Biomarkers, June 2008; 13(4): 435–449

informa
healthcare

Metabonomics in cancer diagnosis: mass spectrometry-based profiling of urinary nucleosides from breast cancer patients

ANTJE FRICKENSCHMIDT[1], HOLGER FRÖHLICH[2], DINO BULLINGER[1], ANDREAS ZELL[2], STEFAN LAUFER[3], CHRISTOPH H. GLEITER[1], HARTMUT LIEBICH[4], & BERND KAMMERER[1]*

[1]*Department of Pharmacology and Toxicology, Division of Clinical Pharmacology, University Hospital Tübingen, Otfried-Müller-Str. 45, D-72076 Tübingen, Germany,* [2]*Center for Bioinformatics Tübingen (ZBIT), Sand 1, D-72076 Tübingen, Germany,* [3]*Institute of Pharmacy, University of Tübingen, Auf der Morgenstelle 8, D-72076 Tübingen, Germany and* [4]*Medical Clinic, University Hospital Tübingen, Otfried-Müller-Str. 10, D-72076 Tübingen, Germany*

Abstract
Modified nucleosides are formed post-transcriptionally in RNA. In cancer disease, the cell turnover and thus RNA metabolism is increased, yielding higher concentrations of excreted modified nucleosides. In the presented study, urinary ribonucleosides were used to differentiate between breast cancer patients and healthy volunteers. The nucleosides were extracted from urine samples using affinity chromatography and subsequently analyzed via liquid chromatography ion trap mass spectrometry (LC-IT-MS). The peak areas were related to the internal standard isoguanosine and to the urinary creatinine level. For bioinformatic pattern recognition we used the support vector machine. We examined 113 urine samples from breast cancer patients (stage Tis-T4) and 99 control samples from healthy volunteers. We achieved a sensitivity of 87.67% and a specificity of 89.90% when including 31 nucleosides. The medical metabonomics concept based on the urinary nucleoside profile reveals a significantly improved classification compared with currently applied breast cancer biomarkers such as CA15-3.

Keywords: *Medical metabonomics, medical metabolomics, modified nucleosides, ion trap mass spectrometry, support vector machine, cancer diagnosis*

(Received 17 November 2007; accepted 22 February 2008)

Introduction

Metabonomics is a relatively new field in analytics, which has rapidly expanded in recent years, building on from the established genomics and proteomic concepts. In genomics, the genome including all DNA segments of an organism is studied. Proteomics deals with the proteins expressed by the genome, with particular emphasis

Correspondence: Bernd Kammerer, University Hospital Tübingen, Department of Clinical Pharmacology, Otfried-Müller-Str. 45, D-72076 Tübingen, Germany. Tel: +049-7071-29-72265. Fax: +049-7071-29-5035. E-mail: Bernd.Kammerer@uni-tuebingen.de

ISSN 1354-750X print/ISSN 1366-5804 online
DOI: 10.1080/13547500802012858

on structure and function. Metabonomics supplements these, by taking metabolic changes resulting from genomic or environmental alterations into account.

The importance of metabolite identification and the possibility of differentiating between physiologically variable groups (state) based on the metabolite pattern in biological fluids has been described by several groups (Plumb et al. 2002, Gamache et al. 2004, Griffin 2004, Kim et al. 2004, Lindon et al. 2004, Wilson et al. 2005). The analytical data used were mainly obtained by nuclear magnetic resonance imaging (NMR) (Kleno et al. 2004, Beger 2005, Lenz et al. 2005) or mass spectrometry (MS) (Plumb et al. 2002, 2003, Wang et al. 2005, Wilson et al. 2005, Yang et al. 2005). Metabonomics has been applied to several areas including drug development (Plumb et al. 2002), biomarker elucidation for disease diagnosis (Gamache et al. 2004, Kleno et al. 2004, Yang et al. 2005), nutrition research (Kim et al. 2004) and examination of metabolic pathways in plants (Roessner et al. 2002). The most commonly used statistical method for classification in these studies was principal component analysis.

Yang *et al.* (2004) applied principal component analysis to 113 mostly unidentified peaks appearing in HPLC-UV-chromatograms, comprising all cis-diol structures extracted from urine samples of hepatitis and hepatocirrhosis patients. In a more recent publication, they expanded the method by coupling with tandem MS (Yang et al. 2005). In both cases, the integrated peak areas were compared to both an internal standard and the creatinine level in urine prior to principal component analysis.

In our study, we analyzed nucleosides in urine samples from stage Tis to T4 breast cancer patients and healthy volunteers. Most of the known modified nucleosides are formed post-transcriptionally in RNA catalyzed by different enzymes (Bjoerk et al. 1987). They are formed during RNA metabolism and excreted in urine as end products due to a lack of specific phosphorylases. In healthy adults, the nucleoside excretion is constant, while in children up to the age of 16, the excretion rates are higher (Sander et al. 1986, Itoh et al. 1993, Prankel et al. 1995, Liebich et al. 1997). As the RNA turnover seems to be impaired in cancer patients, modified nucleosides have been evaluated as possible tumour markers (Sasco et al. 1996, Liebich et al. 1997, Xu et al. 2000, Dieterle et al. 2003). Methylated nucleosides in particular play an important role, e.g. the levels of 1-methylinosine and pseudouridine are higher in urine from breast cancer patients (Tormey et al. 1980). In recent studies, Dudley et al. (2003) identified the rare nucleoside 5′-deoxycytidine as potential urinary biomarker for head and neck cancer. A number of modified nucleosides occurring in urine are shown in Figure 1.

