\cdot Exp\left(\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2t)\right)\cdot erfc\left(\frac{\mu+\lambda\sigma^2-t}{\sqrt{(2)\sigma}}\right), \sigma > 0, \lambda > 0 \quad \text{(S7)}$$

where φ and Φ are the probability density and cumulative distribution functions of the standard normal distribution $ND(0,1)$, respectively; $erfc(t) = \frac{2}{\sqrt{\pi}}\int_t^\infty e^{-x^2}dx$.

■

Figure legends

Figure S1. The flowchart of the proposed peak finding procedure for analysis of GC×GC-TOF MS data.

Figure S2. The contour plots of GC×GC-TOF MS data obtained from an experiment and its convoluted total ion chromatogram (TIC) using NEB models. (a)The entire GC×GC-TOF MS data. (b)The sub-region considered in this study which is magnified for the inlet green box in (a). (c) is the plot of the original and convoluted TICs of the sub-region as depicted in (b). (d) is the magnified TICs of the green box in (c).

Figure S3. The nonzero peak region indices for each cut-off value of odds. (a) When the cut-off value is 1. (b) When the cut-off value is 10. (c) When the cut-off value is 100. The grey line represents the contour plot of GC×GC-MS data, and the red box represents the nonzero peak region detected.

Figure S4. The detected nonzero peak regions and peaks before/after peak merging using the sub-region. The plots are depicted when the cut-off value of odds is 1 ((a),(b)), 10 ((c),(d)), and 100 ((e),(f)), respectively. The plots before merging are in (a), (c), and (e). The plots (b), (d), and (f) are after merging.

Figure S5. The contour plots of the entire GC×GC-MS data used for peak detection and its convoluted total ion chromatogram (TIC) using NEB models. (a) The entire GC×GC-MS data with the inlet used for peak detection. (b) The magnified region of the inlet green box in (a). (c) is the plot of the original and convoluted TICs of the entire data (b). (d) is the magnified TICs of the green box in (c).

Figure S6. The analysis of MS similarity within the selected peak region. (a) The black solid line with the solid circle represents the observed normalized intensities. The blue, red, and green dotted lines indicate the MS similarities with the points around each of the three local maxima from the left, respectively. (b) The heatmap of the correlation matrix of MS similarities among the observed data points. The white and the dark green indicate the highest and the lowest MS similarities, respectively.

Figure S7. The contour plots of the entire GC×GC-MS data for another replicated data set. The entire GC×GC-MS data with the inlet used for peak detection in the upper, and the bottom is the magnified region of the inlet green box in the upper.

Figure S8. The detected nonzero peak regions and peaks by trial-and-error optimization before/after peak merging using another replicated data set. The plots are depicted when the optimization is performed with MSE before peak merging (a) and after peak merging (b).

Figure S1. The flowchart of the proposed peak finding procedure for analysis of GC×GC-TOF MS data.

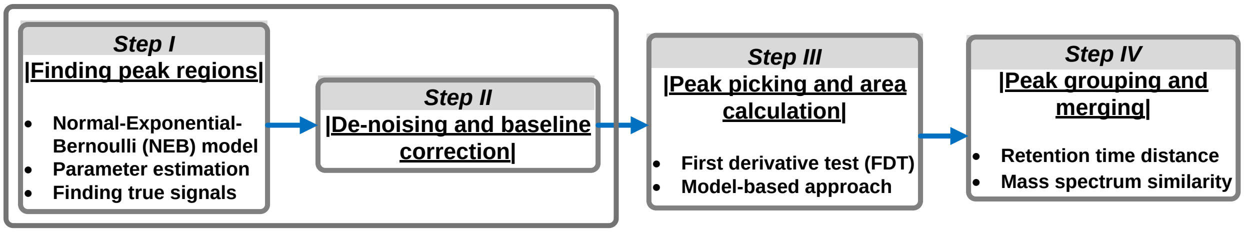

Step I
|Finding peak regions|
Normal-Exponential-Bernoulli (NEB) model
Parameter estimation
Finding true signals
Step II
|De-noising and baseline correction|
Step III
|Peak picking and area calculation|
First derivative test (FDT)
Model-based approach
Step IV
|Peak grouping and merging|
Retention time distance
Mass spectrum similarity

Figure S2. The contour plots of GC×GC-TOF MS data obtained from an experiment and its convoluted total ion chromatogram (TIC) using NEB models. (a)The entire GC×GC-TOF MS data. (b)The sub-region considered in this study which is magnified for the inlet green box in (a). (c) is the plot of the original and convoluted TICs of the sub-region as depicted in (b). (d) is the magnified TICs of the green box in (c).

