Available online at www.sciencedirect.com
ScienceDirect
Journal of Pharmaceutical and Biomedical Analysis 43 (2007) 413–420
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Increasing conclusiveness of metabonomic studies by cheminformatic preprocessing of capillary electrophoretic data on urinary nucleoside profiles

E. Szymańska [a], M.J. Markuszewski [a], X. Capron [b], A.-M. van Nederkassel [b], Y. Vander Heyden [b], M. Markuszewski [c], K. Krajka [c], R. Kaliszan [a,*]

[a] *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland*
[b] *Department of Analytical Chemistry and Pharmaceutical Technology, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium*
[c] *Department of Urology, Medical University of Gdańsk, Kliniczna 1a, 80-402 Gdańsk, Poland*

Received 4 July 2006; received in revised form 16 August 2006; accepted 16 August 2006
Available online 26 September 2006

Abstract

Nowadays, bioinformatics offers advanced tools and procedures of data mining aimed at finding consistent patterns or systematic relationships between variables. Numerous metabolites concentrations can readily be determined in a given biological system by high-throughput analytical methods. However, such row analytical data comprise noninformative components due to many disturbances normally occurring in analysis of biological samples. To eliminate those unwanted original analytical data components advanced chemometric data preprocessing methods might be of help. Here, such methods are applied to electrophoretic nucleoside profiles in urine samples of cancer patients and healthy volunteers. The electrophoretic nucleoside profiles were obtained under following conditions: 100 mM borate, 72.5 mM phosphate, 160 mM SDS, pH 6.7; 25 kV voltage, 30 °C temperature; untreated fused silica capillary 70 cm effective length, 50 μm I.D. Different most advanced preprocessing tools were applied for baseline correction, denoising and alignment of electrophoretic data. That approach was compared to standard procedure of electrophoretic peak integration. The best results of preprocessing were obtained after application of the so-called correlation optimized warping (COW) to align the data. The principal component analysis (PCA) of preprocessed data provides a clearly better consistency of the nucleoside electrophoretic profiles with health status of subjects than PCA of peak areas of original data (without preprocessing).

Keywords: Metabonomics; Modified nucleosides; Urogenital cancer; Metabolic profiling; Data preprocessing

1. Introduction

Bioinformatics is the application of computer sciences and mathematics to the management and analysis of biological datasets to aid the solution of biological problems [1,2]. Nowadays, in the post-genomic era, large databases containing metabonomic, proteomic and transcriptomic data are created and attention should be focused to their storage, management, analysis as well as extraction and ultimate application of systematic information they convey [3]. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of "omics" databases are increasing even faster. Therefore, appropriate characterization and classification of data processing tools as well as creation of new computational procedures (algorithms) is unavoidable [3].

According to the definition of the Metabolomics Society (http://www.metabolomicssociety.org), metabolomics is the study of metabolic changes that encompasses metabolite target analysis, metabolite profiling, metabolic fingerprinting, metabolic profiling and metabonomics. Metabonomics can be understood as comprehensive analysis of endogenous metabolites changes in biological fluids and tissues that result from

* Corresponding author at: Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland. Tel.: +48 58 3493260; fax: +48 58 3493262.
E-mail address: roman.kaliszan@amg.gda.pl (R. Kaliszan).

0731-7085/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2006.08.015

disease or therapeutic treatment. Since metabolites are the final products of cellular regulatory processes, their quantitative levels can be regarded as the ultimate response of biological systems to genetic and environmental changes [4]. Data obtained from metabolome analysis can be used for various aims, like simulation of the biological activity with genes coded in genome, production of valuable metabolites by gene technology [5] and diagnosis of various pathological states [6–9]. A specific feature of metabolomics is its reductive nature. Currently it focuses on ca. 2400 compounds, compared to 25,000 genes and about one million proteins and peptides to be considered in genomic and proteomic studies, respectively [10,11]. Of course, particular metabolite is usually involved in several pathways. It can be rationally assumed that metabolite profile patterns might be characteristic for specific diseases, however. Nowadays, one can imagine determination of all the metabolites by high-throughput automatized and roboticized analytical techniques, followed by a fast and reliable pattern recognition by generic model fitting or classification algorithms. That should result in predictive data mining. However, such a diagnostics "philosopher stone" would be impractical, if at all possible. Instead, considering of limited sets of metabolites appears advisable in a more or less specific disease diagnostics.

The approach needs not to rely on any reasoning or understanding the mechanisms of the processes. However, it must be shown to provide correct predictions or classification in cross-validation samples. For that aim proper preprocessing of analytical data seems to be of utmost importance to provide eventually consistent patterns or systemic relationships between variables and then to validate the conclusions by applying the identified patterns to new subsets of data.

In case of heterogeneous diseases, like cancer, a panel of biomarkers (metabolites) determined through the use of multiple high throughput platforms, might provide reliable information on health status of the patients, which is normally not provided by a single variable (biomarker) [6,9].

To be useful, biomarkers not only must distinguish between subjects with a given diseases and those without it, but also their assay methods should be validated and readily employed. Researches from different laboratories should use the same experimental protocol and compare their profiles against those of others in universal database.

