Mixture Similarity Peak Alignment for Two-Dimensional Gas Chromatography Mass Spectrometry Based Metabolomics

Seongho Kim[1*], Aiqin Fang[2], Bing Wang[2], Jaesik Jeong[3], Xiang Zhang[2*]

[1] Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40292, USA; [2] Department of Chemistry, University of Louisville, Louisville, KY 40292, USA; [3] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA

GC×GC-MS data

Two sets of GC×GC-MS data were used in this study. One is a mixture of compound standards and the other is a spiked-in sample. In the first dataset (Dataset I), a mixture of 76 compounds (8270 MegaMix, Restek Corp., Bellefonte, PA) and C7-C40 saturated alkanes (Sigma-Aldrich Corp., St. Louis, MO) spiked with a deuteronated six component semi-volatiles internal standard (ISTDF) mixture (Restek Corp., Bellefonte, PA) at a concentration of 2.5 μg/mL were analyzed on a LECO Pegasus 4D GC×GC-MS instrument (LECO Corporation, St. Joseph, MI, USA) equipped with a cryogenic modulator. The GC×GC-MS analyses were repeated 10 times under 5 ^{o}C/min temperature gradient, resulting in a total of 10 datasets.

As for the spiked-in sample (Dataset II), a 100 μL rat plasma sample was mixed with 900 μL of organic solvent mixture (methanol/water 8:1, v/v) and vortexed for 15 s. After sitting at 20 °C for 30 min, the mixture was centrifuged with 16000 ×g at 4 °C for 15 min. Supernatants from the mixture were collected and evaporated to dryness with a SpeedVac and then redissolved in 100 μL of pyridine. Fifty micro liters of the metabolite extract was treated with 100 μL of 50 mg/mL ethoxyamine hydrochloride pyridine solution for 30 min at 60 °C. Subsequently, the extracts were derivatized with 100 μL of MTBSTFA for 1 h at 60 °C. The derivatized sample was spiked with ISTD mixture at a concentration of 2.5 μg/mL right before the GC×GC-MS analysis. Then the compounds were analyzed five times on GC×GC-MS.

In order to find the tentative true alignment peaks, we did compound identification with ChromaTOF software using its two NIST databases, mainlib and replib. After that, we chose the compounds which had the similarity score of the ChromaTOF greater than or equal to 600 for both Dataset I and II. In addition, we also set to the S/N of 100 for Dataset I and of 200 for Dataset II.

Table S1 summarizes each dataset by calculating the number of compounds and the absolute difference between minimum and maximum of each retention time (|*max-min*|). The numbers in parentheses are the original number of peaks before correcting the multiple peaks. To compare the variation between two retention times, the coefficient of variation (CV) of $|max - min|$, which is the ratio of the standard deviation to mean, was estimated. We observed that the second dimension retention time has more variation than the first dimension retention time (the first retention time's CV = 0.0103 and 0.0023; the second retention time's CV = 0.1056 and 0.1097) for both datasets. The scatter plots of Dataset I and II of the first and second dimension retention times are depicted in Figure S2. It should be noted that the identified compounds by ChromaTOF could be wrong. In other words, all the compound names identified are "tentative."

Effect of different distance measures on peak matching

The graphs that the distance from a point *A* is equal to 3 with the four distance measures (0.33 for Canberra distance) are delineated in Figure S1 to help understand the variation among the distance measures.

In case of matching to a point *A* from points *C* and *D*, the point *D* is closer to the point *A* in Canberra distance while the point *C* is closer to the point *A* in other distance measures. Because of the GC×GC-MS instrument configuration, the first dimension retention time is much larger than the second dimension retention time. The first dimension retention time usually ranges from 0 to one hour, depending on multiple experiment parameters such as temperature gradient. The second dimension retention time is defined by the instrument modulation time, which is usually in a range of several seconds, for example, 5 second. Using Canberra distance measure, one may be able to take account of this difference since it normalizes the distance in each retention time by dividing the sum of the retention times, as described in Equation (4) of the main paper.

When the closest peak from the point *A* is searched from points *B* and *E*, the matched point would be the point *E* in case of the Maximum distance and the point *B* for other distance measures because $D_1(A,B) < D_1(A,E)$; $D_2(A,E) < D_2(A,B)$; $D_3(A,B) < D_3(A,E)$; $D_4(A,B) < D_4(A,E)$, where D_i is a certain distance measure described in the main paper (Equations (1) to (4)).

Review of MSort and DISCO peak alignment algorithms

Currently, two algorithms, MSort (Oh et al., 2008) and DISCO (Bing et al., 2010), are available for peak alignments based on peak list of homogeneous two-dimensional gas chromatography mass spectrometry data. Both methods use the peak distance as well as the spectral similarity between two peaks. Two methods are briefly described in the following.

MSort

In MSort, for each peak t_j in the target peak list, r_i is aligned from the current reference chromatogram such that:

$$r_i = argmax_{r_h \in R_S} corr(I_{t_j}, \mathrm{I}_{\mathrm{r_h}} | r_h \in R_S)$$

and

$$R_s = \left\{ r_h \middle| |t_{j,1} - r_{h,1}| \le \delta_1, |t_{j,2} - r_{h,2}| \le \delta_2, corr\left(I_{t_j}, I_{r_h}\right) \ge \rho_{min},\ r_h \in R \right\}, \text{(A1)}$$

where $corr\left(I_{t_j}, I_{r_h}\right)$ is the Pearson's correlation coefficient of the spectra of two peaks t_j and r_h for the spectra similarity, δ_1 and δ_2 are the threshold of the peak distance for the first and the second retention times, respectively, and ρ_{min} is the threshold of the similarity measure. MSort uses the correlation information for the last decision rule when $|R_S| = m_S \ge 2$.

The distance measure used is in fact the same as the maximum distance which is D_2 (Equation (2) in the main paper), although it is not clarified in the paper. Thus the expression (A1) can be reformulated as

$$R_s = \left\{ r_k \middle| D_2(t_j, r_h) \le \delta, corr\left(I_{t_j}, I_{r_h}\right) \ge \rho_{min},\ r_h \in R \right\}, \text{(A2)}$$

where δ is the threshold of the peak distance and $m_S = |R_S| \le |R| = m$.

DISCO

Bing et al. (2010) introduced a peak alignment method entitled DISCO for both of the homogeneous and heterogeneous two-dimensional gas chromatograms. We here focus on the only homogeneous case of DISCO since the heterogeneous alignment is beyond the scope of this work. DISCO finds first the landmark peaks among all the chromatograms using the Euclidean distance and the Pearson's correlation based similarity measure. The landmark peaks found are used to reduce the search space of the non-landmark peaks. In detail, the two-dimensional domains of the reference and the target chromatograms are divided into several rectangles according to the first and the second dimension retention times of the landmark peaks that are present in both of the reference and the target chromatograms.