We determined the nucleosides using liquid chromatography/electrospray ionization ion trap mass spectrometry (LC-ESI-IT-MS) after extraction by affinity chromatography. Only cis-diols were extracted using phenylboronic acid gel, which binds cis-diols reversibly. The separation was performed using reversed phase liquid chromatography, based on a method developed by Liebich et al. (1997) which was adapted to make it compatible with mass spectrometric detection. We used a Bruker Esquire High Capacity Ion Trap (HCT-IT MS) in positive ionization mode with the mass range 50–500 Da for detection. The identification of the compounds of interest had been performed in previous studies using an auto-LC-MS^3 method to identify the nucleosides not only according to retention time, but also according to their characteristic fragmentation pattern (Kammerer et al. 2005b). Further structural

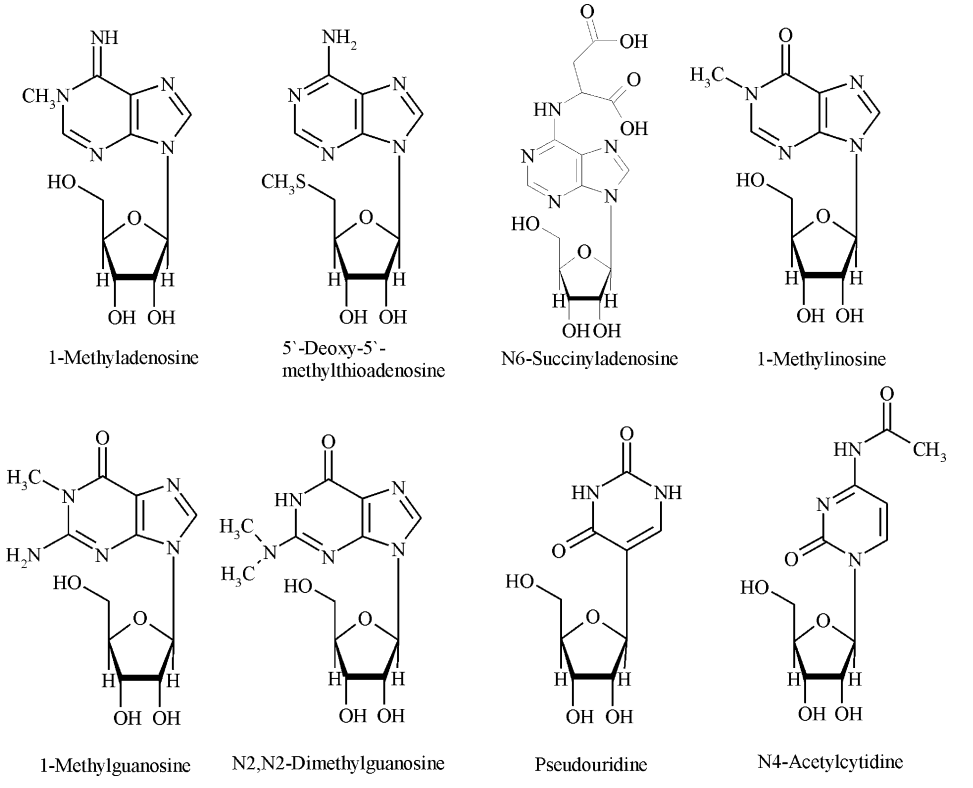

Figure 1. Structures of some modified nucleosides occurring in urine.

elucidation of nucleosides is also possible using matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS (Kammerer et al. 2005a) and by measuring accurate masses using LC-ESI-oa-TOF MS (oai orthogonal acceleration) (Bullinger et al. 2005).

The aim of our study was to classify 212 urinary samples from patients with breast cancer and control samples from healthy volunteers on the basis of the LC-MS-results. In preliminary studies, we identified 37 compounds in urine samples of breast cancer patients as nucleosides/ribose containing compounds (Kammerer et al. 2005b). Thirty-one of these compounds were considered suitable for statistical interpretation. The peak areas of the extracted ion chromatograms (EIC) of all compounds were compared with the internal standard isoguanosine and to the creatinine level for standardization.

The final classification of the 212 samples into breast cancer patients and control samples was evaluated by means of a leave-one-out cross-validation procedure (Duda et al. 2001). The underlying statistical model was a support vector machine (SVM) (Cortes & Vapnik 1995, Schölkopf & Smola 2002) trained with an RBF-kernel function to allow non-linear class separations. SVMs are learning algorithms capable of handling large numbers of input vectors (Cortes & Vapnik 1995, Schölkopf & Smola 2002, Guyon & Elisseeff 2003). They have been applied to various domains including characterization of molecules (Holloway et al. 2005, Saeh et al. 2005, Lepp et al. 2006) as well as microarray expression (Furey et al. 2000) and proteomics

(Honda et al. 2005, Liu 2006) in conjunction with disease diagnosis. They are generally used to generate a classification based on a varying number of descriptors. SVMs have previously been used for cancer diagnosis based on mass spectrometric data (Duan & Rajapakse 2005, Rajapakse et al. 2005).

In our study we achieved a cross-validated sensitivity of 87.67% and a specificity of 89.90% when including all 31 nucleosides. A further automatic selection of the most relevant compounds yielded a sensitivity of 89.39% and a specificity of 88.89%.

Methods

Chemicals and materials

Methanol was of HPLC hypergrade and formic acid of analytical grade (both from Merck, Darmstadt, Germany). Double distilled water from an in-house distillation system was used for chromatography and preparation of solutions. The nucleosides used as reference standards for HPLC separation were 5,6-dihydrouridine (DHU), pseudouridine (Ψ), cytidine (C), uridine (U), adenosine (A), 1-methyladenosine (m^1A), N^6-methyladenosine (m^6A), 5′-deoxy-5′-methylthioadenosine (MTA), inosine (I), 1-methylinosine (m^1I), guanosine (G), 1-methylguanosine (m^1G), N^2-methylguanosine (m^2G), N^2_2-dimethylguanosine (m^2_2G), 3-methyluridine (m^3U), 5-methyluridine (m^5U), xanthosine (X), N^4-acetylcytidine (ac^4C), N^6-threonylcarbamoyladenosine (t^6A), N^2,N^2,7-trimethylguanosine ($m^{2,2,7}G$) and adenylosuccinic acid (sodium salt). All nucleosides were from Sigma (Taufkirchen, Germany) except m^2_2G, $m^{2,2,7}G$ and t^6A which were obtained from Biolog (Bremen, Germany). The internal standard isoguanosine was kindly donated by Prof. J.H. Kim of Seoul University, South Korea. Affigel 601 was purchased from Biorad (Bremen, Germany).