The optimal practice in analysis of biological samples should include selection of appropriate analytical methods and collection of analytical data, followed by application of multivariate data processing models, such as principal component analysis (PCA), partial least squares (PLS) or parallel factor analysis (PARAFAC) for explanatory purposes. All these steps should be robust and fast enough to deal with many disturbances normally occurring in analysis of various biological samples. That is essential in metabonomic studies, where database may comprise hundreds or thousands of variables. Usually, variations observed in metabonomic measurements are due to complexity and diversity of analyzed biofluid samples (matrix effects: sample-to-sample), mechanical drift (fluctuations: run-to-run) as well as imperfections of analytical methods in the long-term and large-scale analysis projects.

Capillary electrophoresis (CE) is one of the most important analytical methods in modern life sciences laboratories [12–15]. It is employed widely in search for cancer biomarkers [16,17]. Advantages of CE, that make it particularly valuable in metabonomic studies are: high resolution power, relatively short time of analysis and small quantities of both the sample and the background solutions needed for assay. However, CE in comparison to high performance liquid chromatography (HPLC) or gas chromatography (GC), produces less reproducible results, what may pose a problem in long lasting projects. Variations in migration time – a function of electroosmotic flow (EOF) inside the capillary, sample loading, wall interactions and physical errors (such as injection irreproducibility or temperature variations) – may lead to poorly reproducible data and preclude their appropriate interpretation [14,15]. To overcome this problem specific chemometric approaches may be of value for migration time adjustments and peak alignment. After a proper chemometric transformation, the data originating from various sources could be compared and relevant information might be extracted and further investigated by specific advanced explanatory/inductive cheminformatics.

In this study, different chemometric methods were compared in preprocessing of CE data obtained in metabonomic studies. The data were from CE analysis of nucleoside profiles – potential biomarkers of cancer [18–20] – in urine samples from cancer patients and healthy controls. The applicability of different pretreatment tools, as well as impact of preprocessing on evaluation of internal relationships of the data, was investigated. Various baseline correction, denoising and peak matching algorithms have been used. PCA of an original dataset and a derivative dataset, obtained after implementation of individual pretreatment methods, evidences the advantage of the proposed preprocessing of electrophoretic data for conclusiveness of metabonomic studies. The study has a pilot methodological character as it has been done on a relatively small group of 28 subjects. However, preliminary results indicate evident trends in clustering of cancer patients separately from healthy volunteers thus encouraging extension of both the number of subjects and the metabolites assayed. Chemometric preprocessing of the employed analytical data to reduce noninformative components appears advisable.

2. Materials and methods

2.1. Analytical procedures

Spontaneous urine samples from healthy adults and cancer patients from the Department of Urology, Medical University of Gdańsk, Gdańsk, Poland were collected after their informed consents and the studies were performed in accordance to the principles embodied in the Declaration of Helsinki. Cancer patients included in the study were after diagnosed kidney, prostate and bladder cancer. After collection, urine samples were frozen immediately and stored at −24 °C. Directly before the analysis of nucleoside profiles and creatinine concentration the samples were thawed at room temperature.

Urinary nucleosides were analyzed with application of the following stages: sample pretreatment, solid phase extraction (SPE) and capillary electrophoretic (CE) separation and quantification. Additionally, during a separate electrophoretic experiment [21] the concentration of creatinine for every sample was determined. Analytical methods were previously developed and validated in the Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gdańsk, Poland.[1]

CE experiments were carried out on a Beckman Coulter MDQ P/ACE 5510 system (Beckman Instruments, Fullerton, CA, USA), fitted with a diode array UV-absorbance detection (190–600 nm), a temperature-controlled capillary compartment (liquid cooled) and a temperature controlled autosampler (air cooled). Electrophoretic data were acquired (acquisition rate of signals was 4 Hz) and analyzed by 32 Karat Software (Beckman). The electrophoretic dataset was obtained [21] by application of following conditions: 100 mM borate, 72.5 mM phosphate, 160 mM SDS, pH 6.7; 25 kV voltage, 30 °C temperature during analysis; injection 5 s × 0.5 psi; capillary: untreated fused silica 70 cm length to detector, 50 μm I.D. In case of creatinine concentration determination exactly the same electrophoretic conditions were used instead of injection time 10 s × 0.5 psi.