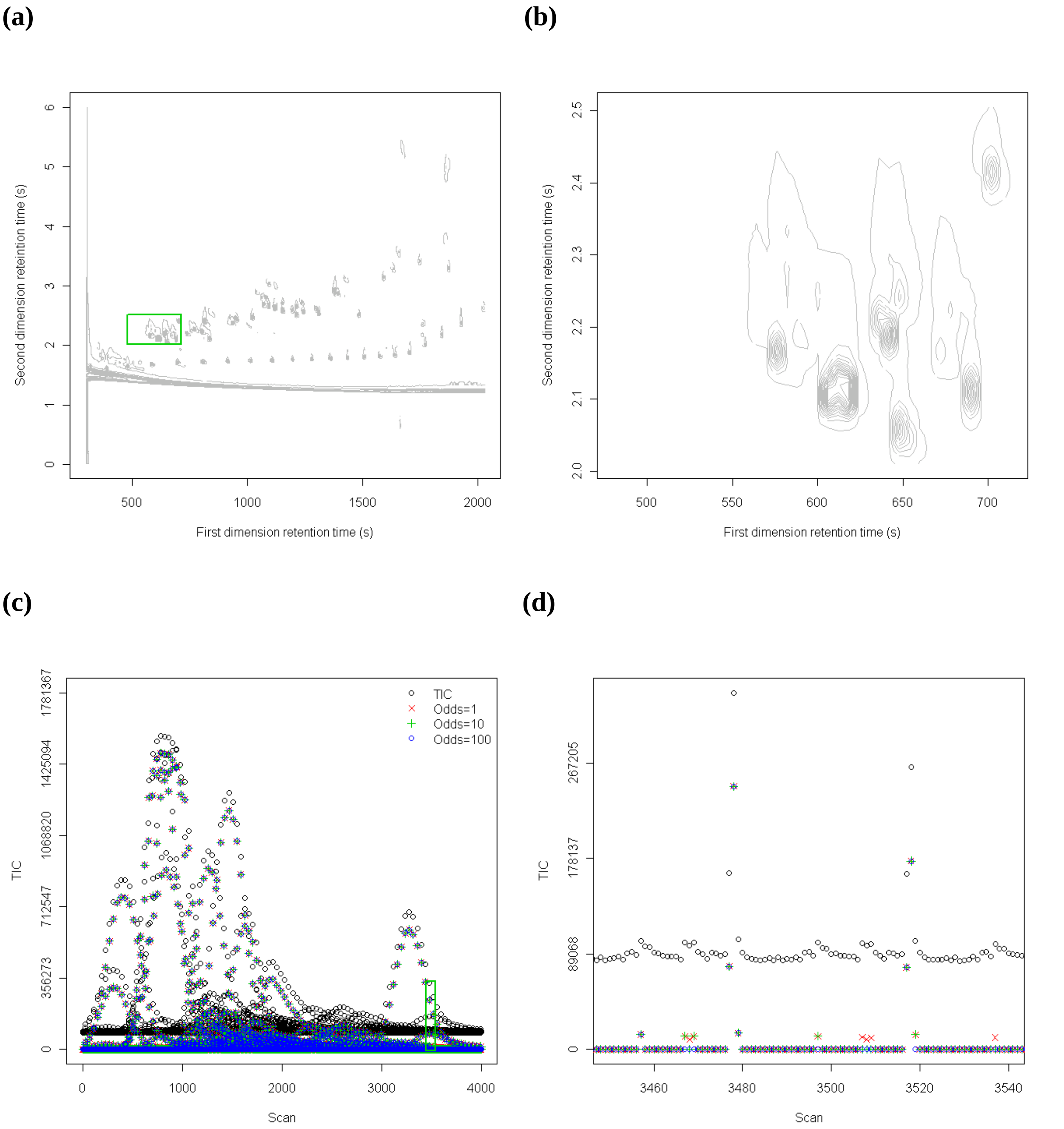


Figure S3. The nonzero peak region indices for each cut-off value of odds. (a) When the cut-off value is 1. (b) When the cut-off value is 10. (c) When the cut-off value is 100. The grey line represents the contour plot of GC×GC-MS data, and the red box represents the nonzero peak region detected.