Instrumentation

The LC equipment consisted of an HPLC 1100 system (Agilent, Waldbronn, Germany) including a vacuum degasser (G 1379 A), a binary pump (G 1376 A), a thermostatted autosampler (G 1330 A and G 1313 A), a column oven (G 1316 A) and DAD detector (G 1315 B). The nucleosides were detected by a Bruker Esquire HCT Ion Trap mass spectrometer equipped with an electrospray (ESI) source. The system was controlled and data were collected by the Bruker software Esquire Control 5.1 and Data Analysis 3.1.

The separation of the nucleosides was performed on a LiChroCART Superspher 100 RP-18 endcapped column (150 × 2.0 mm, 4 μm; Merck, Darmstadt, Germany) with LiChroCART Superspher 100 RP-18 precolumn (1 × 2.0 mm, 4 μm; Merck).

High-resolution FTICR MS spectra of semipreparatively isolated nucleosides were measured in positive ionization mode using an APEX II ESI-FTICR MS mass spectrometer equipped with a 4.7 T magnet (Bruker Daltonics, Bremen, Germany). Samples were introduced via syringe pump (Cole Parmer, Vernon Hills, IL, USA) infusion at a flow rate of 80 μl h^{-1} via a 60° off-axis grounded capillary sprayer needle (Analytica of Branford Inc.). The capillary exit voltage was adjusted to 15–25 V. For internal and external calibration, a homologous series of polyethyleneglycols (PEG 400) was used. For data acquisition and post-processing, the XMASS version 5.0.10 software (Bruker Daltonics) was applied. Calculating the corresponding molecular formulae, we were able to identify the metabolites N^2,N^2,7-trimethylguanosine

($C_{13}H_{20}N_5O_5$ $[M+H^+]$, ppm error: 0.527) and N^6-succinyladenosine ($C_{14}H_{18}N_5O_8$ $[M+H^+]$, ppm error: 0.784).

Samples

We examined 113 urine samples of female breast cancer patients with different stages of cancer (Tis to T4) and 99 samples of female healthy volunteers. All samples were randomly collected at the Frauenklinik, University Hospital Tübingen and stored at -80°C until extraction. The urine samples were taken and examined in agreement with the participants. The clinical trial has been approved by the local ethics committee of University Hospital Tübingen. The breast cancer patients included in this study were aged 29–87 years, the healthy control subjects were aged 16–78 years. All samples were taken preoperatively with no neoadjuvant hormonal, irradiation or chemotherapies applied.

Sample preparation

The nucleosides were extracted from urine samples using phenylboronic acid gel. The isolation procedure was carried out according to Liebich et al. (1997). The urine samples were alkalinized to pH 8.5 with 2.5% ammonia solution and centrifuged at 2400 g for 10 min. Subsequently, 10 ml of the obtained supernatant was mixed with 0.5 ml internal standard solution (0.25 M isoguanosine in water) and then put on the column (500 mg affigel 601, column dimensions: 150×15 mm), preconditioned with 45 ml ammonium acetate solution (pH 8.5, 0.25 M). After washing with 20 ml ammonium acetate solution (pH 8.5, 0.25 M) and 6 ml methanol-water (1:1, v/v), elution was carried out with 25 ml 0.1 M formic acid in methanol-water (1:1, v/v). Afterwards, the column was washed with 25 ml 0.1 M formic acid in methanol-water (1:1, v/v) and 25 ml methanol-water (1:1, v/v) and then reconditioned with 45 ml ammonium acetate solution (pH 8.5, 0.25 M) for the next sample. The collected elution solvent was removed using a rotary evaporator and the residue was dissolved again in 0.5 ml of 25 mM KH_2PO_4. The creatinine level, which the nucleoside areas were related to, was determined by a modified Jaffé method (Bartels et al. 1975).

Chromatographic conditions

Fifteen microlitres of the extracted samples were injected into the HPLC system. The separation of the nucleosides was performed using a gradient as described in Table I.

MS parameter optimization

The mass spectrometric parameters were tuned for optimization of the sensitivity in positive electrospray mode (see Table I). This was done by direct infusion of solutions of 16 standard nucleosides (10 μg ml^{-1}). As the standard nucleosides showed differences in their optima for skimmer voltage, capillary exit voltage, octupole 1 and 2 voltages, trap drive and octupole RF amplitude, we optimized these parameters separately for all nucleosides and set them to an average value to find a compromise for all nucleosides, especially considering those occurring in lower concentrations in urine.

Table I. LC and MS settings.

LC conditions	
Solvent system	
A	5 mM ammonium formate buffer, pH 5.0
B	methanol/water (3/2, v/v) + 0.1% formic acid
Gradient system	
0 min	1% B
15 min	15% B
40 min	60% B
50 min	1% B
55 min	1% B
Equilibration time between runs	10 min, 1% B
Flow rate	125 µl min^{-1}
Run time	55 min
MS settings	
Mode	Positive
Scan range	50–500 Da
Capillary voltage	5 kV
End plate offset voltage	−500 V
Nebulizer gas (nitrogen)	35 psi
Dry gas (nitrogen)	7.0 l min^{-1}
Dry temperature	325°C
Capillary exit voltage	75 V
Skimmer voltage	15 V
Octopole 1	7.5 V
Octupole 2	1.7 V
Octupole RF amplitude	110 Vpp
Trap drive	35.7

The skimmer voltage was set to a value at which the signal intensity was maximal while the nucleoside fragmentation was extensively prevented.

A multiple reaction monitoring method would have been more accurate and selective, but impossible to realize with such a high number of nucleosides, as the number of compounds which may be selected is limited to 10.

As we wanted to include compounds which are very likely nucleosides, but as we were not able to identify them, we could not perform a standard quantification due to the missing standard substances. Thus, we determined values which were related to an internal standard as well as to the urinary creatinine level.