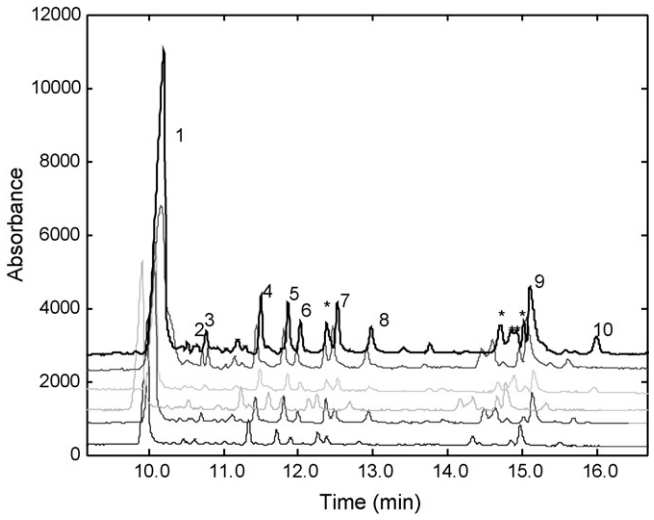

Fig. 1. Six example electrophoretic profiles from the obtained dataset before preprocessing. Peaks: 1, pseudouridine; 2, uridine; 3, cytidine; 4, 5-methyluridine; 5, inosine; 6, N^4-acetylcytidine; 7, guanosine; 8, adenosine; 9, N^2,N^2-dimethylguanosine; 10, xanthosine; (*), unidentified peaks. The electropherogram was obtained under following conditions: 100 mM borate, 72.5 mM phosphate, 160 mM SDS, pH 6.7; 25 kV voltage, 30 °C temperature during analysis; capillary: untreated fused silica 70 cm length to detector, 50 μm I.D.

2.2. Creation of dataset

The electropherograms were imported from 32 Karat Software into Matlab 6.5 software environment for Windows (Mathworks, Natick, MA, USA) as three-dimensional matrix (time, absorbance, wavelength). The objects of preprocessing – data points from 2200 to 4000 (240 data points correspond to 1 min of analysis) – were selected as sections of electropherograms including most peaks (of known and unknown identity) and wavelength of 254 nm was chosen as the most representative (Fig. 1). On that basis, dataset matrix with rows corresponding to each analyzed sample was created. It contains data from 28 CE analyses of urine samples: 18 from cancer patients and 10 from healthy controls.

2.3. Chemometric analysis of data

Different preprocessing tools were compared. These included baseline correction, denoising and alignment of data. In the baseline correction method, the selection of minimum points in data point window and estimation of new baseline by linear interpolation was employed. The proper signal denoising requires knowledge of the nature of the noise occurring in the data and characteristics of the peaks in the time and frequency domains. To limit the noise, the efficient digital signal processing algorithms, such as the Savitzky–Golay (SG) implementation of polynomial least-squares filters, the Cooley–Tukey algorithm for the fast Fourier transform (FFT), derivative filters or discrete wavelet transform [22,23] were applied.

At the stage of alignment the research concentrated on the development of two different methods: dynamic time warping (DTW) and correlation optimized warping (COW), as these are the most common techniques used in preprocessing of chromatographic data [24–28]. In selection of appropriate alignment algorithm three principles were considered according to Johnson et al. [29]:

(1) The algorithm must preserve the chemical selectivity differences between different profiles and limit run-to-run retention/migration time shift.
(2) The algorithm must be fast and less memory-demanding to deal with large number of data sets in a short period of time.
(3) The resulting precision of retention/migration time estimation should be significantly improved in comparison with that initially provided by the instrumentation.

After selection and application of the best set of preprocessing methods the principal component analysis was performed and the results compared with those of PCA of the original data (before preprocessing). All the calculations at the preprocessing of data and chemometric evaluation of the obtained results were done in Matlab 6.5. Algorithms of the COW, denoising by Savitzky–Golay implementation of polynomial least-squares filters and the Cooley–Tukey algorithm for the fast Fourier transform were from the Department of Analytical Chemistry and Pharmaceutical Technology, Vrije

[1] M.J. Markuszewski, E. Szymanska, K. Bodzioch, R. Kaliszan, M. Markuszewski, K. Krajka, Metabonomic analysis of nucleosides and modified nucleosides in urine of healthy and cancer patients, in: 16th International Symposium on Pharmaceutical and Bioanalytical Analysis, Baltimore, MA, USA, 2005, p. 57 (abstract book); E. Szymanska, M.J. Markuszewski, K. Bodzioch, R. Kaliszan, M. Markuszewski, K. Krajka, Combined use of separation data and bioinformatics in metabolic analysis of urological clinical patterns, in: 11th International Symposium on Separation Sciences, Pardubice, Czech Republic, 2005, pp. 62–63 (abstract book).

University, Brussels, Belgium. Algorithm of baseline correction was from the Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gdańsk, Poland. Routines for DTW and discrete wavelet transform were as freely available at the websites: http://www.models.kvl.dk and http://www.jstatsoft.org/v06/i06/codes/ThreshWaveb.m.

3. Results and discussion

The main goal of pretreatment of first-order data is to reduce the time and cost of processing of complex chemical data, increase data quality and, most important, increase objectivity of the information extracted from them. The choice of preprocessing methods is highly dependent on the type of data. In present research, electrophoretic data were considered. These data exhibit specific features, such as sharp signals and discontinuities within a wide range of both time and frequency domains. That makes them a special object of preprocessing methods. Many chemometric tools usually applied in processing of chromatographic data do not necessarily work properly with electrophoretic data. At the current state of the art, the knowledge on successful application of bioinformatic tools in preprocessing of electrophoretic data is highly limited.

The experimental dataset, obtained as described in Section 2 underwent the following processes: baseline correction, denoising, selection of target sample, optimization of COW/DTW parameters, alignment of the whole data set, normalization of obtained results by known creatinine concentrations and, finally, PCA analysis.