Then DISCO searches the non-landmark peaks that have the Pearson's correlation coefficient greater than a predefined cutoff value in the rectangle. If there is no match satisfied with the cutoff value, the search space is extended into the adjacent rectangles to find the aligned peaks. In DISCO, the Euclidean distance is used as the final decision rule while MSort uses the correlation as the final criteria. Overall, the aligned peak $r_i \in$ R of the peak $t_j \in$ T is found in DISCO using the rules below:

$$r_i = argmin_{r_h \in R_S} D_1(t_j, r_h),$$

where $R_S = \left\{ r_h \middle| corr\left(I_{t_j}, \mathrm{I_{r_h}}\right) \geq \rho_{min}, r_h \in X \right\}$, $\mathrm{D_1}$ is the Euclidean distance (Equation (1) in the main paper), $corr(I_{t_j}, I_{r_h})$ is the Pearson's correlation coefficient of the spectra, and $m_S = |R_S| \leq |R| = m$.

References

1. Oh, C., Huang, X., Regnier, F. E., Buck, C., and Zhang, X. **Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm**. *Journal of Chrmatography A*, 2008, 1179, 205-215.
2. Wang, B., Fang, A., Heim, J., Bogdanov, B., Pugh, S., Libardoni, M., and Zhang, X. **DISCO: distance and spectrum correlation optimization alignment for two dimensional gas chromatography time-of-fligh mass spectrometry-based metabolomics**. *Analytical Chemistry*, 2010, 82, 5069-81.

Table S1. The summary of GC X GC/TOF-MS datasets. (a) A total of 10 datasets were generated under the temperature gradients of 5 °C/min for the mixture of 76 compound standards. (b) A total of 5 datasets were generated for a spiked-in sample. The number of compounds and the absolute difference between the minimum and the maximum of each retention time are calculated.

	A mixture of compound standards (Dataset I)										**A spiked-in sample (Dataset II)**				
RUN ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**
The number of compounds	78 (180)*	76 (186)	76 (161)	75 (151)	74 (151)	73 (145)	74 (172)	76 (163)	77 (168)	75 (174)	466 (759)	456 (733)	436 (694)	452 (727)	418 (661)
|Max-Min|**	2948	2888	2948	2948	2923	2863	2928	2948	2948	2948	1974	1974	1964	1969	1974
|Max-Min|***	3.089	3.135	3.050	3.029	3.115	2.370	3.036	2.508	3.458	3.055	4.158	3.412	4.099	3.406	3.359

*, the number of peaks found by ChromaTOF before choosing a peak out of multiple peaks; **, the absolute difference between the minimum and the maximum of the first retention times; ***, the absolute difference between the minimum and the maximum of the second retention times

Table S2. The maximum F1 scores of each peak alignment method of pairwise peak alignments for Dataset I and II. The mean and standard error (SE) of TPR, PPV, and F1 are reported along with the 95% confidence intervals for PPV and F1. The constant k is the cutoff value of the distance-based window for DW-PAS, ρ is the cutoff value of the similarity-based window for SW-PAD, w is the weight of the mixture similarity for PAM.

Dataset	Method	k	ρ (w)	Distance	TPR		PPV		F1		95% CI of PPV		95% CI of F1	
					Mean	SE	Mean	SE	Mean	SE	lower	upper	lower	Upper
I	PAD	3	0.5	Canberra	0.9507	0.0035	0.9803	0.0022	0.9652	0.0026	0.9759	0.9846	0.9600	0.9704
	PAS				0.8772	0.0077	0.9292	0.0063	0.9023	0.0069	0.9168	0.9415	0.8888	0.9157
	DW-PAS			Euclidean	0.9535	0.0045	0.9704	0.0032	0.9618	0.0037	0.9642	0.9767	0.9546	0.9690
	SW-PAD			Canberra	0.9720	0.0030	0.9814	0.0024	0.9766	0.0026	0.9767	0.9860	0.9716	0.9817
	PAM			Canberra	0.9751	0.0024	0.9870	0.0021	0.9810	0.0020	0.9830	0.9911	0.9771	0.9849
II	PAD	15	0.93	Manhattan	0.5289	0.0215	0.4278	0.0151	0.4729	0.0177	0.3982	0.4574	0.4382	0.5076
	PAS				0.6912	0.0113	0.5474	0.0098	0.6109	0.0102	0.5282	0.5667	0.5908	0.6310
	DW-PAS			Canberra	0.6785	0.0122	0.5271	0.0104	0.5932	0.0110	0.5068	0.5474	0.5716	0.6149
	SW-PAD			Manhattan	0.5973	0.0142	0.5936	0.0134	0.5954	0.0136	0.5673	0.6199	0.5687	0.6221
	PAM		0.05	Euclidean Maximum Manhattan	0.7012	0.0110	0.5475	0.0101	0.6148	0.0104	0.5278	0.5673	0.5945	0.6351

Table S3. Estimates of the parameters $\theta = (d, w)$ of each peak pair of Dataset I and II using the OP-PAM method. Dataset I has 9 pairs of 10 peak lists and four pairs of five peak lists are existed in Dataset II. The estimates $\hat{\theta} = (\hat{d}, \hat{w})$ and the likelihood function L(T, R) are presented for each peak pair of both Dataset I and II. The starting value of w was 0.5 for all the cases.

Dataset	Indices of align pair		Estimates ($\hat{\theta}$)		Likelihood ($L(T, R)$)
	Reference (R)	Target (T)	Distance ($\hat{d}$)	$\hat{w}$	
I	**1**	**2**	Canberra	0.9656	219.64
	2	**3**	Canberra	0.6914	213.93
	3	**4**	Canberra	0.9510	218.47
	4	**5**	Canberra	0.5115	209.50
	5	**6**	Canberra	0.8682	202.08
	6	**7**	Canberra	0.9023	207.31
	7	**8**	Canberra	0.1954	209.14
	8	**9**	Canberra	0.6085	215.19
	9	**10**	Canberra	0.5508	219.38
II	**1**	**2**	Canberra	0.5515	931.99
	2	**3**	Canberra	0.7119	932.33
	3	**4**	Canberra	0.5558	930.44
	4	**5**	Canberra	0.6557	900.11

Figure S1. Graphical representation of different distance measures. The graphs whose distance from a point *A* is equal to 3 are depicted with Euclidean, Maximum and Manhattan distance measures. In case of Canberra distance, the distance is 0.33. The square, circle, rhombus, and quadrilateral-like shapes represent the Maximum, Euclidean, Manhattan, and Canberra distance measures, respectively.