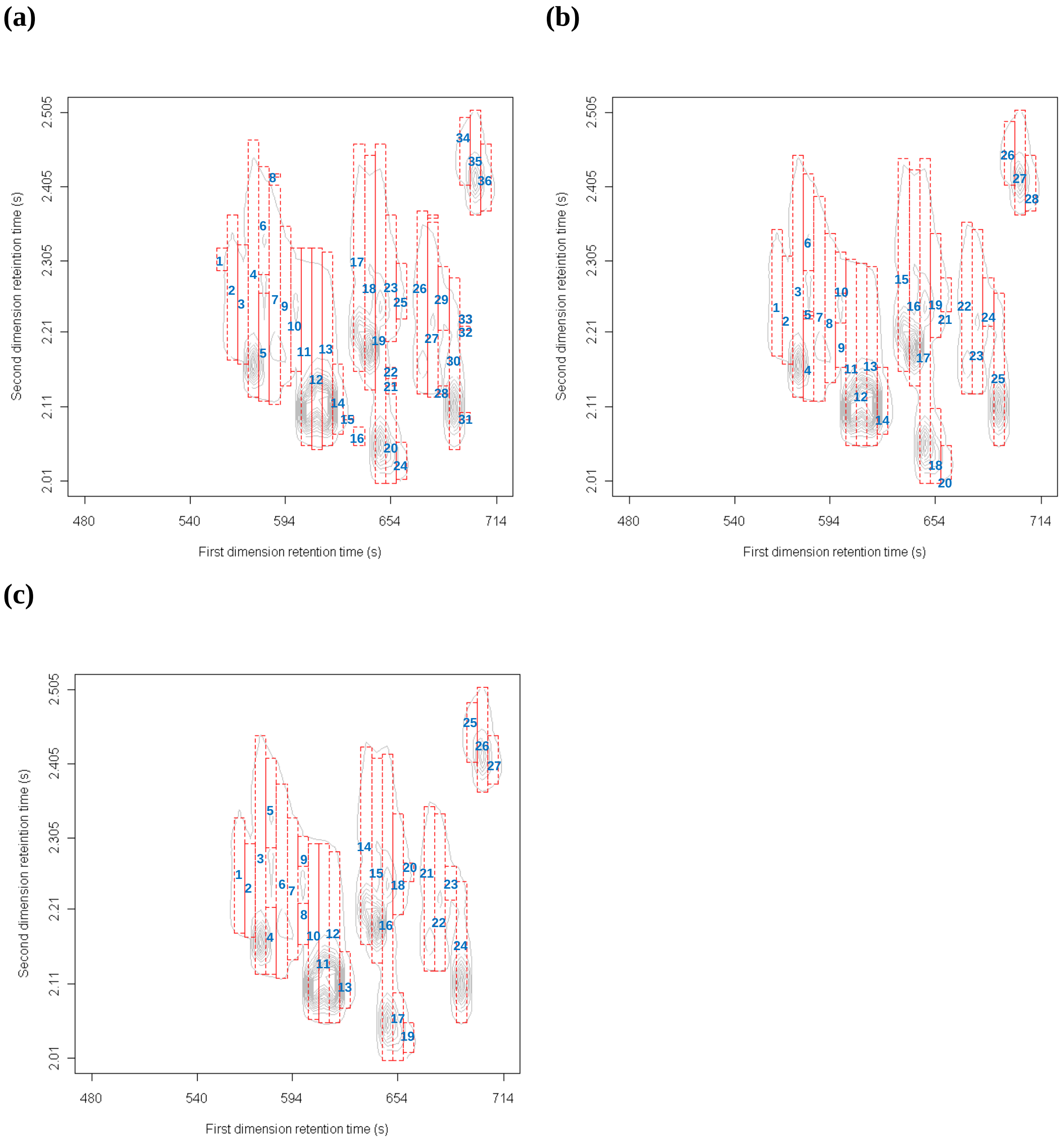
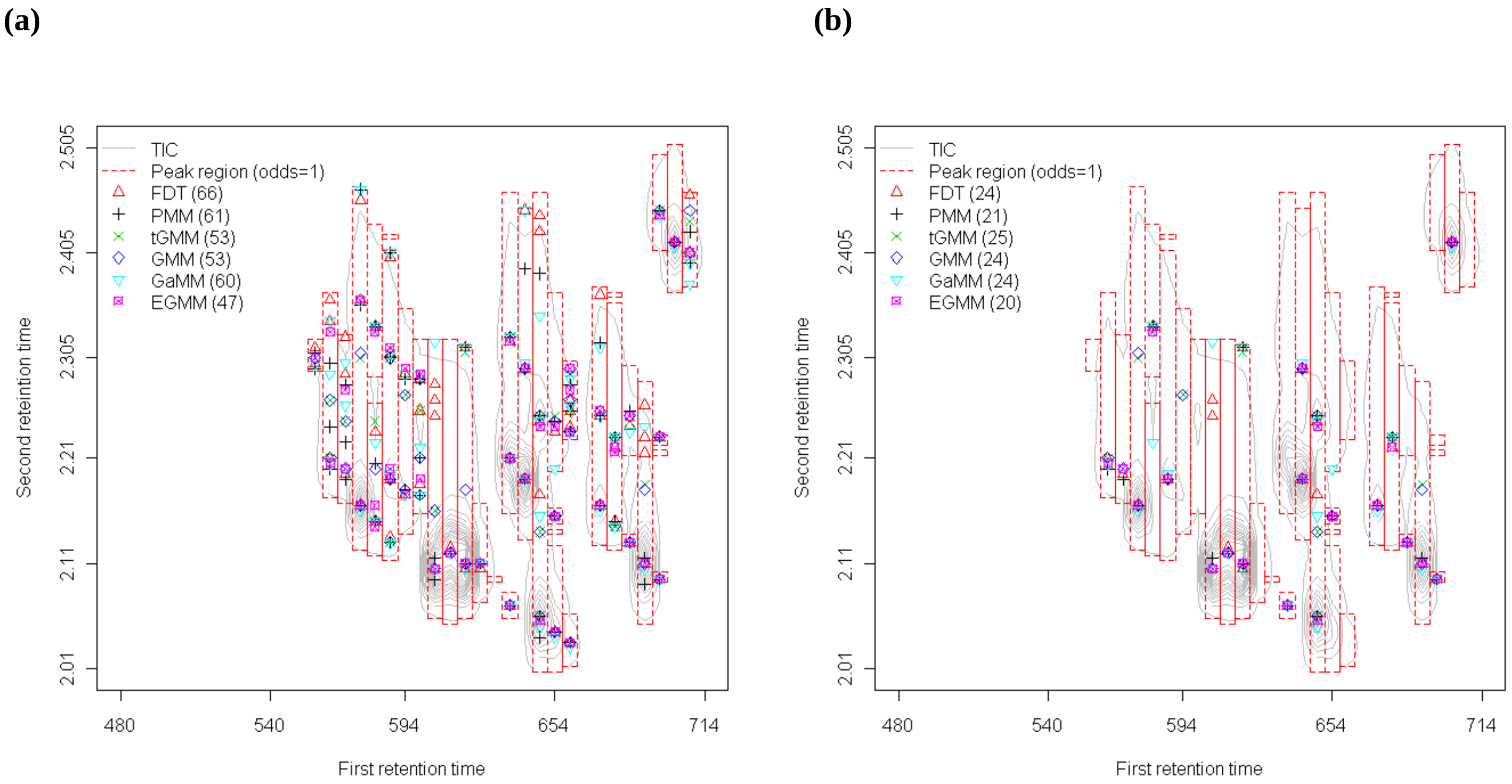
