Metabolomics of biofluids : from analytical tools to data interpretation

Nevedomskaya, E.

Chapter 2

Alignment of capillary electrophoresis–mass spectrometry datasets using accurate mass information

Nevedomskaya E., Derks R., Deelder A.M., Mayboroda O.A., Palmblad M.

Analytical and Bioanalytical Chemistry **2009**, 395, 2527–2533

ABSTRACT

Capillary electrophoresis–mass spectrometry (CE–MS) is a powerful technique for the analysis of small soluble compounds in biological fluids. A major drawback of CE is the poor migration time reproducibility, which makes it difficult to combine data from different experiments and correctly assign compounds. A number of alignment algorithms have been developed but not all of them can cope with large and irregular time shifts between CE–MS runs. Here we present a genetic algorithm designed for alignment of CE–MS data using accurate mass information. The utility of the algorithm was demonstrated on real data, and the results were compared with one of the existing packages. The new algorithm showed a significant reduction of elution time variation in the aligned datasets. The importance of mass accuracy for the performance of the algorithm was also demonstrated by comparing alignments of datasets from a standard time-of-flight (TOF) instrument with those from the new ultrahigh resolution TOF maXis (Bruker Daltonics).

INTRODUCTION

Capillary electrophoresis (CE) is an ideal technique for separation of small soluble polar compounds that are present in biological fluids.(1) There are also other advantages of CE for analysis of biological fluids, such as relatively short separation times with good resolution and low sample consumption.(2) CE is often criticized for its low loading capacity. However, pre-concentration techniques such as pH-mediated stacking (3) can overcome this drawback. If a mass spectrometer is used as a detector (CE–mass spectrometry (MS)), additional information on mass and isotopic distribution (4) is provided, which enables compounds and potential biomarkers to be identified. For comparison of multiple samples, elution or migration time precision is also very important. This is a serious concern for CE which, especially when bare-fused silica capillaries are used, lacks reproducibility of migration time.(5) Low reproducibility of migration time affects not only identification of compounds and their synchronization between samples but also statistical analysis. Misalignment introduces variation in the data that will noticeably affect results of multivariate statistics (6) and for studies involving numerous samples, as typically encountered in clinical research, manually assisted alignment of CE–MS datasets is not feasible.

The data produced by CE–MS is three-dimensional: intensity as a function of time and mass-to-charge ratio. There are two main strategies for alignment of this type of data: (1) to group features together in matrices that can be further statistically analyzed or (2) to transform all time axes to a common axis with further analysis of aligned signals. (7) The former works well for protein and peptide data, as some peaks can be identified and used as internal standards for quantification and correction, but is problematic in case of metabolomics.(8) In addition, for complex and overlapped electropherograms, peak assignment is less reliable.(9) For the second strategy, there are already many algorithms and software packages available. However, most of these programs have been developed for liquid chromatography- mass spectrometry (LC-MS) and cannot deal with the large and irregular time shifts typically encountered in CE–MS. Furthermore, a majority of these programs use only chromatographic information, aligning base peak or total ion chromatograms using different time warping procedures. As has been mentioned by Daszykowski *et al.*(10), the next step in the development of alignment methods should take advantage of mass as well as chromatographic information.

Another issue for data processing tools is to have them platform independent, to be able to share results within the scientific community. Commercial software often works with data from certain instruments using their specific formats, making them vendor dependent. Free software is working with data formats that can be generated by a large number of tools

and programs for format conversion, such as mzXML.(11) Currently, mzXML is the preferred format to generate aligned CE–MS data as many programs exist which can read this format for further analysis. In this paper we describe the adaptation and application of an algorithm originally developed for LC–MS and LC–MS/MS (12) for alignment of CE–MS datasets—msalign2. Previously published algorithm was developed for alignment of LC–MS and LC–MS/MS data generated by two different mass analyzers (for example, high resolution data of FTICR and low resolution data of ion-trap). The latter was used for confident identification of peptides, masses of which were then matched to masses in LC–MS dataset. The new msalign2 is an alignment method for hyphenated MS applications. It is not limited to only capillary electrophoresis but can be as well used for any hyphenated technique, for instance LC–MS. CE–MS was chosen as it represents the most challenging task for alignment, and there is a demand for this type of software.