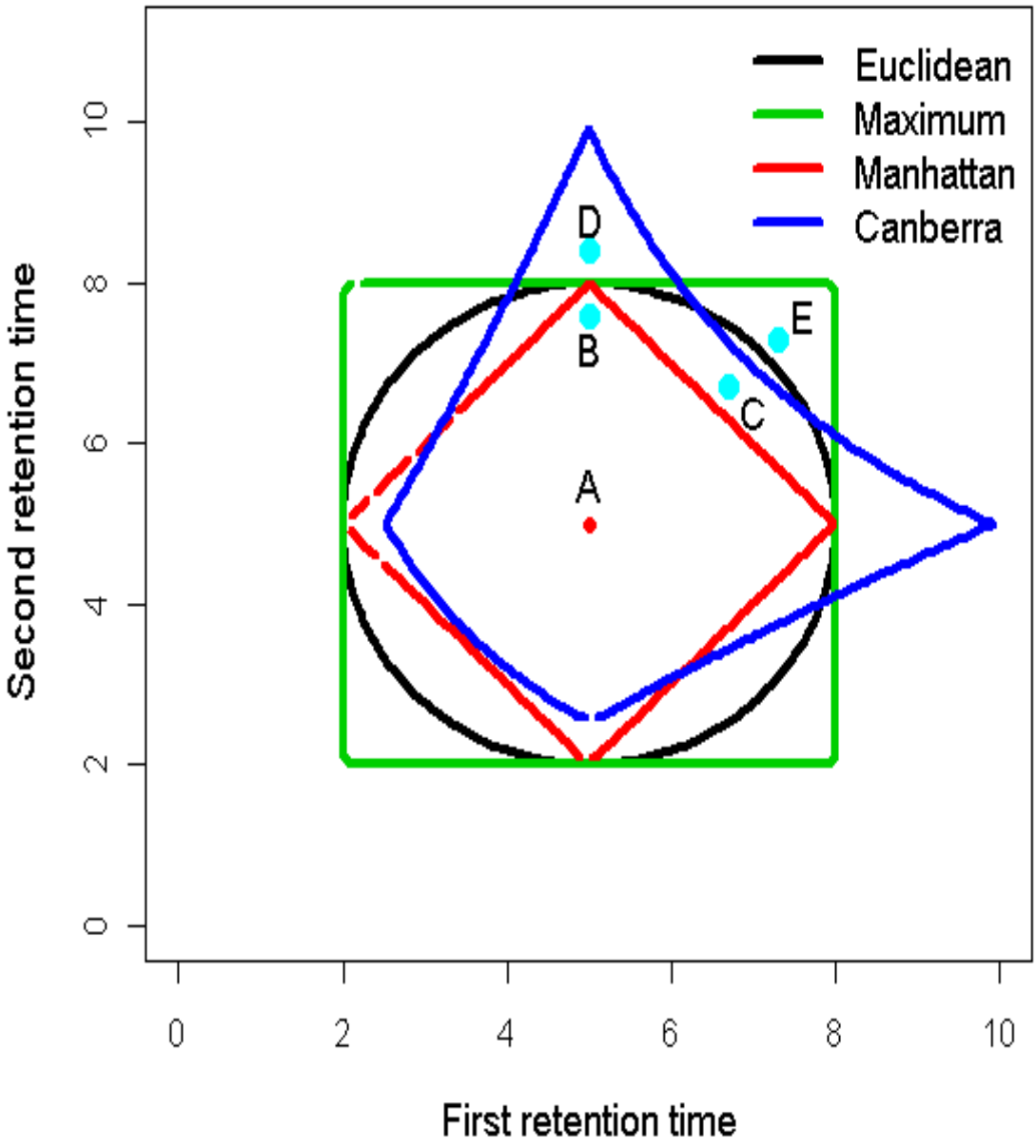
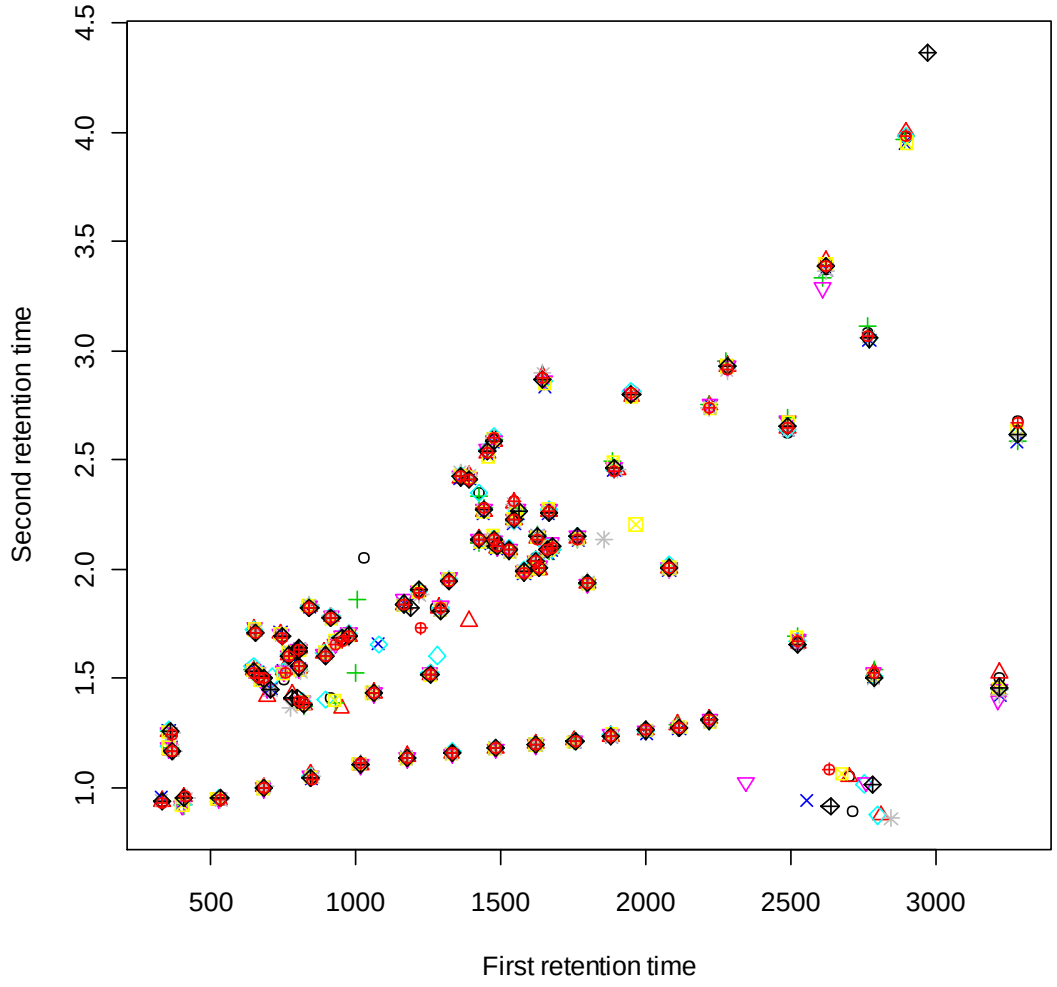