3.1. Baseline correction and denoising of data

At first step various baseline correction and denoising algorithms were applied. Baseline correction provides flatter baselines and averages the baseline to zero. This improves the accuracy of integrals, the appearance of the signal and the quality of a result when subtracting one electropherogram from another. In our case, baseline correction was obtained by simply finding the minimum points in data points window (in this study 400 data points) for all possible window placements, considering them as new baseline and linear interpolation of other points [24].

Denoising separates the desired part of signal (correlated with the properties of analyzed sample) from unwanted part of signal (noise), what makes the further processing steps, such as warping, more effective. Three different algorithms, normally used in processing of chromatographic data, were implemented: SG smoothing, FFT of Cooley–Tukey algorithm and discrete wavelet transform [22,23].

The employment of the two first methods was unsuccessful because, besides limitation of noise, also shapes of electrophoretic peaks were changed. This fact could be related to inaccuracy of sine and cosine basis and low-order polynomial functions to process electrophoretic data functions used by fast Fourier transform and Savitzky–Golay smoothing, respectively.

Then, discrete wavelet transform methods, which are recommended for pretreatment of electrophoretic data by Perrin et al. [23], were implemented. Soft thresholding with sigma value 3 (standard deviation of additive Gaussian White Noise [23]) occurred to be accurate in our case. However, limitation of noise by the developed method appeared to have negligible impact on proper quantification of data and presentation of data in PCA (the percent of explained variance by three first principal components and location of points were very similar). Therefore, denoising step could be omitted in this case.

3.2. Alignment of data

The aim of various alignment (warping) methods is to correct or eliminate the shift in discrete data signals in such a way that the output data could be directly used by other appropriate chemometric tools for visualization and data mining. There are several methods applied in signal aligning which base on different similarity criterions and disparate ways of creation of aligned signals [22–34].

DTW and COW, which are used in our studies, are alignment methods that seem to work for broad ranges of signals [24–26]. COW is a special case of DTW and the main differences between these two methods are presented in Table 1.

Dynamic time warping nonlinearly warps the two trajectories in such a way that similar events are aligned and minimum distance between them is obtained [25,26]. In DTW several constraints should be specified to avoid excessive corrections, which could occur in the simplest implementations. Types and

Table 1
The main differences between dynamic time warping (DTW) and correlation optimized warping (COW) [24,25]

Method	Dynamic time warping (DTW)	Correlation optimized warping (COW)
Way of alignment	Sequence of elementary transitions going from one end point to other	Piecewise linear stretching and compression
Similarity criterion	Squared Euclidean distance between target profile and aligned profile	Correlation coefficient between corresponding part in target profile and aligned profile
Reconstruction of profiles	Linear interpolation or averaging	Linear interpolation
Parameters to define	Local continuity constraints; band-constraints	$N^{\dagger}$, $t^{\ddagger}$

† Number of points in each segment of aligned sample.
‡ Slack parameter: maximum length increase or decrease in this segment.

values of selected constraints and synchronization steps predetermine the results of warping and its quality. Local continuity constraints (called "rules") define the corrective function of the whole algorithm and are collected in the lookup table ($T^{(n,m)}$). In the lookup table $T^{(n,m)}$ n is the largest block distance covered by any of the rules in the table and m is the maximum number of horizontal/vertical consecutive transitions allowed by the table [26]. The second type of constraints (band-constraints) is responsible for maximum compression/expansion in time points of the sample and reference to their original lengths. In synchronization step, all the points in optimal path can be used to obtain the warped signal (symmetric synchronization) or some points can be aligned with the same point of the reference signal (asymmetric synchronization).

In comparison to other preprocessing methods, COW is known as less flexible and more "peak shape preserving" and easy to employ by simple optimization of the parameters settings (only two parameters have to be set) [24–26]. In this method also dynamic programming is used but solution space of optimization is restricted to only two parameters: the number of points in each segment of the aligned sample N and the maximum length increase or decrease in this segment t-slack parameter. After tracking back the optimal path and setting all borders of segments in right positions, the warped/aligned sample segments are reconstructed by linear interpolation and they are the best-matching signals for the predefined set of parameters N and t.

After selection of denoising and baseline correction tools, profile #4 (see Table 2) was selected as the target of alignment as the most representative one (it contains all peaks present in other samples and its peaks are situated near the center of the distributions of the corresponding peaks from other profiles). The main object of alignment, the shift, was characterized in Table 2. It could be seen that maximal peak shift was about 200 data points for that dataset, with mean value of 48 data points for one of the last peaks in the aligned section. That strictly corresponds to 1.78% relative standard deviation of migration time and means an acceptable reproducibility in analysis of biological samples. However, such a shift is approximately three times higher than that reported for chromatographic data, which were processed by various warping techniques [23–34]. Its reduction may be treated as a chemometric challenge.