The algorithm and ancillary software is implemented in C and R and is available as open source (http://www.ms-utils.org/msalign2/). The algorithm has been shown to work on real CE–MS datasets of urine that are representative of data from a biomedical study. The results have been compared with another alignment tool in the open-source package XCMS (13) in terms of efficiency and relevance of further statistical analysis, visually inspecting and comparing principal component analysis (PCA) results. Our algorithm showed reduced variance in the data and performed better for multivariate analysis.

THEORY

Two CE–MS datasets can be aligned by matching masses across samples and fitting a curve to these matches. The curve represents the relation between electropherograms.

The shifts in migration time are not linear.(14) Non-linearity can be introduced by changes in conductivity and electroosmotic flow or the sheath liquid flow driven by a mechanical pump. That is why the natural solution to alignment problem in CE–MS (and LC–MS) is a piece-wise function of time that can cope with these irregularities.

The problem of finding a function best fitting measured data is an optimization problem. Genetic algorithms are one class of methods for solving this problem. These algorithms were developed in the 1960s from earlier published computer simulations of evolution and artificial selection.(15) A genetic algorithm (GA) is able to find exact or approximate solutions for optimization problems.

GA operates on a population of possible solutions for a problem, called chromosomes. The starting population is created randomly and then goes through a number of generations being transformed by the operators of inheritance: mutation, selection, and recombination. Chromosomes are encoded in such a way that they are suitable for applying

these operators. A function for computing the quality of each chromosome is required. Using this fitness function best candidate solutions are selected in each generation and allowed to reproduce. In the end the global optimum or its close approximation is found.

For aligning CE–MS datasets, a candidate solution is set by breakpoints of the piece-wise function. The fitness function $F(s_i)$ for a candidate solution s_i is calculated as:

$$F(s_i) = \sum_{m=1}^{N} e^{-\frac{(y_m - y(x_m))^2}{2\sigma^2}} - kn_i \qquad (1)$$

where m are peaks out of all N matches with retention times x_m in the dataset to be aligned, y_m in the reference dataset; n_i is the number of breakpoints in chromosome s_i, k is a cost per breakpoint, and σ^2 is the residual variance between the datasets. The fitness function is the sum of likelihoods of observing peak m at a retention times x_m and y_m in the two datasets with the piece-wise function evaluating $y(x_m)$. The cost for a breakpoint is introduced so that the number of breakpoints in the alignment is not too large. The residual variance can be provided by the user based on the knowledge of the analytical system used, or, if absent, is estimated automatically by the software, as previously described.

In each generation, half of the chromosomes in the population with the lowest fitness are replaced by copies of half of the chromosomes with the highest fitness, applying three types of mutations: insertion (randomly adding a breakpoint anywhere in the alignment interval), deletion (removing a breakpoint), and shifting a breakpoint by a small random amount in both dimensions. After a single pass through the GA, the solution of the highest fitness was chosen as the alignment of the two datasets.

The genetic algorithm was run for 1,000 generations with a population size of 300 candidate solutions with maximum number of breakpoints of 12 and the cost for breakpoint set to 0.5.

MATERIALS AND METHODS

Chemicals. Methanol (MeOH) HPLC-grade (Biosolve B.V., Netherlands), ultrapure water (18.2 MΩ/cm), and formic acid (FA) (Fluka, Germany) were used for solvent preparation. NH_4OH was from Sigma-Aldrich, NaOH from J.T. Baker.

Urine samples. Urine was collected and pooled from two groups of mice wild-type 129 and Swiss mice and stored at −20 °C. Urine (4 μL) was mixed with 4 μL of MeOH, 11 μL of water, and 1 μL of BGE, centrifuged to eliminate any possible sediment remaining, and put into vials for injection into CE instrument.