Integration procedure

We used extracted ion chromatograms (EIC) of the corresponding masses for manual integration, using Bruker DataAnalysis 3.1. The EICs were processed with a Gauss function smoothing algorithm contained in DataAnalysis software. For analytical and physiological normalization, the integrated peaks areas of the analyzed metabolites were related to the peak area of the internal standard isoguanosine and the determined urinary creatinine level (in µmol ml^{-1}) prior to the bioinformatical analysis. The creatinine level is a constant parameter which nucleoside concentrations in urine are usually related to for comparing different samples (Bartels et al. 1975, Liebich et al. 1997, Liebich et al. 2005)

Learning the class separation between healthy and cancer patients with support vector machines

The basic concept of a SVM classification is depicted in Figure 2. The idea is to separate two classes of points (in our case nucleoside concentration profiles of healthy and breast cancer patients) by means of an optimally separating hyperplane. This hyperplane is constructed such that the margin (i.e. the distance of the hyperplane to the point closest to it) is maximized. Non-linear classifications can be performed by mapping the data in a non-linear fashion into some higher dimensional space and then calculating the optimal hyperplane there.

The mapping is usually achieved by means of so-called kernel functions, which are a special class of similarity measures. A kernel function that is often employed is the radial basis function kernel (RBF-kernel):

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \tag{1}$$

In contrast to techniques like principal component analysis (PCA), which only aim to reduce the dimensionality of the data, SVMs construct a decision function of the form

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{2}$$

where the α_i are so-called Lagrangian multipliers and b a bias term. Both are learned from a set of n training data points $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \Re^d \times \{+1, -1\}$. Hereby $\mathbf{x}_i$ represents the ith input vector, which is labelled with $y_i \in \{+1, -1\}$ to characterize its class membership. SVMs are predictive models, i.e. given an unclassified vector $\mathbf{x}$ the decision function (2) can be asked to forecast its class membership.

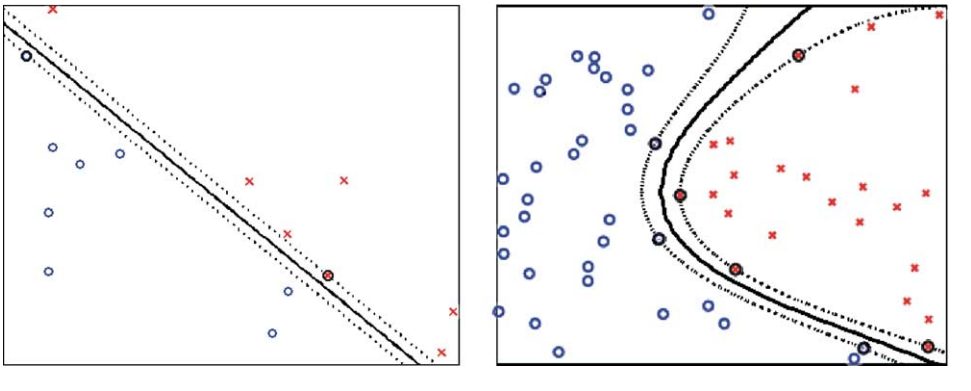

Figure 2. (A) Example of a support vector machine (SVM) classification between crosses and circles by means of an optimal separating hyperplane (sold black line). The hyperplane only relies on the points lying on the margin (dashed lines). These are the so-called support vectors (marked by extra circles). (B) Example of a non-linear SVM classification between crosses and circles.

Results and discussion

Evaluation of the analytical method

Proof of reproducibility. A standard solution containing 17 nucleosides and the internal standard was injected 11 times. These values were normalized to the internal standard isoguanosine and the average and standard deviation were calculated.

The standard deviation for the nucleosides was between 1.7 and 8.5% except for guanosine (10.2%), which was considered acceptable for further analysis.

Additionally, a urine sample of a healthy volunteer extracted as described above was injected five times and the deviation was calculated for the above mentioned 17 standard nucleosides as well as 19 further compounds which were already identified as nucleosides or ribosyl derivatives (Kammerer et al. 2005b). The deviations were between 1.7 and 8.4% and thus comparable to those determined for the standard solution.

Proof of linearity. For proof of linearity, a urine sample was extracted as described above and diluted 1:2, 1:5, 1:10, 1:50 and 1:100. These solutions were mixed 1/1 with the internal standard solution and injected in the HPLC system. The integrated peak areas of the EICs were related to the internal standard and the results processed by linear regression. Only five nucleosides did not show linearity. These five compounds (cytidine, putative ribosyl derivatives with *m/z* 255, *m/z* 258, uridine and 5-methyluridine) were excluded from further analysis, yielding 32 compounds to be examined. The regression coefficients of these compounds were between 0.965 and 0.999.

The ribosyl derivative with a mass of 300 Da was excluded later on because of a bad peak form in the chromatograms of some samples. In the end, 31 nucleosides/ribosyl derivatives were included in the evaluation using a support vector machine. For the sake of simplicity, in the following they are identified with numbers 1,..., 31. The final assignment of the nucleosides number to its *m/z* value and name (if known) is given in Table II. A 3D image of the integrated EICs of one urine sample of a healthy volunteer is shown in Figure 3.

Prediction of cancer vs non-cancer with support vector machines

In the present study we evaluated how well a SVM classifier with a RBF-kernel function, could learn the discrimination of the 99 healthy volunteers from the 113 breast cancer patients based on the concentration profile of the 31 nucleosides measured in our experiments (see Table II). For this purpose we calculated the so called leave-one-out error (LOO), which is an almost unbiased estimator of the true generalization performance of a machine learning algorithm (Luntz & Brailovsky 1969). The LOO is computed by sequentially leaving out one of our 212 samples and training the model on the rest. Each time the model is asked to predict the class (healthy or not) of the left out patient. Within each round of the LOO procedure, we normalized all nucleoside concentrations to mean 0 and standard deviation 1 on the current training set and then applied the resulting scaling factors to the left out test sample. Furthermore, all necessary parameter tunings for the SVM were done only on the current training set (i.e. within the LOO procedure).

Table II. Assignment of nucleosides and other ribosyl derivatives to numbers used for evaluation (m/z); index ordered by increasing retention time.