Figure S2. The scatter plots of the first and the second retention times of the peak list of each dataset. The scatter plots for two sets of data are depicted in (a) for Dataset I and (b) for Dataset II. Dataset I and II have 10 and 5 GCxGC chromatograms, respectively.

(a)



(b)

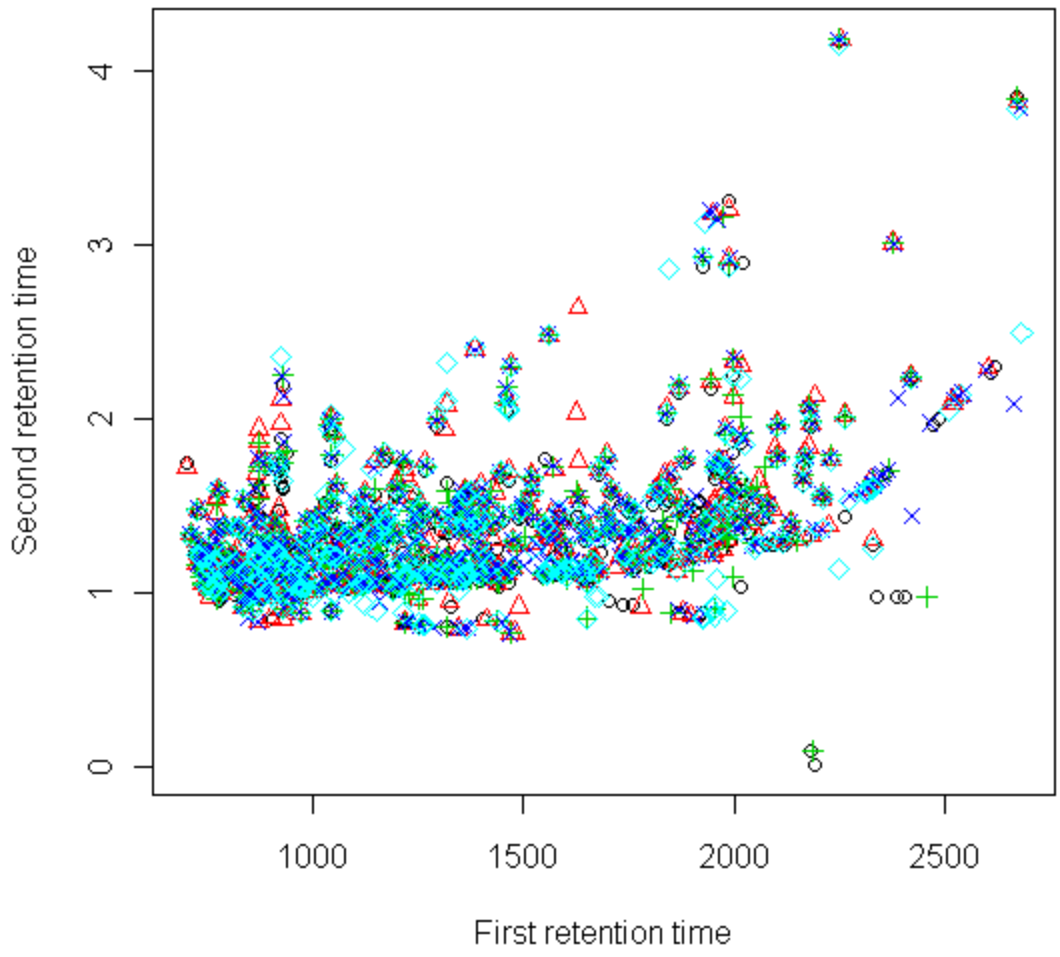
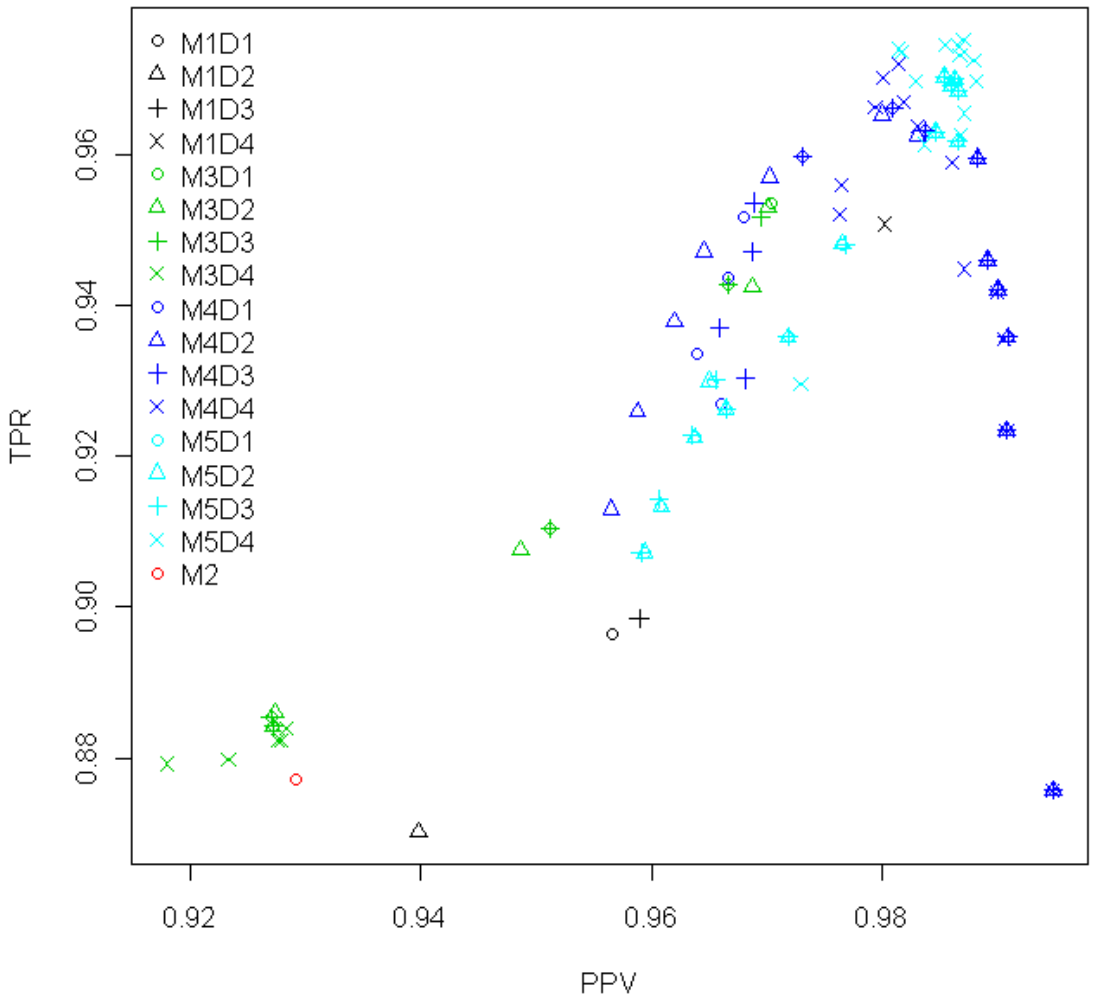