Instrumentation. CE was performed on a PA 800 (Beckman Coulter, Fullerton, CA, USA) as described before.(16) Uncoated fused silica capillaries (BGB-Analytik, Germany)

of total length of 100 cm with 50 μm inner diameter were used for separation. MeOH (20%) with 2 M FA was used as background electrolyte. Sample injection was performed hydrodynamically with pH-mediated stacking: a small plug (50 mbar, 9 s) of 12.5% NH_4OH was injected before the sample plug (50 mbar, 90 s).

The second set of mouse urine has been measured with 0.1 M NaOH washing step between the runs.

MS was performed using two types of time-of-flight (TOF) mass spectrometers: the micrOTOF (Bruker Daltonics, Bremen, Germany) and the new ultrahigh resolution TOF (UHR-TOF), maXis from the same vendor. The acquisition and spraying parameters were optimized so that the total areas on both instruments were identical. Transfer parameters were optimized by direct infusion of an ESI tuning mix (Agilent Technologies, Waldbronn, Germany). Spectra were collected with a time resolution of 1 s. CE–MS coupling was realized by a co-axial sheath liquid interface (Agilent Technologies, Waldbronn, Germany) with methanol–water–formic acid (50:50:0.1, *v/v/v*) as sheath liquid. The following spray conditions were used: sheath liquid flow, 4 μL/min; dry gas temperature, 180 °C; nitrogen flow, 4 L/min; nebulizer pressure, 0.5 bar. Electrospray in positive ionization mode was achieved and ESI voltage was −4.5 kV.

Data analysis. Electropherograms were aligned using in-house-developed genetic algorithm running 1,000 generations and by XCMS (The Scripps Research Institute, La Jolla, USA). Data were normalized using non-parametric algorithm as described earlier.(17) The results of alignment were analyzed by PCA in SIMCA-P+ software (Umetrics, Umeå, Sweden). All calculations were performed on a standard office PC (Core 2 Quad, 2.4 GHz, 2 GB RAM). The alignment of 20 datasets took about 40 min and used up to 200 MB RAM. This time is tens of times less than the time needed for acquisition of the data, so it does not represent a bottleneck in the whole analysis pipeline.

RESULTS AND DISCUSSION

The algorithm works pair-wise, operating on two CE–MS datasets in the mzXML format, which makes it platform independent and suitable for alignment of data generated by any type of CE–MS instrumentation. To demonstrate the alignment performance of the algorithm, 20 electropherograms were aligned. Figure 1 represents the result of alignment.