The DTW and COW algorithms with various parameters were implemented and the results are presented in Fig. 2. The best results of COW were obtained with application of $N = 160$, as number of sections, and $t = 4$, as slack parameter (Fig. 2D). The length of section was then set equal to 12 points and time for calculation for each electrophoretic profile was approximately 50 s. The quality of this alignment was high for most of the profiles, resulting in correlation coefficients above 0.9. Only in few cases, there was the need to use other sets of parameters (actually, the variation between the target profile and the profile to be aligned was higher than 50 data points for four profiles). The division of profile in 12 points pieces and alignment of every piece by stretching and compression could cause marked deformation of data such as deformation of peaks shape and area under peaks. However, in our dataset, where peaks are narrower than standard peaks in chromatograms (10–20 data points' width), no strong deformation of that type was observed.

In the development of DTW method more attention was paid to the application of two types of that algorithm: dynamic time warping with slope constraints and dynamic time warping with COW-like constraints. Limitation of flexibility of dynamic time warping was essential because unconstrained DTW produced artifacts and altered the shape of peaks present in this data. Dynamic time warping gave comparable results to those of COW only when rigid constraints were used. Warping with lookup table $T^{(20,4)}$ (i.e., with rules spanning also 12 data points and 4 points as maximal number of consecutive horizontal or vertical transitions in the warping path) was applied with 10% band constraints. In the synchronization step, interpolation or averaging

Table 2
Characterization of shift in dataset

Peaks[†]	Target profile		Unaligned profiles			Aligned profiles	
	Migration time (min)	Number of data points	Shift range (data points)[‡]	Mean shift (data points)	R.S.D. (%)	Shift range (data points)[‡]	R.S.D. (%)
1	10.07	2416	−46 to +83	±15	1.05	−2 to +1	0.029
2	10.53	2528	−63 to +41	±19	1.00	−5 to +6	0.086
3	10.65	2557	−93 to +34	±21	0.57	−3 to +1	0.032
4	11.37	2728	−107 to +37	±23	1.12	−3 to +2	0.031
5	11.74	2817	−41 to +115	±17	1.18	−2 to +1	0.02
6	11.92	2861	−119 to +47	±25	1.20	−3 to +6	0.028
*	12.28	2948	−130 to +50	±27	1.24	−1 to +1	0.021
7	12.40	2976	−130 to +50	±28	1.26	−2 to +2	0.024
8	12.85	3083	−141 to +65	±30	1.39	−3 to +7	0.069
*	14.86	3767	−191 to +59	±41	1.49	−1 to +3	0.023
9	14.95	3590	−92 to +198	±35	1.60	−4 to +4	0.024
10	15.55	3732	−200 to +76	±48	1.78	−3 to +3	0.033

The shift was compared for 12 electrophoretic peaks (10 identified and 2 unidentified) by calculation of variation in number of data points and variation in migration time according to target profile (profile #4) before and after alignment.

[†] Peak numbers: 1, pseudouridine; 2, uridine; 3, cytidine; 4, 5-methyluridine; 5, inosine; 6, N^4-acetylcytidine; 7, guanosine; 8, adenosine; 9, N^2,N^2-dimethylguanosine, 10, xanthosine; (*), unidentified peaks.

[‡] The negative sign (−) refers to a forward shift and a positive sign (+) refers to a backward shift as compared to profile #4 (target profile).