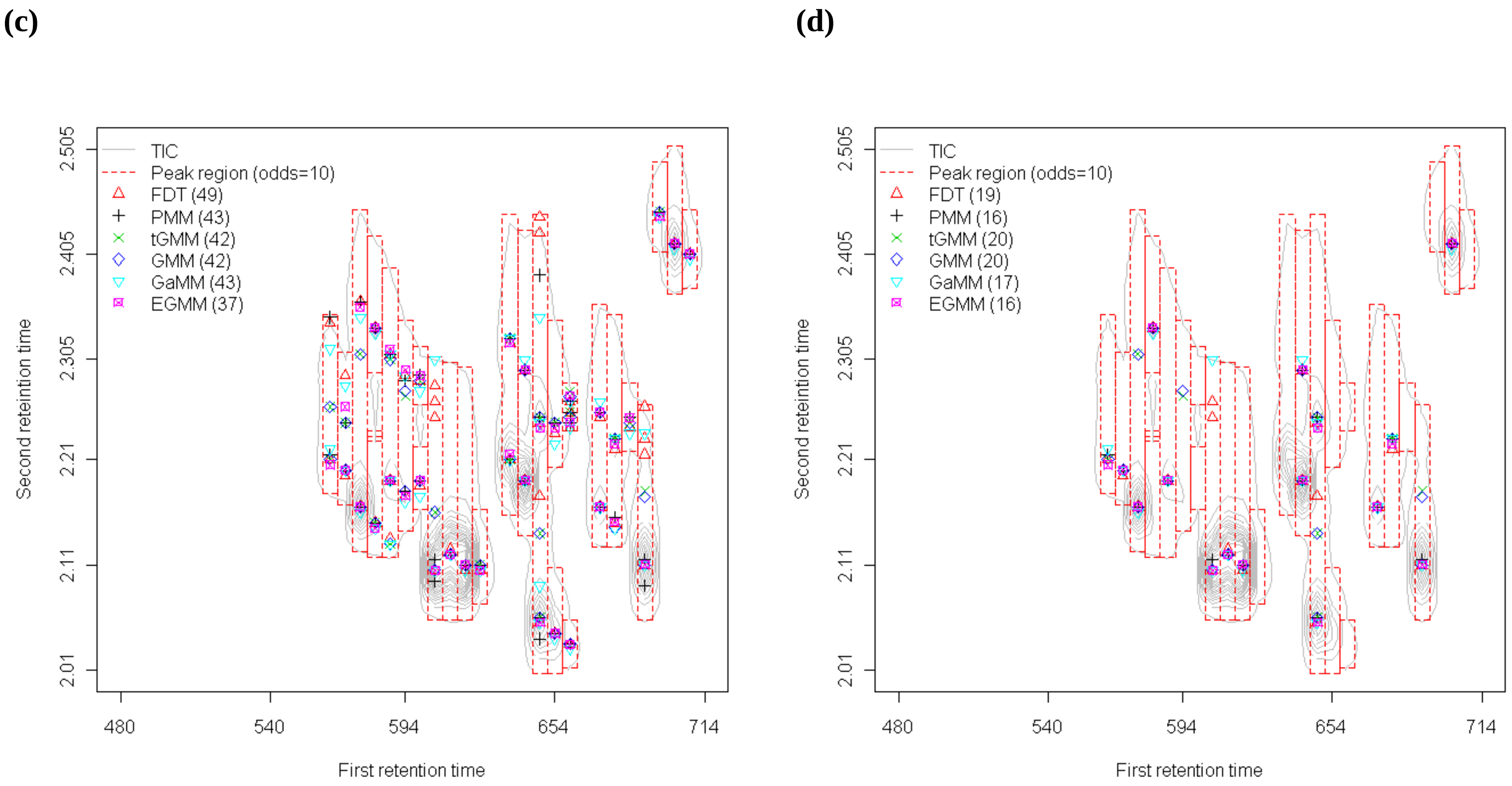