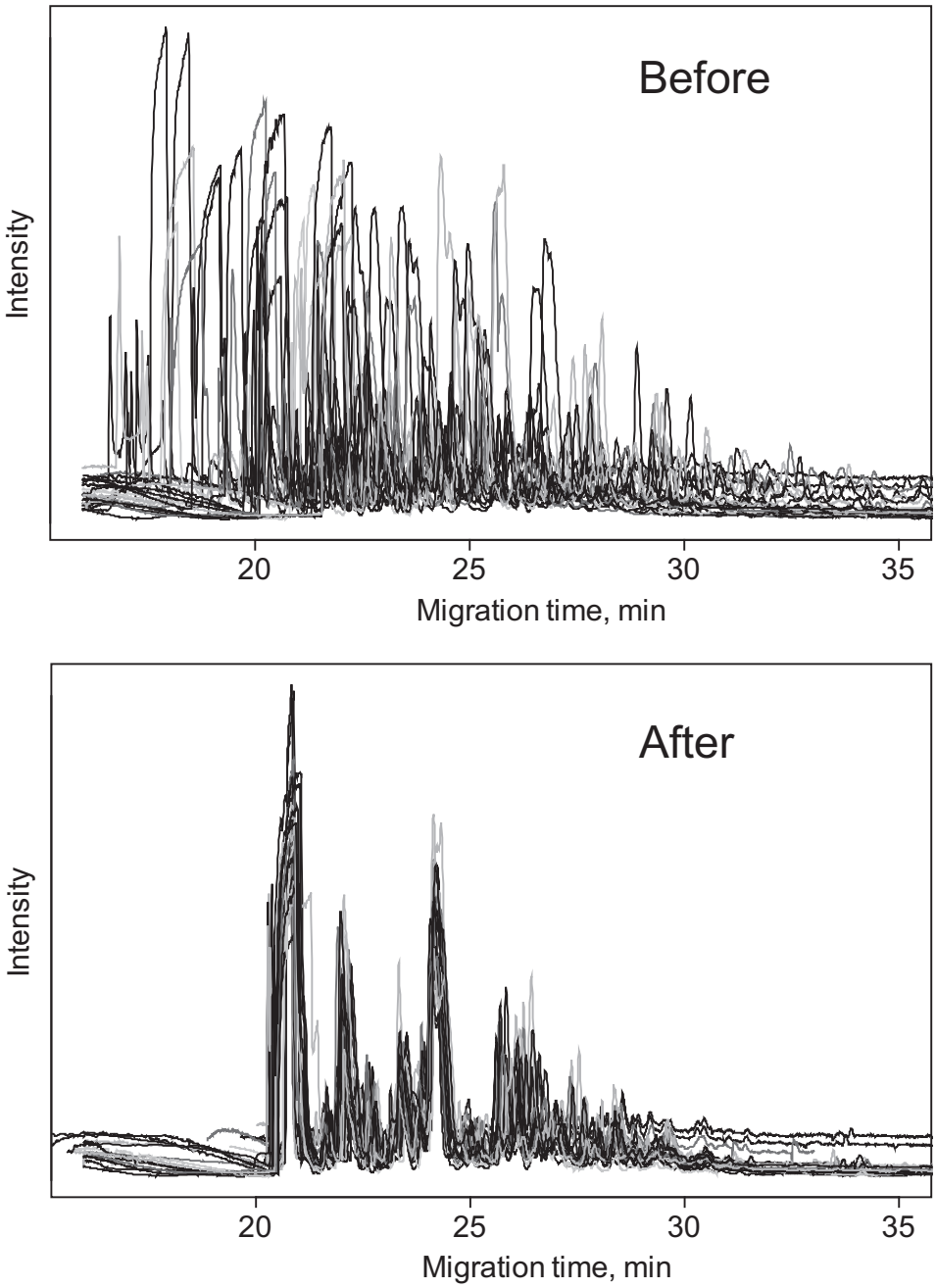

Figure 1. Total ion electropherograms of 20 CE–MS datasets before and after alignment. In the region where compounds migrate the alignment works for each single dataset.

A visual inspection already shows a significant improvement in peak positions. As previously mentioned, migration time shifts are not linear, so the evaluation of alignments should be done not on a single peak but better on several peaks in different parts of the electropherogram. The relative standard deviations (RSD) in migration time were calculated for three peaks at the beginning, middle, and end of the electropherogram and ranged from 5.6% to 6.5% before alignment. The RSD was higher for peaks, which are closer to the end of electropherogram. This happens because toward the end of the run adsorption on the walls of the capillary accumulates, and shifts in migration times increase.(18) After the alignment, the RSD significantly improved and varied from 0.12% to 0.99%. Only analytical window was used for alignment, excluding the time at the beginning of the run where the sodium clusters are migrating. It can be seen on the aligned chromatograms that at the beginning of the run, the variation in migration time is still present, whereas in the rest of the electropherogram, peaks are very well aligned.

To show how alignment improves further statistical analysis, a case study consisting of two groups of measurements was selected. Two-group scenario of data analysis is common in clinical research where there are typically groups of samples of patients and controls or patients at different time points or with different treatments. For our study we used two groups, each consisting of nine electropherograms of pooled mouse urine from two strains of mice (129 and Swiss). The samples from the second group were measured with a NaOH washing step between samples, which introduces a systematic difference between the groups and allows to show the robustness of the algorithm. The washing step is important in CE–MS. When a biological matrix is introduced, the capillary is contaminated, leading to large shifts in migration time, clogging, or even breakdown of the capillary. This is especially the case when bare-fused silica in a low pH system is used. There are two principal solutions to this problem. The first is changing the capillaries after a certain number of runs, introducing some additional variation, and then aligning all the datasets using one of the available algorithms. Alternatively, it is possible to regenerate the capillary with sodium hydroxide after each run. The second option is more time consuming as it requires washing steps with NaOH followed by water. However as we also show here, the washing strategy gives better analytical results with less variation in the data. The variation in the data without the washing step cannot be entirely eliminated even when advanced alignment techniques are used.