Second retention time
0
1
2
3
4
1000
1500
2000
2500
First retention time

Figure S3. The overall performance of the pairwise peak alignment for Dataset I and II. The scatter plots of the true positive rate (TPR) versus the positive predictive value (PPV) are depicted for (a) Dataset I and (b) Dataset II. The black, red, green, blue, and cyan colored points are of PAD (M1), PAS (M2), SW-PAD (M3), DW-PAS (M4), and PAM (M5), respectively. The Euclidean (D1), Maximum (D2), Manhattan (D3), and Canberra (D4) distances are depicted as circles, triangles, pluses ("+"), and crosses ("x"), respectively.

(a)



(b)

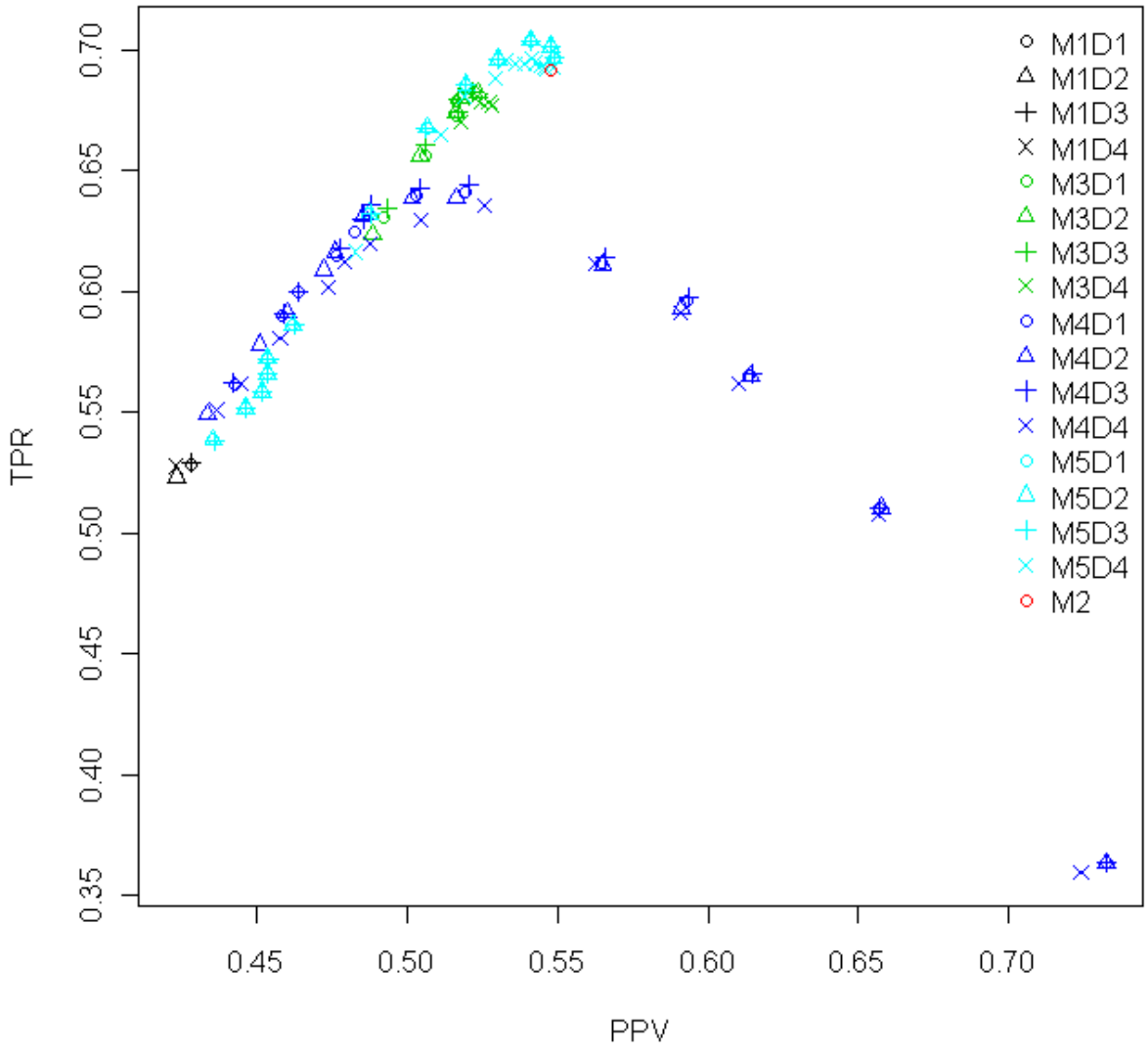

TPR
PPV
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
M1D1
M1D2
M1D3
M1D4
M3D1
M3D2
M3D3
M3D4
M4D1
M4D2
M4D3
M4D4
M5D1
M5D2
M5D3
M5D4
M2

Figure S4. The boxplots between the F1 score and the cutoff (including the weight) values. The upper row is for Dataset I and the boxplots for Dataset II are in the bottom row. The first column is between F1 score and the cutoff value, *k*, of the distance-based window (DW-PAS). The middle column is between F1 score and the cutoff value, ρ, of the similarity-based window (SW-PAD). The last column is between F1 score and the weight, *w*, of the mixture similarity measure (PAM).