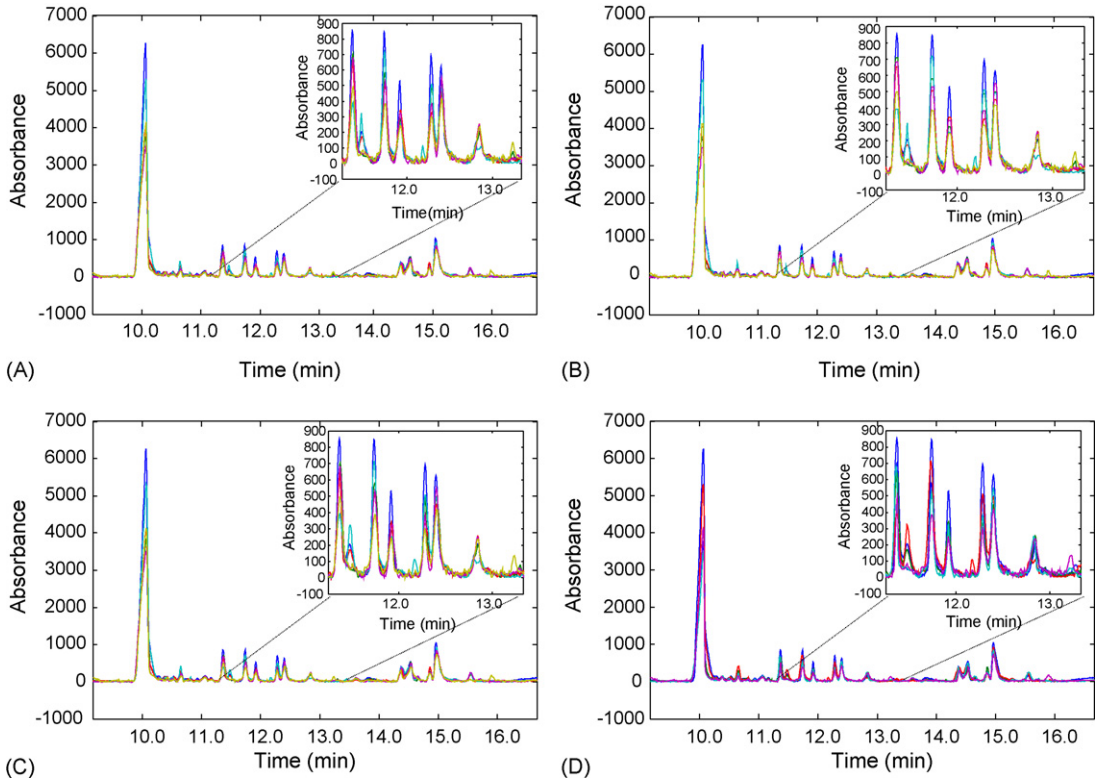

Fig. 2. Six example electrophoretic profiles after warping with different DTW and COW algorithms. A–C, after DMW warping $T^{(20,4)}$ with interpolation (A), averaging (B), defining endpoints (C); D, after COW warping $N = 160$, $t = 4$. Electrophoretic conditions as in Fig. 1.

was employed (Fig. 2A and B). Additionally, further restrictions to dynamic time warping constraints, by defining end points of rules in DTW, were applied and the algorithm analogous to COW with only difference in optimization criterion was formulated (Fig. 2C).

Consequently to applied DTW parameters, reconstructions of aligned profiles were different. Averaging in synchronization step caused disturbances in peak height. That means that when the sample peak is larger than the matching in target profile, the former is cut. Conversely, when the sample peak is smaller, its top element is repeated until two sides of the synchronized peak match those on reference, resulting in plateau (Fig. 2B). No such artifacts could be noticed after employment of interpolation instead of averaging (Fig. 2A and C). The results of the presented DTW warping are quite similar to those obtained by COW with small changes in peak height (Fig. 2B) and width (especially width of the second peak at closer view of aligned data at Fig. 2A and B). Effects of DTW warping with defined endpoints and interpolation as synchronization are identical to that obtained by COW, in spite of different optimization criteria (Euclidean distance in DTW and correlation coefficient in COW).

Calculation time for all the applied alignment methods was different. It was, accordingly, 105, 110 and 97 s per one profile for cases A–C. That is about two times more than time of the corresponding correlation optimized warping. On that basis, COW with $N = 160$ and $t = 4$ was selected for further preprocessing of dataset.

Now, the obtained derivative of data was normalized by creatinine concentration which has been known for each sample, because concentrations of all the considered analytes are in linear relationship with the concentration of urine expressed by the value of extracted creatinine [18].

3.3. Principial component analysis

Finally, PCA on transformed and normalized by creatinine concentration nucleoside profiles was performed to reveal the structure of data and evaluate the differences between two groups of profiles: healthy controls and cancer patients. Additionally, PCA analysis of peak areas obtained by standard integration of 17 peaks (14 peaks presented in Fig. 1 and 3 different peaks present in the part of electropherograms cut before preprocessing) from original electropherograms was performed for the sake of comparison. The principal components are displayed as a set of scores, which highlights the clustering and outliers. PCA scatterplots (PC1 versus PC2 and PC1 versus PC3) are presented in Fig. 3. Large values (>80%) of explained variance in the first three principal components proves the usefulness of PCA in mining the information contained in electrophoretic profiles. The division between the groups of healthy and cancer patients can be clearly noticed on the scatter plots obtained after preprocessing of electrophoretic data (Fig. 3 A and B) although it is also seen after standard integration and calculation of peaks area before the alignment (Fig. 3C and D). PCA analysis of preprocessed data (Fig. 3 A and B) is evidently more informative (the percent of explained variance is higher and structure of data is more clear in that case), even though it is performed on shorter electrophoretic profiles and comprises not all the data