We compared the alignment of 18 electropherograms by our piece-wise alignment and by one of the existing methods — XCMS. XCMS was chosen because it is a very powerful and quite widely used package, for which our tool can be a useful complement. The

workflow of the alignment in XCMS is based on completely different principles and includes peak detection and matching prior to time correction.

The ubiquitous creatinine peak was used for visual inspection and control of the alignment efficiency by two programs. The extracted ion chromatograms of creatinine (m/z 114.12±0.05) are shown in Fig. 2. Both algorithms perform quite satisfactorily but it is apparent that XCMS leaves some of the peaks unaligned, whereas msalign2 successfully aligned all datasets.

This may be because the large number of changeable parameters in XCMS makes it difficult for the user to optimize the process and find the ideal settings for given sets of data. In contrast, msalign2 has a minimum number of parameters, most of which never have to be changed between one sample and the next. The free parameters are mass measurement error, background threshold, start and end scans for the time interval for alignment, and expected variance in elution time. The last is optional, and if the user does not supply it, the program will automatically estimate the variance. Mass measurement error is given in ppm and depends on the type of mass spectrometer being used and not the separation technique. The background threshold should be chosen such that the number of matched masses across samples is on the order of 1,000 to get reliable alignments with a reasonable computational time. Start and end scan numbers are used to align only the informative part of the electropherograms, disregarding intense signals that commonly appear at the beginning and end of each run, such as sodium clusters and compounds that have attached to the capillary wall and are released during the washing step. The web-application contains, besides the alignment algorithm, a tool to estimate the background threshold needed to get specified number of matched features for the alignment.

Another possibility why some misalignment appears in the case of XCMS is that the package was primarily designed for LC–MS and can give suboptimal results for CE–MS, which may have much larger shifts in time domain. Time shifts are crucial for the matching step performed in XCMS before the alignment. Nevertheless it does not decrease the applicability and usefulness of the XCMS package as it includes not only an alignment algorithm but also powerful peak picking that might for instance be used after performing alignment with msalign2.