Figure S4. The detected nonzero peak regions and peaks before/after peak merging using the sub-region. The plots are depicted when the cut-off value of odds is 1 ((a),(b)), 10 ((c),(d)), and 100 ((e),(f)), respectively. The plots before merging are in (a), (c), and (e). The plots (b), (d), and (f) are after merging.




(c)
TIC
Peak region (odds=10)
FDT (49)
PMM (43)
tGMM (42)
GMM (42)
GaMM (43)
EGMM (37)
Second reteintion time
2.505
2.405
2.305
2.21
2.11
2.01
480
540
594
654
714
First retention time
(d)
TIC
Peak region (odds=10)
FDT (19)
PMM (16)
tGMM (20)
GMM (20)
GaMM (17)
EGMM (16)
Second reteintion time
2.505
2.405
2.305
2.21
2.11
2.01
480
540
594
654
714
First retention time

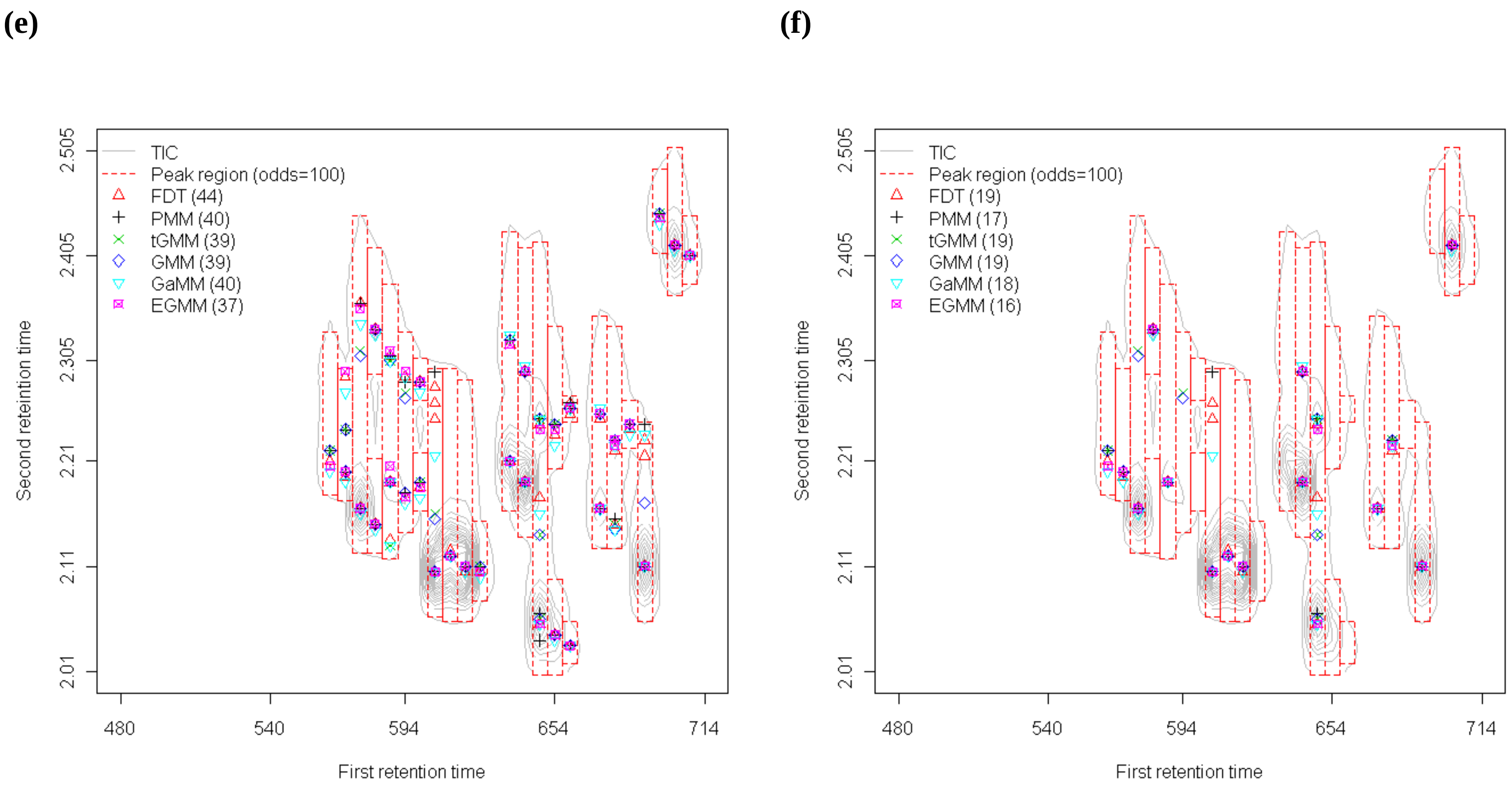

(e)
TIC
Peak region (odds=100)
FDT (44)
PMM (40)
tGMM (39)
GMM (39)
GaMM (40)
EGMM (37)
Second reteintion time
2.01
2.11
2.21
2.305
2.405
2.505
480
540
594
654
714
First retention time
(f)
TIC
Peak region (odds=100)
FDT (19)
PMM (17)
tGMM (19)
GMM (19)
GaMM (18)
EGMM (16)
Second reteintion time
2.01
2.11
2.21
2.305
2.405
2.505
480
540
594
654
714
First retention time

Figure S5. The contour plots of the entire GC×GC-MS data used for peak detection and its convoluted total ion chromatogram (TIC) using NEB models. (a) The entire GC×GC-MS data with the inlet used for peak detection. (b) The magnified region of the inlet green box in (a). (c) is the plot of the original and convoluted TICs of the entire data (b). (d) is the magnified TICs of the green box in (c).