	RT (min)	Mass $(M+H)^+$	Compound[a]
1	4.2	259	**1-Ribosylimidazole-4-acetic acid**
2	4.9	247	**5,6-Dihydrouridine**
3	5.3	245	**Pseudouridine**
4	8.2	228	*1-Ribosylpyridin-4-one*
5	8.7	346	*3-(3-Amino-3-carboxypropyl)-uridine*
6	15.9	271	1-Ribosylpyridin-2-one-5-carboxamide
7	19.9	269	**Inosine**
8	20.7	298	**7-Methylguanosine**
9	20.9	301	Ribosyl derivative 1
10	21.9	293	Ribosyl derivative 2
11	21.9	271	*1-Ribosylpyridin-3-one-4-carboxamide*
12	23.7	296	*1,N^6-Dimethyladenosine*
13	25.0	259	**3-Methyluridine**
14	25.3	285	**Xanthosine**
15	26.4	268	*2-Aminopurin-9-riboside*
16	27.4	351	*5-Methylaminomethyl-2-selenouridine*
17	26.9	384	**N^6-Succinyladenosine**
18	29.4	283	**1-Methylinosine**
19	31.1	298	**1-Methylguanosine**
20	32.5	286	**N^4-Acetylcytidine**
21	32.9	293	Ribosyl derivative 3
22	33.4	298	**2-Methylguanosine**
23	34.8	268	**Adenosine**
24	40.5	398	*2-Methylthio-N^6-(cishydroxyisopentenyl)-adenosine*
25	40.9	326	**N^2,N^2,7-Trimethylguanosine**
26	41.2	312	**N^2,N^2-Dimethylguanosine**
27	45.0	282	**N^6-Methyladenosine**
28	45.8	313	Ribosyl derivative 4
29	46.1	413	**N^6-Threonylcarbamoyladenosine**
30	47.9	298	**5′-Methylthioadenosine**
31	49.6	459	Ribosyl derivative 5

[a]Bold: identified nucleosides; italic: proposals based on mass spectrometric fragmentation and exact mass (unpublished results); other: unknown metabolites.

All computational experiments were carried out within the MATLAB™ programming environment. For performing the SVM trainings, we integrated LIBSVM (http://www.csie.ntu.edu.tw/∼cjlin/libsvm).

It is a common practice to summarize the outcome of a classifiers' prediction in form of a so-called confusion matrix, where each column represents the predicted class for the left out patient and each row represents the actual class (see Table III for illustration). If a healthy patient was correctly classified as healthy, we defined her as a true negative test case, and if a cancer patient was correctly predicted as having cancer, she was said to be a true positive case.

During the LOO procedure each patient was taken as an independent test case once. The LOO-*specificity* of the classifier is the number of true negative cases divided by the number of all healthy patients. Likewise, the LOO-*sensitivity* of the classifier is the number of true positive cases divided by the number of all cancer patients. Both, sensitivity and specificity are taken as a performance measure for our SVM classifier system here. They are commonly used in the machine learning literature (Duda et al.

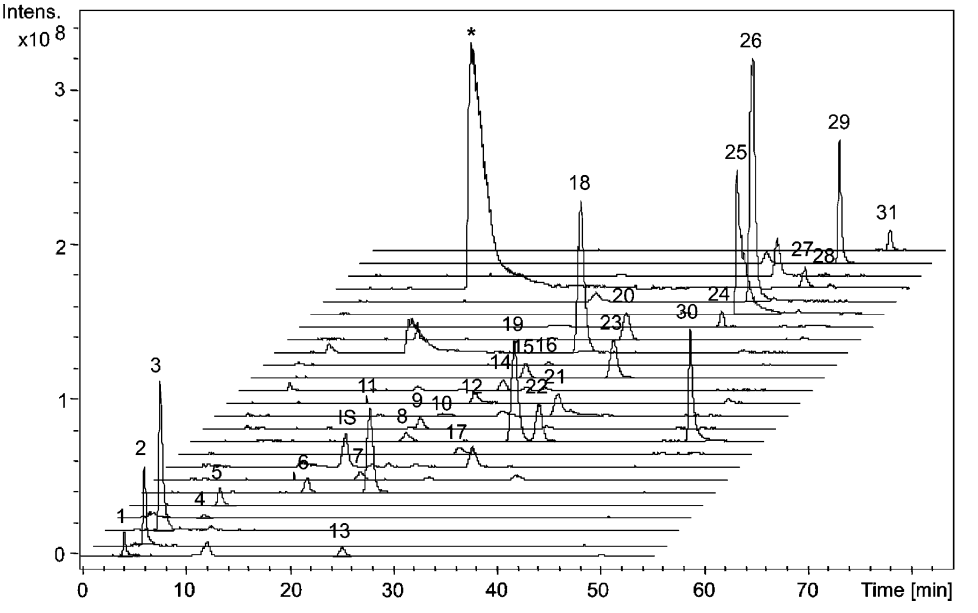

Figure 3. Extracted ion chromatograms (EIC) of all compounds included in the evaluation (for annotations compare Table II). *Modified nucleoside 1-methyladenosine (*m/z* 282, not included due to poor peak shape and linearity).

2001). In this context it should be noted that naturally there is a trade-off between a high sensitivity and a high specificity. While each classifier system can trivially achieve a perfect sensitivity by marking all test cases as positives, it would suffer a specificity of zero in that case. On the other hand a good classifier would gain both, high specificity and sensitivity.

Using the pre-described evaluation procedure we achieved a LOO-sensitivity of 87.67% and a LOO-specificity of 89.90%. In order to get a better impression we computed the statistical information content of each of the 31 nucleosides (see Table II) for the classification (Duda et al. 2001). The statistical information content, also called the *mutual information*, is a quantity that measures the mutual dependence of two random variables (in our case predictor variable X and a classification variable Y). Formally, it can be defined as follows:

$$I(X, Y) = \int_{\substack{(x,y)\in \\ dom(X)\times dom(Y)}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

Table III. Illustration of a confusion matrix for classification. Columns represent the predicted class, whereas rows represent the actual class of a patient.