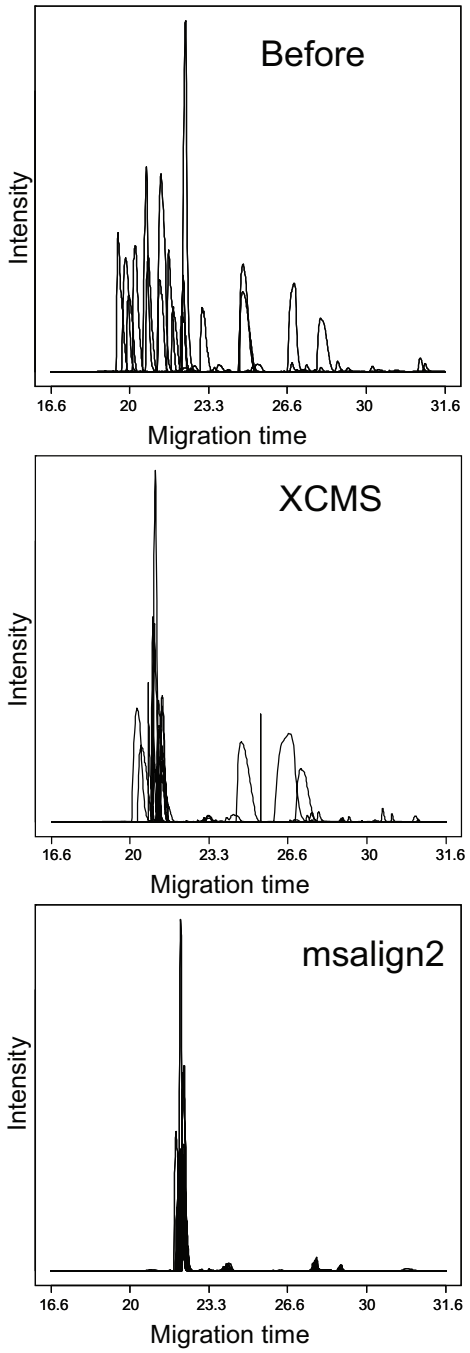

Figure 2. Extracted ion electropherograms of creatinine before alignment, aligned with XCMS, and aligned with msalign2.

The output of XCMS contains a table with detected features (chromatographic peaks) and their intensities throughout all samples. It can be directly used for analysis with any statistical package, for instance SIMCA-P+.(19) The output of our algorithm is an mzXML file with corrected retention times. This makes it easy to explore how well the alignment works. To apply further statistics, one can either perform binning or peak picking from these mzXML files. Here we used the peak picking from the XCMS package, so that the only difference between the two methods is the alignment step. The resulting tables from both alignments were normalized as described above and imported to the SIMCA-P+ software package for multivariate analysis.

PCA is a usual first step in analyzing multivariate data. It shows the correlation structure present in the data and the directions of the largest variance. Tables produced by peak picking step were normalized and used as input for PCA. The scores plots are shown in Fig. 3.

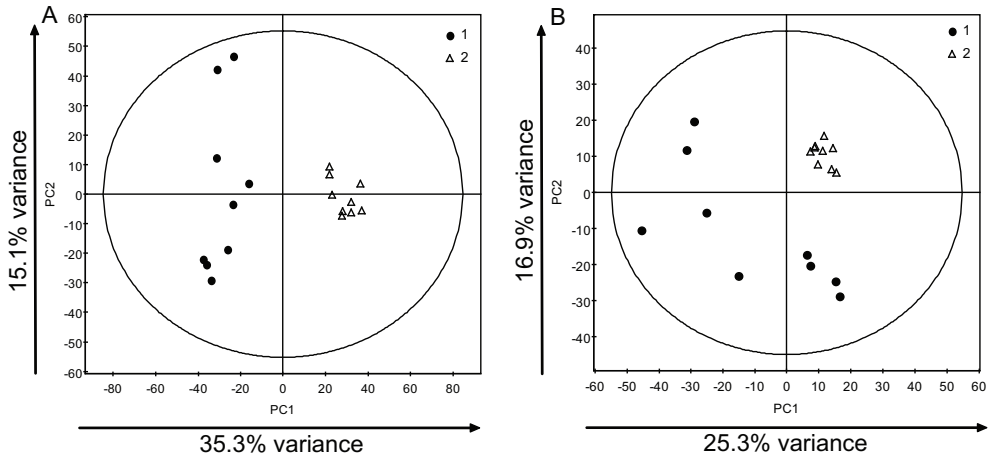

Figure 3. PCA scores plots: A data from peak picking by XCMS after genetic algorithm alignment and B data obtained from XCMS alignment. The variance explained by the components is indicated on arrows along the axes. Group 1 (without washing step) is indicated by black dots; group 2 (with washing step) is indicated by triangles.

The overall picture is the same in both cases, with the two sample groups nicely separated. The datasets acquired with the washing step between each sample show significantly smaller variance. However, more variation (>50%) is explained by first two principal components with the msalign2 than with the XCMS alignment (<50%). This is due to the additional variation introduced by misalignment in XCMS. It is important to reduce systematic variability in the data as far as possible, and not introduce additional

variability by alignment or normalization procedures that can obscure the chemical species correlating with or being responsible for the actual phenomena under study.

As mentioned above, an important feature of our algorithm is that it uses accurate mass information. In theory, the better the mass accuracy, the easier the alignment task. To test this hypothesis we generated electropherograms of the same pooled urine using two different mass spectrometers: a standard orthogonal TOF (Bruker micrOTOF) and the recently released ultrahigh resolution TOF instrument (Bruker maXis). Two main differences between these instruments are the resolving power (40,000 vs 20,000 at *m/z* 600) and mass measurement precision (0.8 vs 3 ppm).