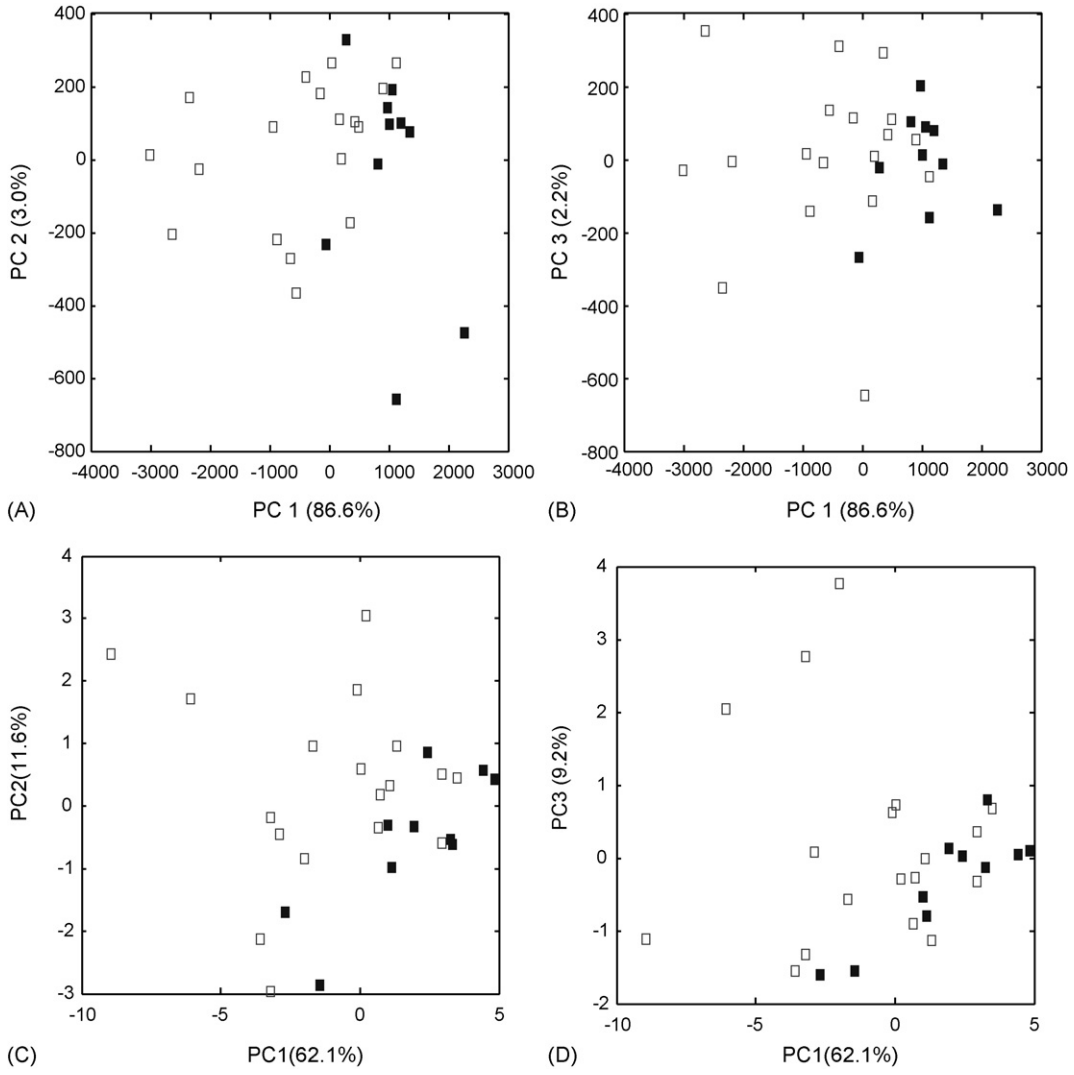

Fig. 3. PCA scores scatter plots ((□) cancer patients; (■) healthy patients). Principal component analysis was performed on preprocessed data (baseline corrected and aligned) by direct PCA evaluation (PC1 vs. PC2 – Fig. 3A; PC1 vs. PC3 – Fig. 3B) and on original data (before the alignment) by integration, calculation, autoscaling and PCA of area of 17 peaks (identified and unidentified peaks from Fig. 1A) (PC1 vs. PC2 – Fig. 3C; PC1 vs. PC3 – Fig. 3D).

collected during the capillary electrophoretic analysis. After data preprocessing (Fig. 3A and B) the data discrimination significantly increased and first principal component, PC1, accounted for 86.6% of data variance, the second principal component, PC2, for 3.0% of data variance, and the third principal component, PC3, for 2.2%. In case of PC analysis performed on original electrophoretic data before preprocessing (Fig. 3C and D) the data discrimination expressed by PC1, PC2 and PC3 accounts for 62.1, 11.6 and 9.2% of data variance, respectively.

That it is an encouraging preliminary finding from the point of view of explanatory potential of metabonomic studies. It appears that application of preprocessing tools not only helps to limit the capillary electrophoretic method disturbances but also to reveal some information present in electrophoretic data, uncovered by standard peak integration. This information might be connected with small peaks and peak's shapes not adequately taken into consideration in the procedure involving peak integration and area calculation. Furthermore, application of proper chemometric tools seems to help to reduce uncertainty and subjectivity introduced by peak quantification and makes possible simultaneous analysis of complex mixtures by such tools like PCA. Preprocessing of data could be a valuable alternative to peak integration in multivariate data analysis applied in metabonomic studies. It could limit shortages of the developed methods and influence of the factors connected with analyzed samples and the equipment used. Thereby, it should help to validate potential markers by a reliable and fast comparison of numerous profiles obtained in the large-scale long-term projects.

4. Conclusions

In the study different preprocessing methods were implemented to evaluate electrophoretic data profiles of urinary nucleosides. The electrophoretic data were collected by analysis of urine samples from healthy and cancer patients and were assumed to contain information about physiological state of the subjects. The aim of the study was to adapt modern data preprocessing methods before further bioinformatic analysis

and to compare this strategy with the standard procedure of peak integration and area calculation.