	Predicted cancer	Predicted healthy
Cancer	True positive	False negative
Healthy	False positive	True negative

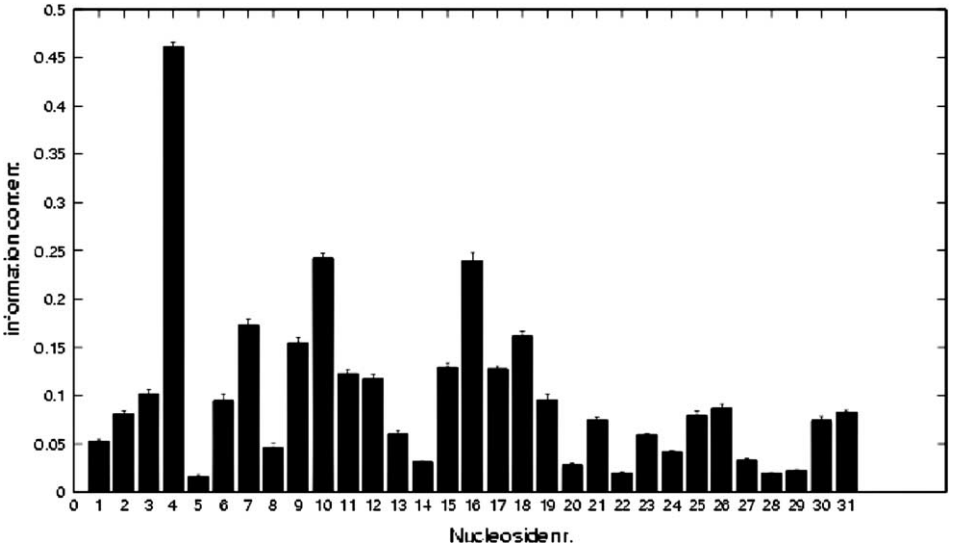

Figure 4. Statistical information content (±standard deviation) of the single nucleosides for the classification (for annotations compare Table II).

The result depicted in Figure 4 clearly shows that the information content among all 31 nucleosides is quite imbalanced and suggests the elimination of those carrying the least information. Compounds with a very high information content are for example the unidentified compounds with *m/z* 228 (no. 4), *m/z* 293 (no. 10) and *m/z* 351 (no. 16), whereas for instance 3-(3-amino-3-carboxypropyl) uridine (no. 5) carries a very small information.

We then investigated which nucleosides in combination yield the highest impact on the classification. It should be noted that this is not as simple as selecting the nucleosides that have the highest information content by themselves. What matters is the choice of the subset of all nucleosides. Indeed it was observed (Guyon & Elisseeff 2003) that even if a nucleoside was completely useless by itself, in combination with others it could eventually increase the prediction quality.

For this purpose we applied the recursive feature elimination (RFE) algorithm (Guyon & Elisseeff 2003), which successively eliminates the nucleoside that least affects the SVM solution (Figure 5).

To ensure a high statistical robustness of the solution against random fluctuations in the dataset, we then only considered those nucleosides as most relevant, which were constantly selected by the RFE algorithm during the LOO procedure. These were the nucleosides with *m/z* 228 (1-ribosylpyridin-4-one, no. 4), 301 (no. 9), 271 (1-ribosylpyridin-3-one-4-carboxamide, no. 11), 268 (putative 2-aminopurin-9-riboside, no. 15), 384 (N^6-succinyladenosine, no. 17), 1-methylguanosine (no. 19), 326 (N^2,N^2-7-trimethylguanosine, no. 25), N^6-methyladenosine (no. 27), *m/z* 313 (no. 28), MTA (no. 30) and *m/z* 459 (no. 31), among which several have not been identified yet. By applying the RFE algorithm on each training set, our LOO-sensitivity was now 89.39% and the LOO-specificity 88.89%, which is similar to the result obtained with all 31 nucleosides included.

To compare with a more traditional analysis, we also plotted the projection of our data onto its first two principal components (Figure 6). As clearly seen, it is impossible

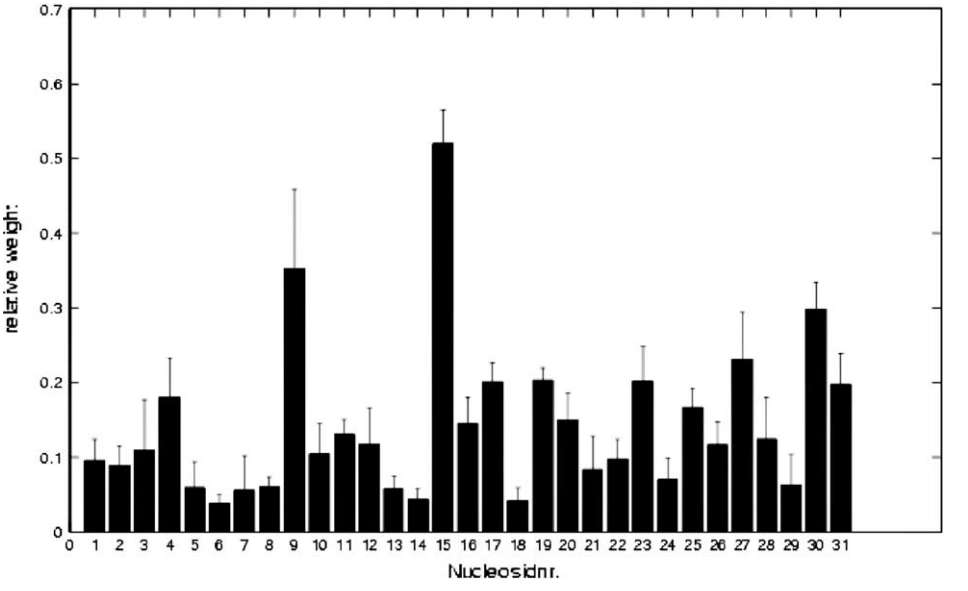

Figure 5. Relative weight (±standard deviation) of each nucleoside (for annotations compare Table II) on the support vector machine (SVM) solution at the initialization phase of the recursive feature elimination (RFE) algorithm. In the next step the nucleoside having the smallest weight is eliminated and all weights are then recomputed.

to separate the two classes in this projection, whereas – as shown before – with our SVM classification the task can be solved reliably through a RBF kernel function using a non-linear projection to a very high dimensional feature space.