Mass electropherograms generated using these machines were indeed found to differ in resolution and mass accuracy. msalign2 performed well on both pairs of data, but as can be seen in Fig. 4, there are significantly fewer mismatched features between the maXis datasets (7%) than between the micrOTOF datasets (more than 40%).

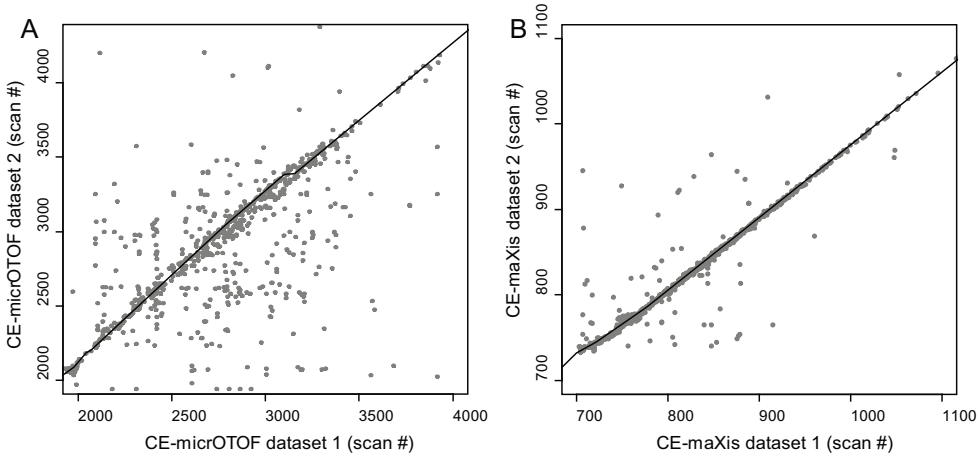

Figure 4. Alignments of CE–MS datasets from two types of instruments—micrOTOF (a) and maXis (b). The matched features are presented as gray dots; black line represents the trend line found by genetic algorithm.

These results demonstrate that more compounds can be correctly identified based on accurate mass with the new UHR-TOF instrument compared to the standard TOF. The alignment problem is easier to solve with better mass accuracy, but the genetic algorithm is sufficiently robust to find the correct global alignment also between the micrOTOF datasets.

The genetic algorithm and supporting software is implemented in C, R and is available as open source on http://www.ms-utils.org/msalign2.

Here we presented the application of our algorithm to CE–MS alignment, which is the worst case among separation techniques in terms of time reproducibility. As our method does not depend on chromatographic parameters and quality, it can as well be used for alignment of LC–MS and GC–MS data.

The analysis of large amounts of data generated by metabolomic, proteomic, peptidomic, or other types of "omic" experiments includes many steps, most of which need sophisticated algorithms and computational implementation. Small mistakes and imperfections on each of these steps can lead to incorrect data interpretation and misleading results. The consequences may be even more dramatic in the field of system biology when the data from different analytical platforms and from different levels of biological organization have to be combined. Alignment of chromato- (electrophero-) grams is just one of the steps of data analysis but is an important one and should be the subject of careful examination and optimization, as was performed in this study.

CONCLUSIONS

Here we present a platform-independent, open-source algorithm for alignment of complex CE–MS datasets. In contrast to other available alignment algorithms it efficiently uses mass information. Performance of MS instrumentation, mass accuracy, and resolving power positively affect alignment results. However, we have clearly shown that the algorithm is robust enough and performs even with relatively "low cost" MS instrumentation. It is shown also that the variation should be reduced not only by means of data processing but also by selecting proper experimental conditions.

As a tool for alignment of CE–MS data, our algorithm outperforms such packages as XCMS resulting in reduced variation that appears in multivariate analysis. On the other hand, alignment is only a step in the data processing pipeline and as such our algorithm is fully complimentary to XCMS.