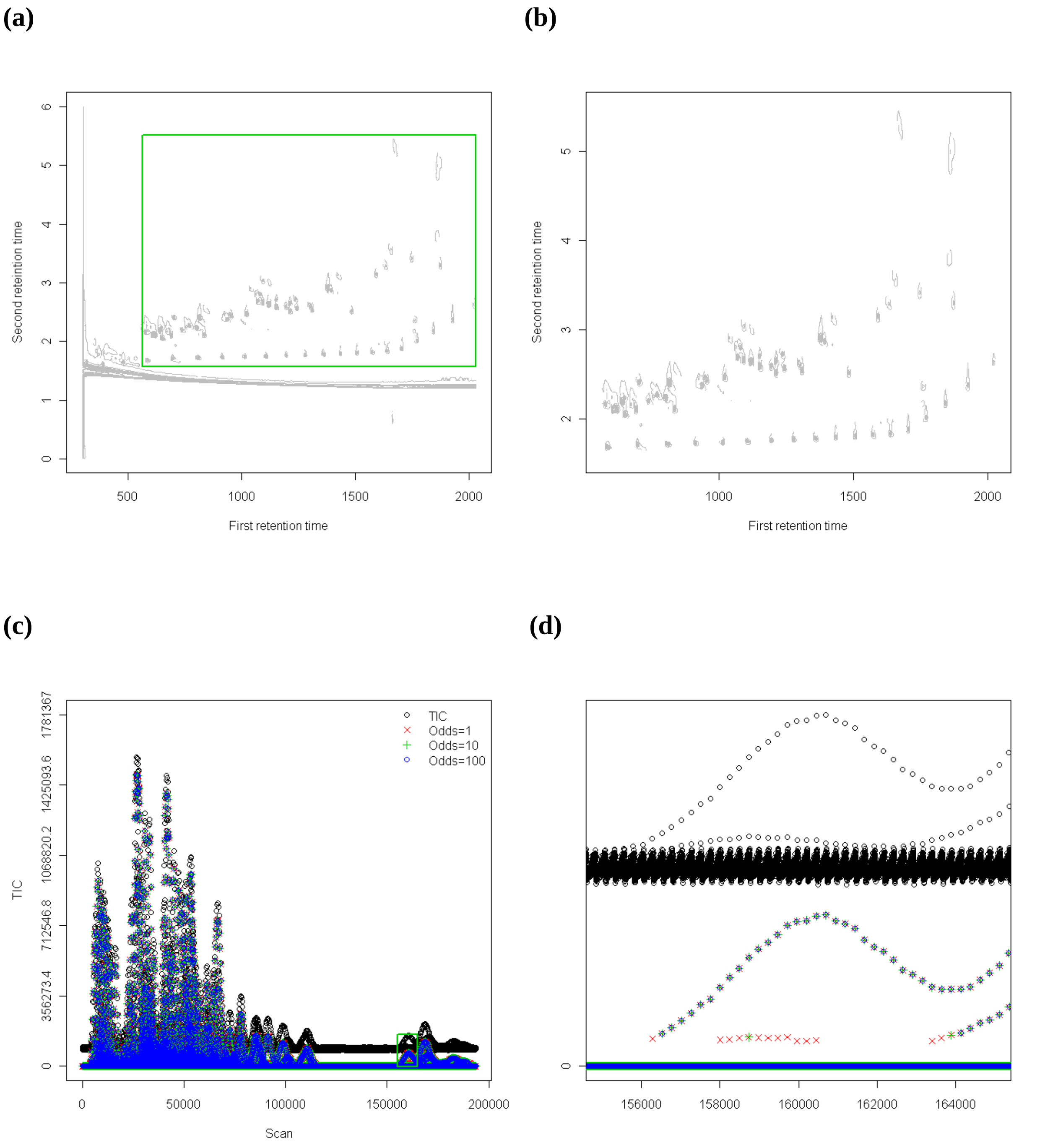


Figure S6. The analysis of MS similarity within the selected peak region. (a) The black solid line with the solid circle represents the observed normalized intensities. The blue, red, and green dotted lines indicate the MS similarities with the points around each of the three local maxima from the left, respectively. (b) The heatmap of the correlation matrix of MS similarities among the observed data points. The white and the dark green indicate the highest and the lowest MS similarities, respectively.

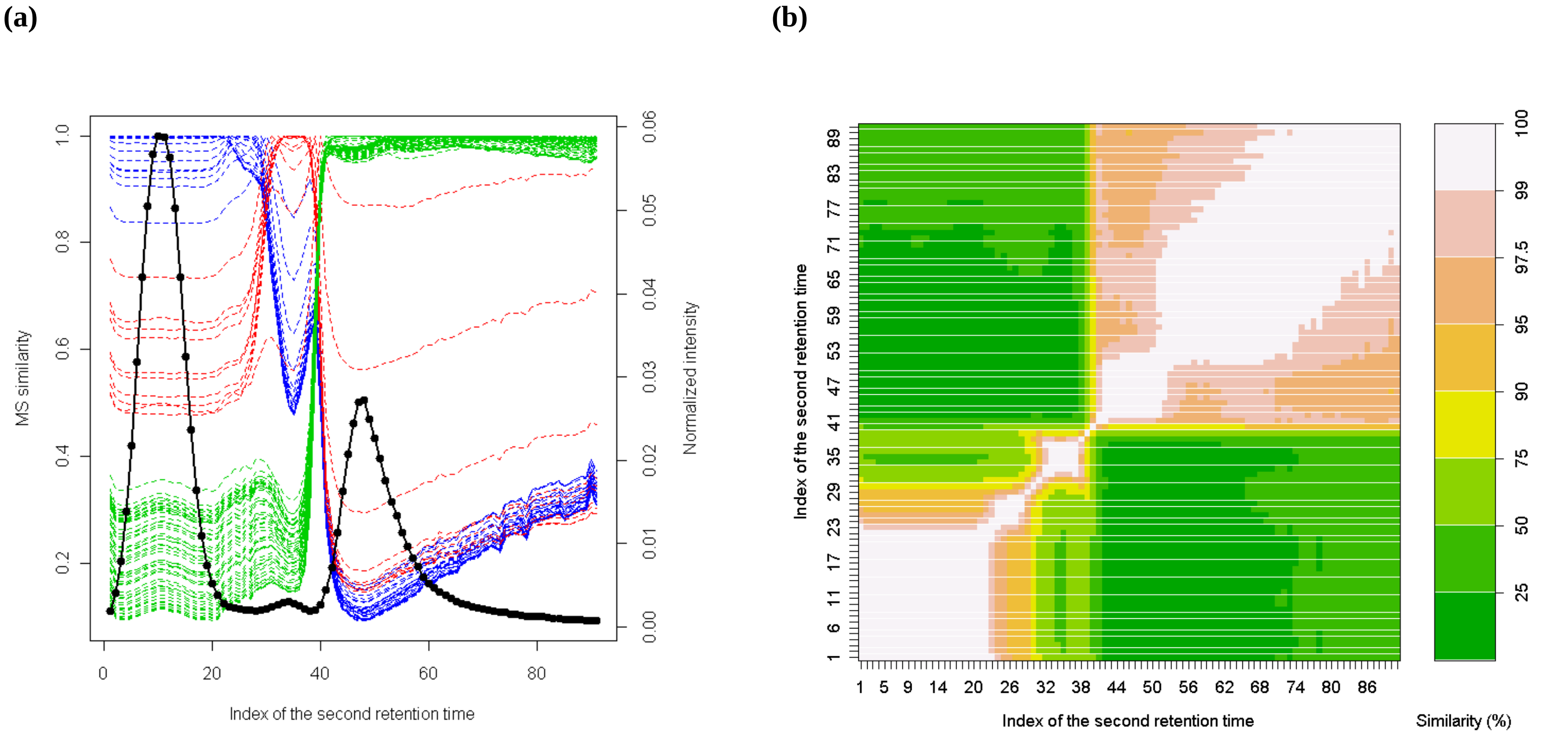


Figure S7. The contour plots of the entire GC×GC-MS data for another replicated data set. The entire GC×GC-MS data with the inlet used for peak detection in the upper, and the bottom is the magnified region of the inlet green box in the upper.