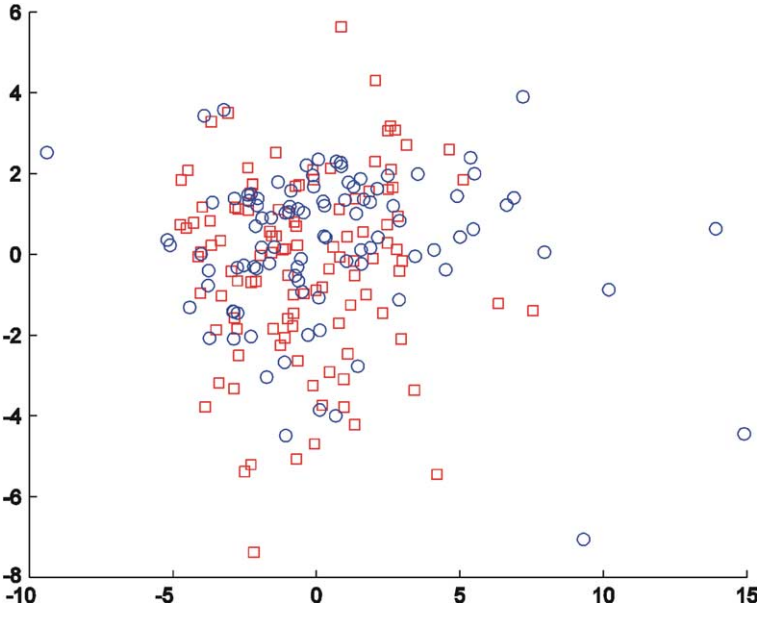

Figure 6. Projection of the data onto the first two principal components. The two classes (circles, healthy patients; boxes, cancer patients) cannot be separated.

Conclusions

Nucleoside modification is an integral part of a complex, sequential and multi-enzymatic process of RNA maturation. In general, the investigation of differing nucleosidic profiles in healthy volunteers and cancer patients, can be regarded as a result of RNA editing and catabolism as well as of different RNA maturation processes in cancerous and healthy tissues.

In this context, a method to differentiate between breast cancer patients and healthy individuals based on the examination of nucleosides/ribosyl derivatives extracted from urine samples by affinity chromatography has been developed. The separation was performed by reversed-phase HPLC coupled to an ion trap mass spectrometer. The integration results of 31 nucleosides, some of which have not been identified yet, while others were characterized by their fragmentation behaviour, were examined by S octopole RF amplitude SVMs, which are a popular classification method in machine learning. Using a leave-one-out cross-validation procedure, we achieved a sensitivity of 87.67% and a specificity of 89.90% of our classifier. By applying the RFE algorithm to automatically identify the most relevant nucleosides within the leave-one-out procedure a sensitivity of 89.39% and a specificity of 88.89% was obtained, which was comparable to the previous results.

These results show a significant improvement in comparison to currently applied tumour markers in breast cancer diagnosis such as the typically used CA15-3 tumour marker (specificity 90–95%, sensitivity 50%). Including unknown compounds with nucleoside structures clearly demonstrates the impairment of the RNA metabolism, which is associated with the disease.

The use of non-invasive methods with both good sensitivity and selectivity like the one presented here could mean an important improvement in cancer diagnosis, as no reliable tumour markers in breast cancer diagnosis are currently available. The diagnosis of other types of cancer faces similar problems, so methods like these presented here may also be applied to other fields of oncology.

Acknowledgements

We thank Prof. H. Seeger and Dr M. Zwirner (Frauenklinik, University Hospital Tübingen) for support in the recruitment process. Furthermore we thank K. Meziane and Dr F. Klaus for the extraction of the urine samples, G. Nicholson (Tübingen University, Institute of Organic Chemistry) for the ESI-FTICR MS measurements and Prof. J.H. Kim (Seoul University, South Korea) for the donation of isoguanosine.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Bartels H, Bohmer M, Heierli C. 1975. Serum creatinine determination without protein precipitation. Clinica Chimica Acta 37:193–197.

Beger RD, editor. 2005. Metabonomics of Cancer. November 1–4. Jefferson, AR, USA: Food and Drug Administration.

Bjoerk GR, Ericson JU, Gustafsson CED, Hagervall TG, Joensson YH, Wikstroem PM. 1987. Transfer RNA modification. Annual Review of Biochemistry 56:263–287.

Bullinger D, Frickenschmidt A, Pelzing M, Zey T, Zurek G, Laufer S, Kammerer B. 2005. Identification of urinary nucleosides by ESI-TOF-MSLC-GC Europe:16–17.

Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning 20:273–297.
Dieterle F, Muller-Hagedorn S, Liebich HM, Gauglitz G. 2003. Urinary nucleosides as potential tumor markers evaluated by learning vector quantization. Artificial Intelligence in Medicine 28:265–279.
Duan K, Rajapakse JC. 2005. SVM-RFE peak selection for cancer classification with mass spectrometry data. Series on Advances in Bioinformatics and Computational Biology 1:191–200.
Duda R, Hart P, Stork G. 2001. Pattern Classification. 2nd edition. New York: Wiley Interscience.
Dudley E, Lemiere F, Van Dongen W, Langridge JI, El Sharkawi S, Games DE, Esmans EL, Newton RP. 2003. Analysis of urinary nucleosides. III. Identification of 5′-deoxycytidine in urine of a patient with head and neck cancer. Rapid Communications in Mass Spectrometry 17:1132–1136.
Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914.
Gamache PH, Meyer DF, Granger MC, Acworth IN. 2004. Metabolomic applications of electrochemistry/Mass spectrometry. Journal of the American Society of Mass Spectrometry 15:1717–1726.
Griffin JL. 2004. The potential of metabonomics in drug safety and toxicology. Drug Discovery Today: Technologies 1:285–293.
Guyon I, Elisseeff A. 2003. An introduction into variable and feature selection. J Machine Learning Research 3:1157–1182.
Holloway DT, Kon M, DeLisi C. 2005. Integrating genomic data to predict transcription factor binding. Genome Informatics 16:83–94.
Honda K, Hayashida Y, Umaki T, Okusaka T, Kosuge T, Kikuchi S, Endo M, Tsuchida A, Aoki T, Itoi T, Moriyasu F, Hirohashi S, Yamada T. 2005. Possible detection of pancreatic cancer by plasma protein profiling. Cancer Research 65:10613–10622.
Itoh K, Aida S, Ishiwata S, Sasaki S, Ishida N, Mizugaki M. 1993. Urinary excretion patterns of modified nucleosides, pseudouridine and 1-methyladenosine, in healthy individuals. Clinica Chimica Acta 217:221–223.
Kammerer B, Frickenschmidt A, Gleiter CH, Laufer S, Liebich H. 2005a. MALDI-TOF MS analysis of urinary nucleosides. Journal of the American Society of Mass Spectrometry 16:940–947.
Kammerer B, Frickenschmidt A, Muller CE, Laufer S, Gleiter CH, Liebich H. 2005b. Mass spectrometric identification of modified urinary nucleosides used as potential biomedical markers by LC-ITMS coupling. Anals of Bioanalytical Chemistry 382:1017–1026.
Kim H, Page GP, Barnes S. 2004. Proteomics and mass spectrometry in nutrition research. Nutrition 20:155–165.
Kleno TG, Kiehr B, Baunsgaard D, Sidelmann UG. 2004. Combination of 'omics' data to investigate the mechanism(s) of hydrazine-induced hepatotoxicity in rats and to identify potential biomarkers. Biomarkers 9:116–138.
Lenz EM, Bright J, Knight R, Westwood FR, Davies D, Major H, Wilson ID. 2005. Metabonomics with 1H-NMR spectroscopy and liquid chromatography-mass spectrometry applied to the investigation of metabolic changes caused by gentamicin-induced nephrotoxicity in the rat. Biomarkers 10:173–187.
Lepp Z, Kinoshita T, Chuman H. 2006. Screening for new antidepressant leads of multiple activities by support vector machines. Journal of Chemical Information & Modeling 46:158–167.
Liebich HM, Di SC, Wixforth A, Schmid HR. 1997. Quantitation of urinary nucleosides by high-performance liquid chromatography. Jouranl of Chromatography A 763:193–197.
Liebich HM, Muller-Hagedorn S, Bacher M, Scheel-Walter H-G, Lu X, Frickenschmidt A, Kammerer B, Kim K-R, Gerard H. 2005. Age-dependence of urinary normal and modified nucleosides in childhood as determined by reversed-phase high-performance liquid chromatography. Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences 814:275–283.
Lindon JC, Holmes E, Bollard ME, Stanley EG, Nicholson JK. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 9:1–31.
Liu Y. 2006. Serum proteomic pattern analysis for early cancer detection. Technology in Cancer Research & Treatment 5:61–66.
Luntz A, Brailovsky V. 1969. On estimation of characters obtained in statistical procedure of recognition. Technicheskaya Kibernetica 3.
Plumb RS, Stumpf CL, Gorenstein MV, Castro-Perez JM, Dear GJ, Anthony M, Sweatman BC, Connor SC, Haselden JN. 2002. Metabonomics: the use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. Rapid Communications in Mass Spectrometry 16:1991–1996.

Plumb RS, Stumpf CL, Granger JH, Castro-perez J, Haselden JN, Dear GJ. 2003. Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. Rapid Communications in Mass Spectrometry 17:2632–2638.

Prankel BH, Clemens PC, Burmester JG. 1995. Urinary excretion of nucleosides varies with age and protein metabolism. Clinica Chimica Acta 234:181–183.

Rajapakse JC, Duan K, Yeo WK. 2005. Proteomic cancer classification with mass spectrometry data. American Journal of Pharmacogenomics 5:281–292.

Roessner U, Willmitzer L, Fernie AR. 2002. Metabolic profiling and biochemical phenotyping of plant systems. Plant Cell Reports 21:189–196.

Saeh JC, Lyne PD, Takasaki BK, Cosgrove DA. 2005. Lead hopping using SVM and 3D pharmacophore fingerprints. Journal of Chemical Information & Modeling 45:1122–1133.

Sander G, Hulsemann J, Topp H, Heller-Schoch G, Schoch G. 1986. Protein and RNA turnover in preterm infants and adults: a comparison based on urinary excretion of 3-methylhistidine and of modified one-way RNA catabolites. Annals of Nutrition & Metabolism 30:137–142.

Sasco AJ, Rey F, Reynaud C, Bobin JY, Clavel M, Niveleau A. 1996. Breast cancer prognostic significance of some modified urinary nucleosides. Cancer Letters 108:157–162.

Schölkopf B, Smola A. 2002. Learning with Kernels. Cambridge: MIT Press.

Tormey DC, Waalkes TP, Gehrke CW. 1980. Biological markers in breast carcinoma – clinical correlations with pseudouridine, N2,N2-dimethylguanosine, and 1-methylinosine. Journal of Surgical Oncology 14:267–273.

Wang C, Kong H, Guan Y, Yang J, Gu J, Yang S, Xu G. 2005. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. Anals of Chemistry 77:4108–4116.

Wilson ID, Plumb R, Granger J, Major H, Williams R, Lenz EM. 2005. HPLC-MS-based methods for the study of metabonomics. Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences 817:67–76.

Xu G, Schmid HR, Lu X, Liebich HM, Lu P. 2000. Excretion pattern investigation of urinary normal and modified nucleosides of breast cancer patients by RP-HPLC and factor analysis method. Biomedical Chromatography 14:459–463.

Yang J, Xu G, Zheng Y, Kong H, Pang T, Lv S, Yang Q. 2004. Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences 813:59–65.

Yang J, Xu G, Zheng Y, Kong H, Wang C, Zhao X, Pang T. 2005. Strategy for metabonomics research based on high-performance liquid chromatography and liquid chromatography coupled with tandem mass spectrometry. Journal of Chromatography A 1084:214–221.