Although in this paper we have focused on CE–MS as this approach represents one of the most challenging tasks for alignment, the algorithm can obviously as well be used for alignment of LC–MS and GC–MS datasets.

REFERENCES

1. Monton,M.R.N., and Soga,T. 2007. Metabolome analysis by capillary electrophoresis-mass spectrometry. Journal of Chromatography A 1168:237-246.

2. Song,E.J., Babar,S.M., Oh,E., Hasan,M.N., Hong,H.M., and Yoo,Y.S. 2008. CE at the omics level: towards systems biology--an update. Electrophoresis 29:129-142.

3. Neususs,C., Pelzing,M., and Macht,M. 2002. A robust approach for the analysis of peptides in the low femtomole range by capillary electrophoresis-tandem mass spectrometry. Electrophoresis 23:3149-3159.

4. Ojanpera,S., Pelander,A., Pelzing,M., Krebs,I., Vuori,E., and Ojanpera,I. 2006. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 20:1161-1167.

5. Garcia-Perez,I., Vallejo,M., Garcia,A., Legido-Quigley,C., and Barbas,C. 2008. Metabolic fingerprinting with capillary electrophoresis. J. Chromatogr. A 1204:130-139.

6. van Nederkassel,A.M., Xu,C.J., Lancelin,P., Sarraf,M., Mackenzie,D.A., Walton,N.J., Bensaid,F., Lees,M., Martin,G.J., Desmurs,J.R. *et al* 2006. Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. J. Chromatogr. A 1120:291-298.

7. Katajamaa,M., and Oresic,M. 2007. Data processing for mass spectrometry-based metabolomics. J. Chromatogr. A 1158:318-328.

8. Aberg,K.M., Alm,E., and Torgrip,R.J. 2009. The correspondence problem for metabonomics datasets. Anal Bioanal Chem.

9. Johnson,K.J., Wright,B.W., Jarman,K.H., and Synovec,R.E. 2003. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. J. Chromatogr. A 996:141-155.

10. Daszykowski,M., Danielsson,R., and Walczak,B. 2008. No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry. J. Chromatogr. A 1192:157-165.

11. Pedrioli,P.G., Eng,J.K., Hubley,R., Vogelzang,M., Deutsch,E.W., Raught,B., Pratt,B., Nilsson,E., Angeletti,R.H., Apweiler,R. *et al* 2004. A common open representation of mass spectrometry data and its application to proteomics research. Nat. Biotechnol. 22:1459-1466.

12. Palmblad,M., Mills,D.J., Bindschedler,L.V., and Cramer,R. 2007. Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. J. Am. Soc. Mass Spectrom. 18:1835-1843.

13. Smith,C.A., Want,E.J., O'Maille,G., Abagyan,R., and Siuzdak,G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78:779-787.

14. America,A.H., Cordewener,J.H., van Geffen,M.H., Lommen,A., Vissers,J.P., Bino,R.J., and Hall,R.D. 2006. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. Proteomics 6:641-653.

15. Fraser,A. 1957. Simulation of genetic systems by automatic digital computers. I. Introduction. Australian J. Biol. Sci. 10:484-491.

16. Mayboroda,O.A., Neususs,C., Pelzing,M., Zurek,G., Derks,R., Meulenbelt,I., Kloppenburg,M., Slagboom,E.P., and Deelder,A.M. 2007. Amino acid profiling in urine by capillary zone electrophoresis - mass spectrometry. J. Chromatogr. A 1159:149-153.

17. Sidorov I.A., Hosack D.A., Gee D., Yang J., Cam M.C., Lempicki R.A., and Dimitrov D.S. 2002. Oligonucleotide microarray data distribution and normalization. Information Sciences 146:67-73.

18. Stutz,H. 2009. Protein attachment onto silica surfaces - a survey of molecular fundamentals, resulting effects and novel preventive strategies in CE. Electrophoresis 30:2032-2061.

19. Nordstrom,A., O'Maille,G., Qin,C., and Siuzdak,G. 2006. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. Anal Chem 78:3289-3295.