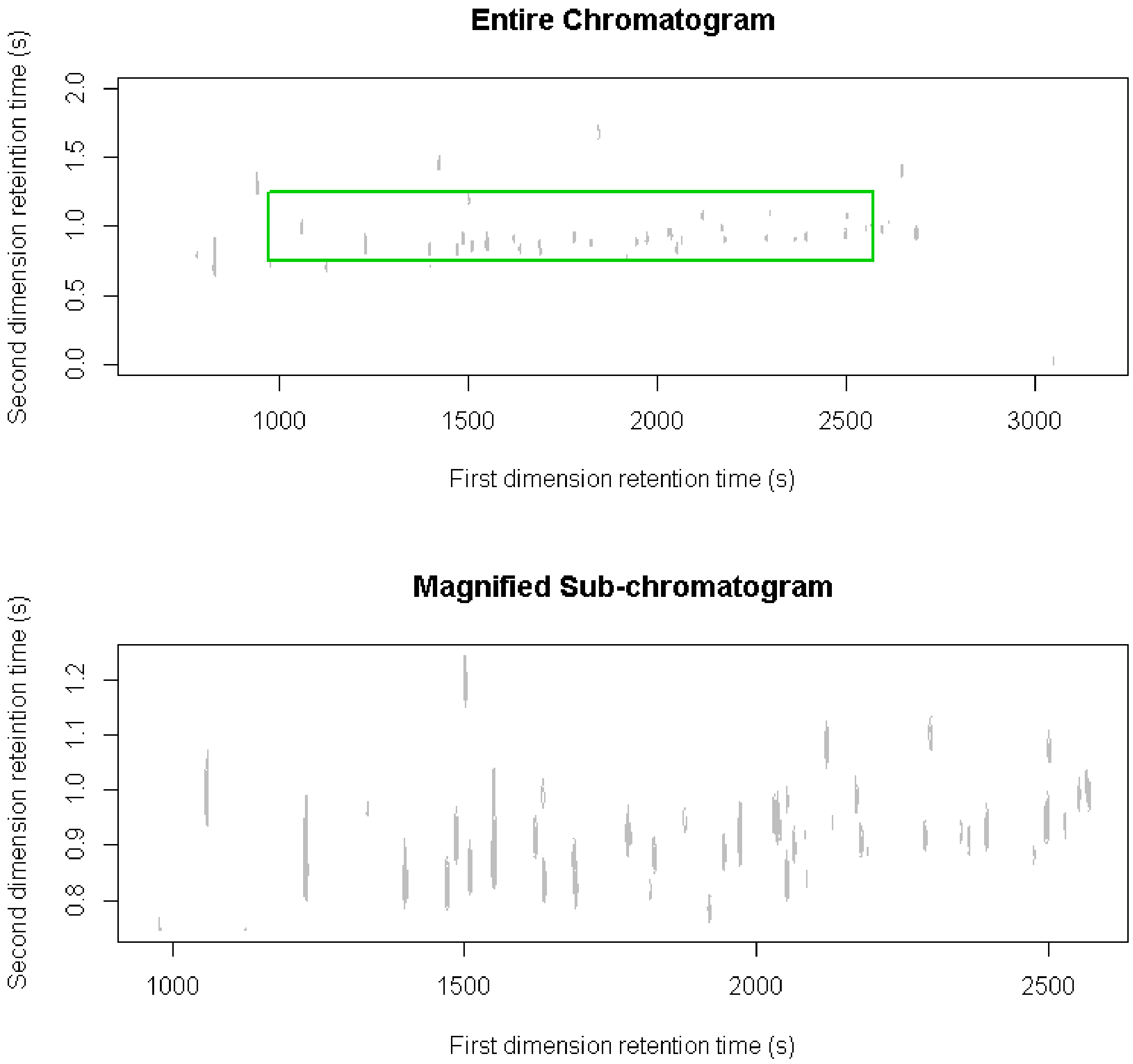
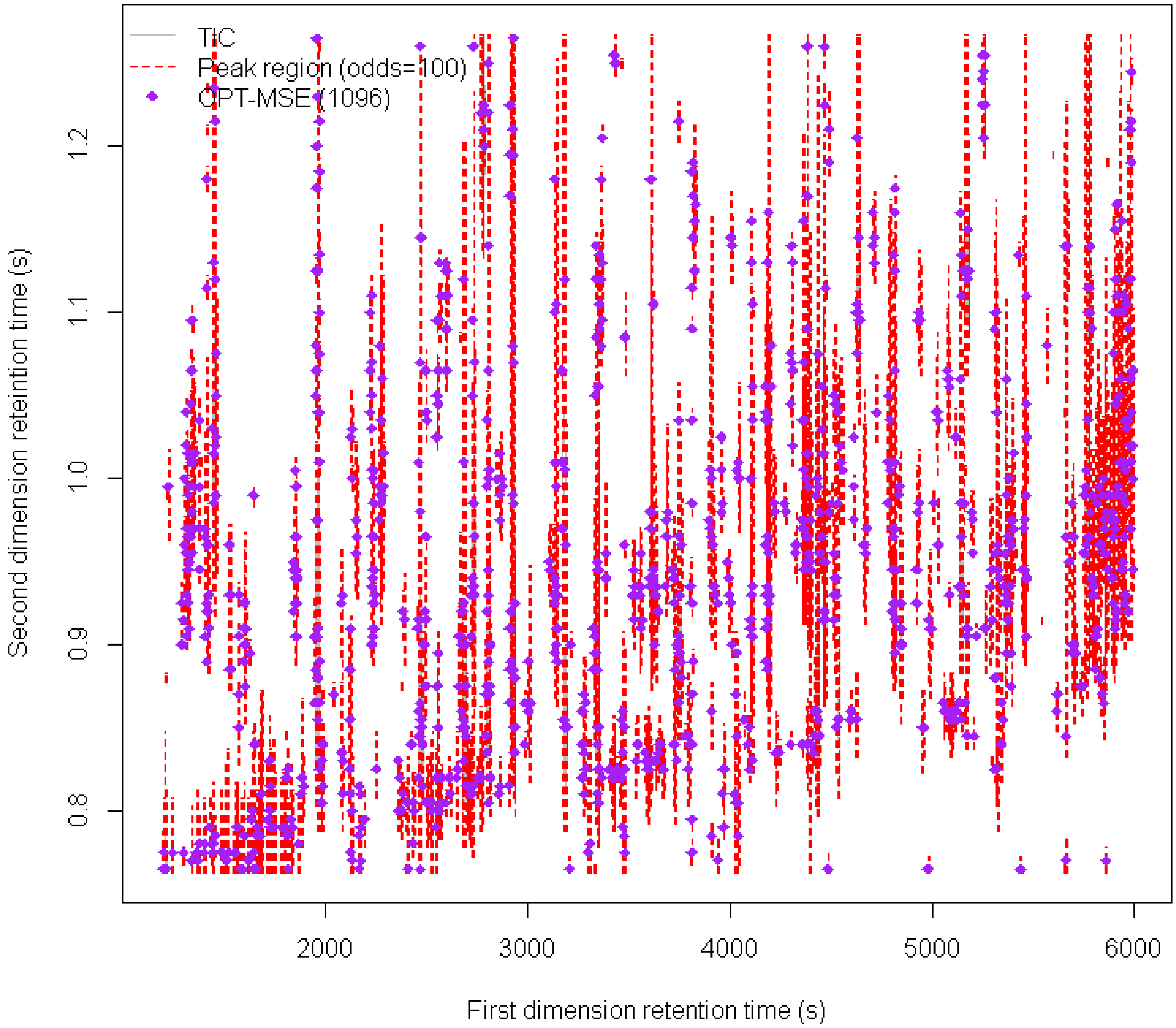
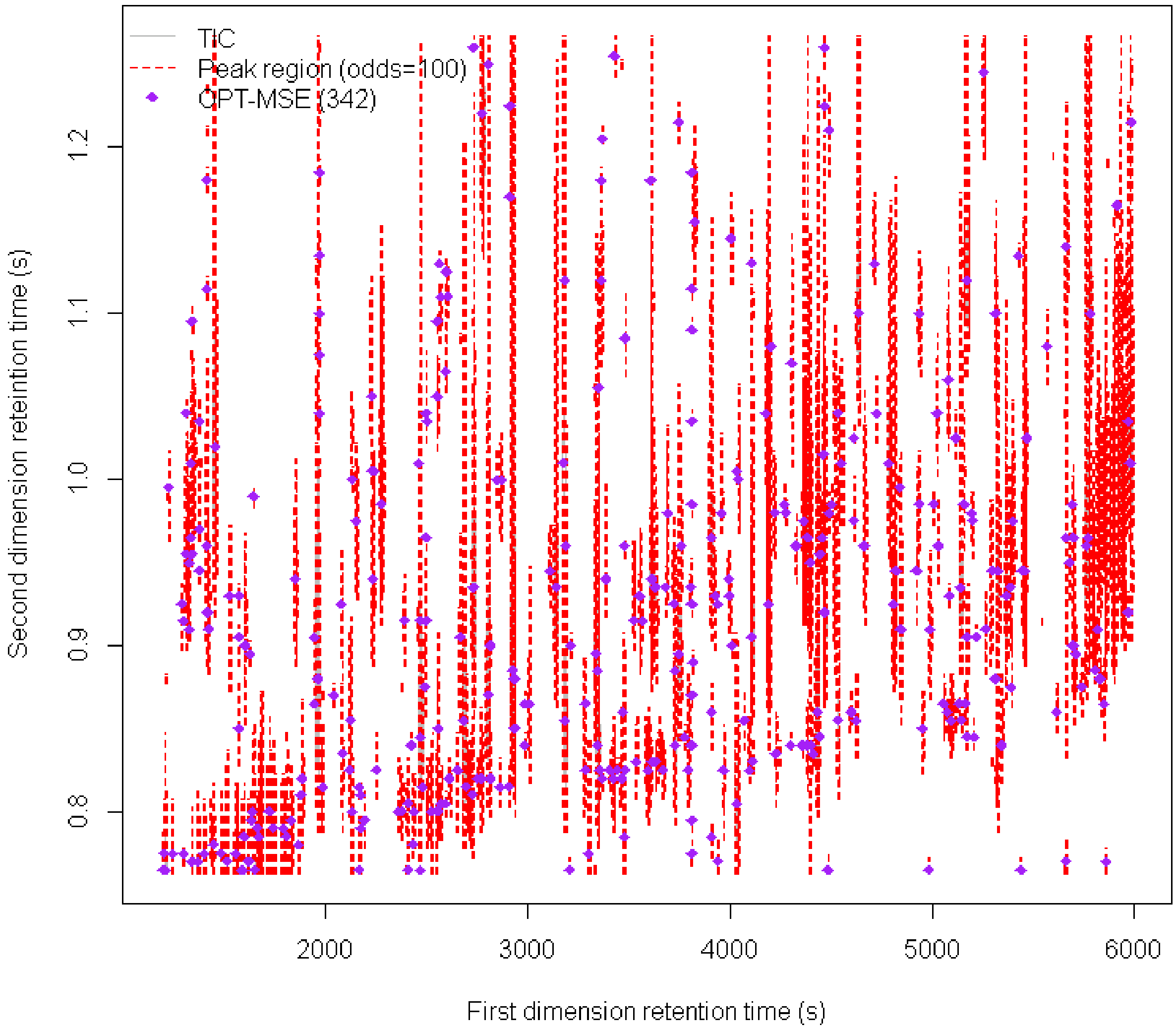


Figure S8. The detected nonzero peak regions and peaks by trial-and-error optimization before/after peak merging using another replicated data set. The plots are depicted when the optimization is performed with MSE before peak merging (a) and after peak merging (b).

(a)



(b)


TIC
Peak region (odds=100)
OPT-MSE (342)
Second dimension reteintion time (s)
0.8
0.9
1.0
1.1
1.2
2000
3000
4000
5000
6000
First dimension retention time (s)