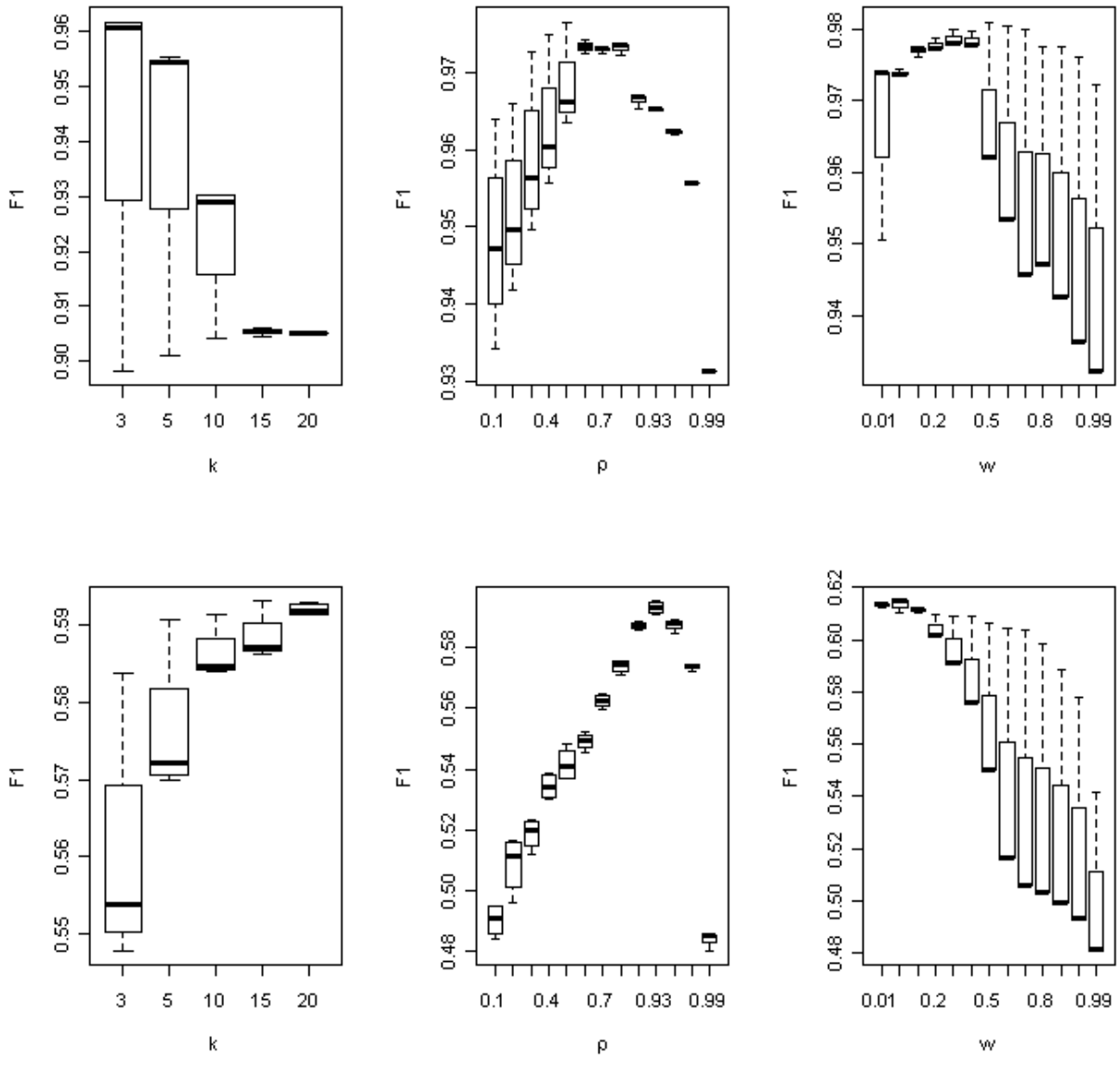
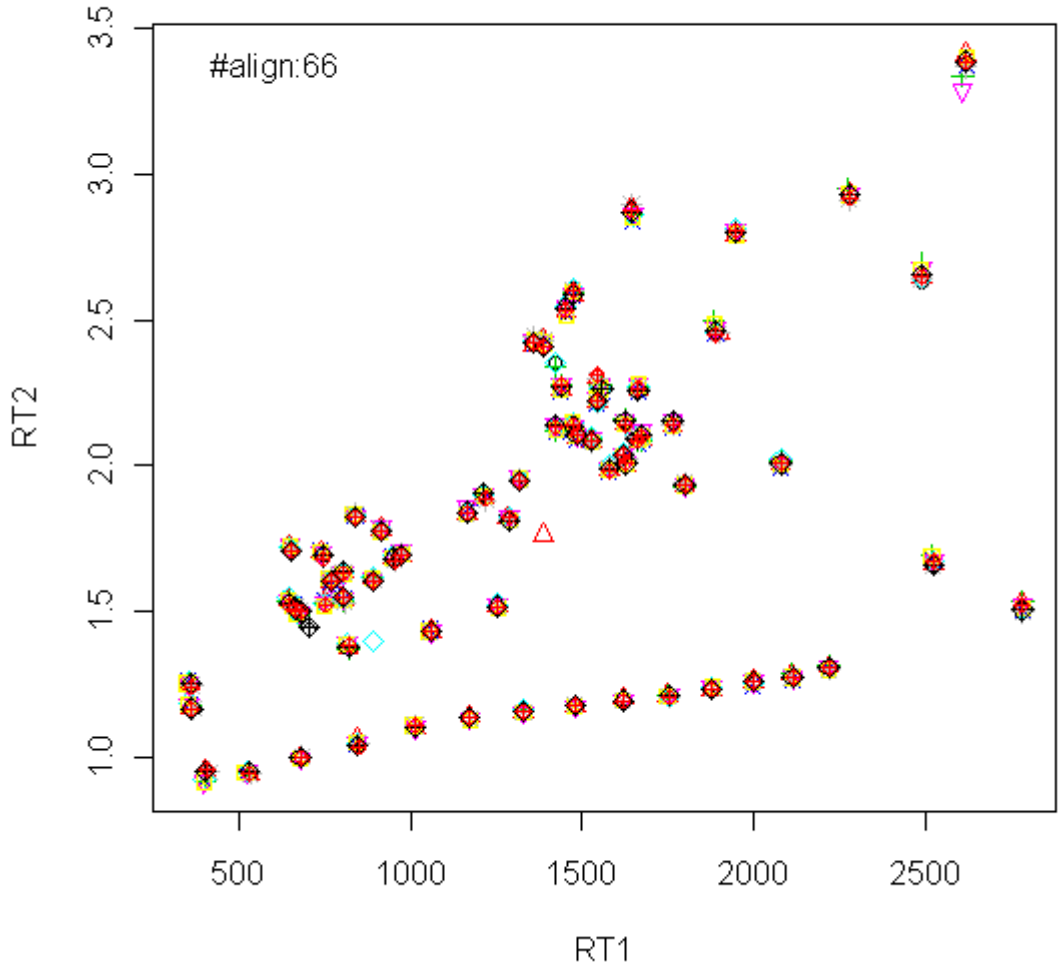