The best results of preprocessing were achieved after application of the COW with the aligned data analyzed by PCA. The PCA of electrophoretic peak areas of original data (without preprocessing) was also performed to compare the results with those obtained for the preprocessed data. The structure of data provided by PCA was different for the preprocessed data and the original data. Preprocessing not only limited the shift in data but also revealed information hidden in the nucleoside profiles which is probably connected with the shape of peaks and such structures of profiles that could not be identified when only areas under peaks were considered. Furthermore, preprocessing of electropherograms seems to be a more objective and less time consuming procedure than peak integration.

Selection of suitable preprocessing methods appears to be a very important step in bioanalytical data evaluation. Certainly, more experience in application of chemometric tools in metabonomics will help not only to explore the systematic information dispersed over numerous data but it should also enhance the reliability of the conclusions drawn and limit the necessary analytical work.

Preliminary results obtained for a limited number of healthy and cancer subjects reveal promising means to separate the two groups based on urinary analytical profiles processed with modern cheminformatics tools. That may open a new pathway to convenient diagnostics.

Acknowledgement

The project was supported by the Komitet Badań Naukowych, Warsaw, Poland (grant KBN No. 3PO5F02724).

References

[1] N.M. Luscombe, D. Greenbaum, M. Gerstein, Methods. Inform. Med. 40 (2001) 346–358.
[2] P.A. Whittaker, Trends Pharm. Sci. 24 (2003) 434–439.
[3] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Trends Biotechnol. 22 (2005) 245–252.
[4] S. Terabe, M.J. Markuszewski, N. Inoue, K. Otsuka, T. Nishioka, Pure Appl. Chem. 73 (2001) 1563–1572.
[5] M.J. Markuszewski, K. Otsuka, S. Terabe, K. Matsuda, T. Nishioka, J. Chromatogr. A 1010 (2003) 113–121.
[6] J. van der Greef, P. Strobant, R. van der Heijden, Curr. Opin. Chem. Biol. 8 (2004) 559–565.
[7] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181–1189.
[8] M.J. Markuszewski, E. Szymańska, R. Kaliszan, in: Z. Brzózka (Ed.), Miniaturyzacja w analityce, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2005, pp. 219–235.
[9] U. Manne, R.-G. Srivastava, S. Srivastava, Drug Discov. Today 14 (2005) 965–976.
[10] D.B. Kell, Biochem. Soc. Trans. 33 (2005) 520–524.
[11] M. Mulchtar, Drug Discov. Today 10 (2005) 1216–1218.
[12] K.D. Altria, D. Elder, J. Chromatogr. A 1023 (2003) 1–14.
[13] J.R. Petersen, A.O. Okorodudu, A. Mohammad, D.A. Payne, Clin. Chim. Acta 330 (2003) 1–30.
[14] B. Casado, C. Zanone, L. Annovazzi, P. Iadarola, G. Whalen, J.N. Baraniuk, J. Chromatogr. B 814 (2005) 43–51.
[15] C. Guillo, D. Barlow, D. Perret, M. Hanna-Brown, J. Chromatogr. A 1027 (2004) 203–212.
[16] Y. Ma, G. Liu, M. Du, I. Stayton, Electrophoresis 25 (2004) 1473–1484.
[17] P. Iadorola, G. Cetta, M. Luisetti, L. Annovazzi, B. Casado1, J. Baraniuk, C. Zanone, S. Viglio, Electrophoresis 26 (2005) 752–766.
[18] G. Xu, H.M. Liebich, Am. Clin. Lab. 3 (2003) 22–32.
[19] J. Yang, G. Xu, Y. Zheng, H. Kong, C. Wang, X. Zhao, T. Pang, J. Chromatogr. A 1084 (2005) 214–221.
[20] S. La, J. Cho, J.-H. Kim, K.-R. Kim, Anal. Chim. Acta 486 (2003) 171–182.
[21] E. Szymańska, M.J. Markuszewski, X. Capron, A.-M. van Nederkassel, Y. Vander Heyden, M. Markuszewski, K. Krajka, R. Kaliszan, submitted for publication.
[22] P.D. Wentzel, C.D. Brown, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd., Chichester, 2000, pp. 9764–9800.
[23] C. Perrin, B. Walczak, D.L. Massart, Anal. Chem. 73 (2001) 4903–4917.
[24] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17–35.
[25] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77–92.
[26] G. Tomasi, F. van der Berg, C. Anderson, J. Chemometr. 18 (2004) 231–241.
[27] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1096 (2005) 177–186.
[28] A.M. van Nederkassel, V. Vijverman, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1085 (2005) 230–239.
[29] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141–155.
[30] J.H. Christansen, J. Mortensen, A.B. Hansen, O. Andersen, J. Chromatogr. A 1062 (2005) 113–123.
[31] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, J. Chromatogr. A 961 (2002) 237–244.
[32] F. Gong, Y.-Z. Liang, Y.-S. Fung, F.-T. Chau, J. Chromatogr. A 1029 (2004) 173–183.
[33] A.-J. Lau, B.-H. Seo, S.-O. Woo, H.-L. Koh, J. Chromatogr. A 1057 (2004) 141–149.
[34] P.H.C. Eilers, Anal. Chem. 76 (2004) 404–411.