Figure S5. The true peak alignment of the entire peak list for Dataset I and II. The scatter plots between the first (RT1) and second (RT2) of the true peak alignment are represented for 10 and 5 homogeneous chromotograms of Dataset I (a) and Dataset II (b), respectively. Dataset I has a total of 66 peaks matched and 146 peaks are aligned for Dataset II.

(a)



(b)

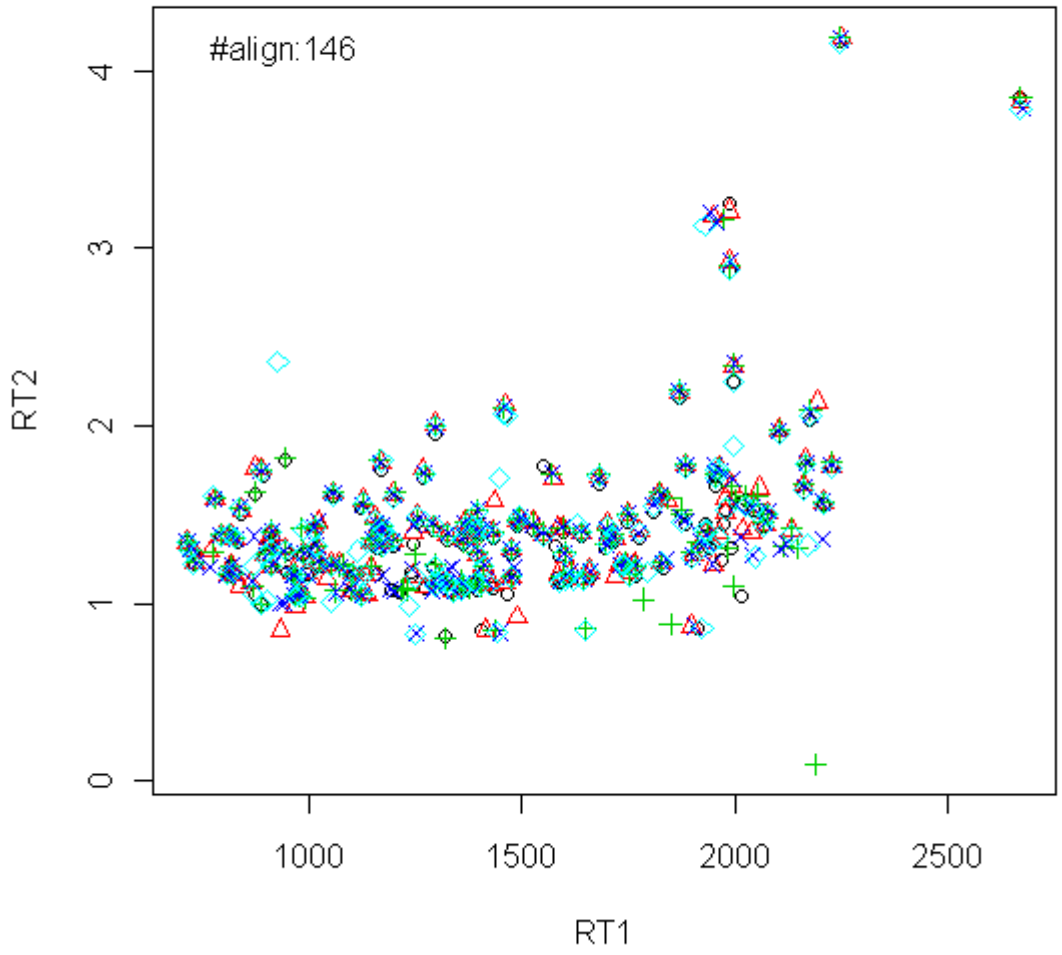
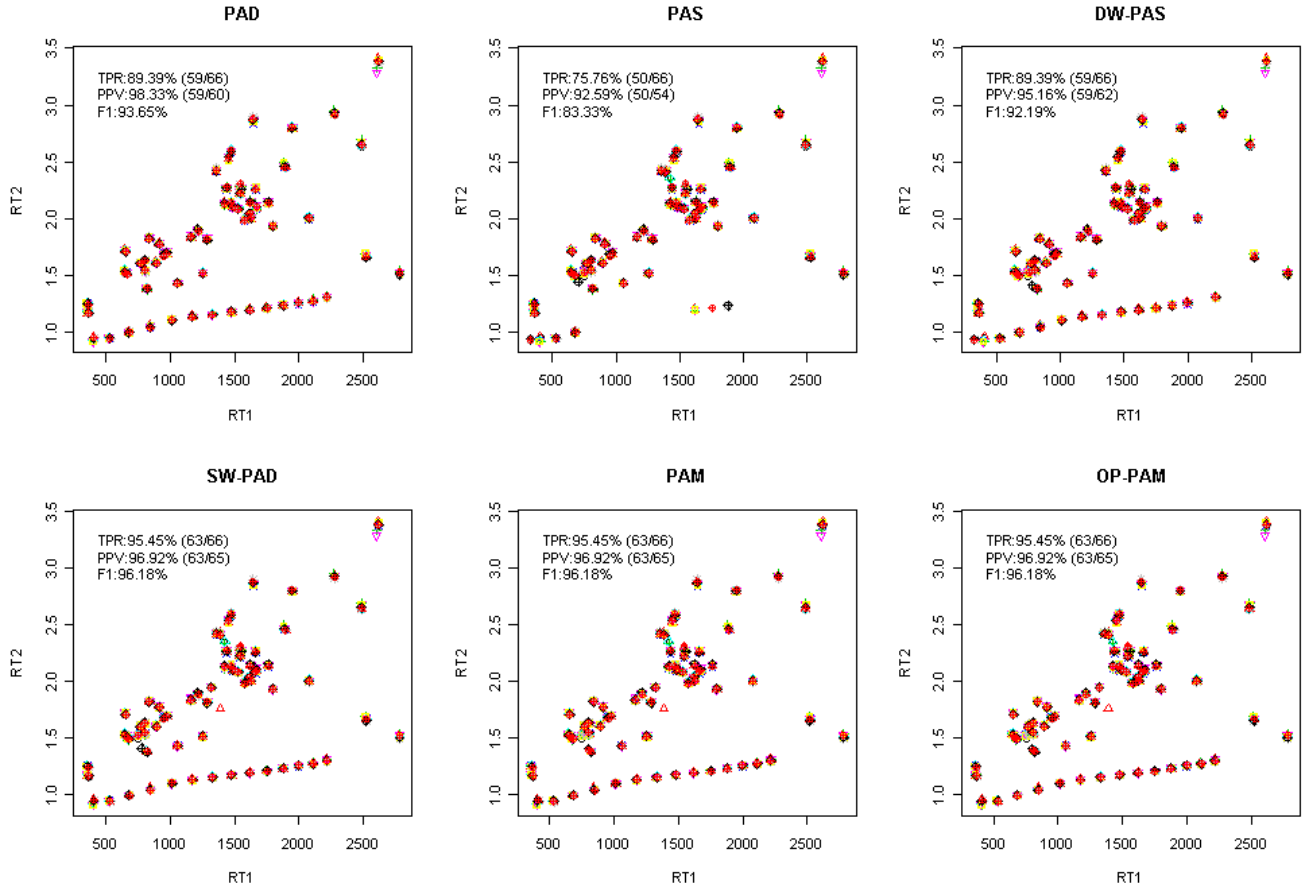
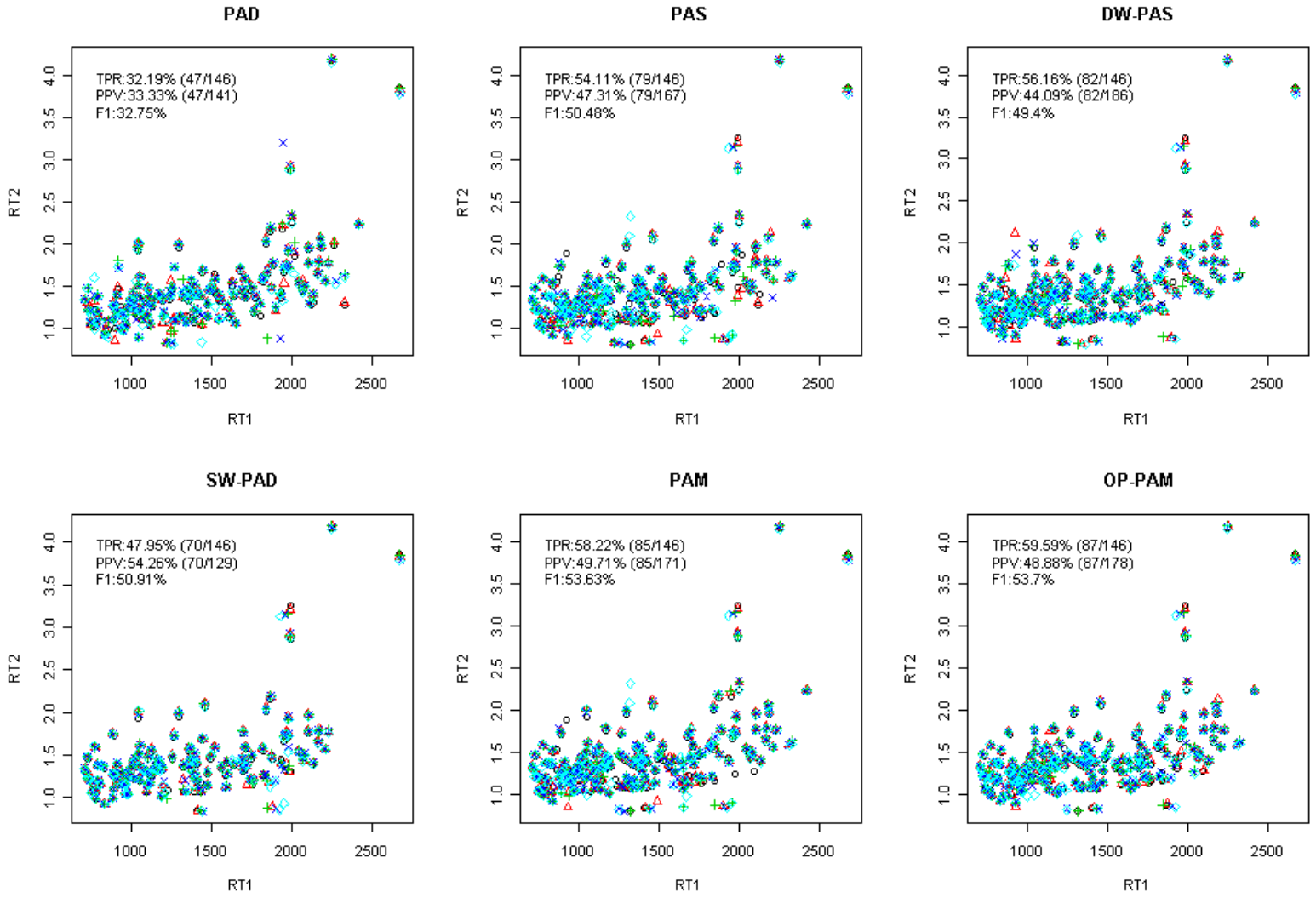


Figure S6. Peak Alignment of the entire peak list. The scatter plots between the first (RT1) and second (RT2) retention times are represented after aligning 10 and 5 homogeneous chromotograms for Dataset I and II, respectively. (a) The six alignment methods are applied to Dataset I. (b) The six alignment methods are applied to Dataset II. The true peak alignments for Dataset I and II can be found in Figure S5. For each peak alignment, TPR, PPV, and F1 score are estimated based on the true peak alignemtns.

(a)



(b)


PAD
TPR:32.19% (47/146)
PPV:33.33% (47/141)
F1:32.75%
PAS
TPR:54.11% (79/146)
PPV:47.31% (79/167)
F1:50.48%
DW-PAS
TPR:56.16% (82/146)
PPV:44.09% (82/186)
F1:49.4%
SW-PAD
TPR:47.95% (70/146)
PPV:54.26% (70/129)
F1:50.91%
PAM
TPR:58.22% (85/146)
PPV:49.71% (85/171)
F1:53.63%
OP-PAM
TPR:59.59% (87/146)
PPV:48.88% (87/178)
F1:53.7%
RT1
RT2