Bidirectional Correlation of NMR and Capillary Electrophoresis Fingerprints: A New Approach to Investigating Schistosoma mansoni Infection in a Mouse Model

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2009

Impact Factor: 5.64 · DOI: 10.1021/ac901728w · Source: PubMed

10 AUTHORS, INCLUDING:

Alexessander Alves
Imperial College London
43 PUBLICATIONS **521** CITATIONS
SEE PROFILE

Jia V Li
Imperial College London
52 PUBLICATIONS **1,529** CITATIONS
SEE PROFILE

Cristina Legido-Quigley
King's College London
63 PUBLICATIONS **1,188** CITATIONS
SEE PROFILE

Carmen Silvia Valente Barbas
Hospital Israelita Albert Einstein
186 PUBLICATIONS **4,210** CITATIONS
SEE PROFILE

Anal. Chem. **2010,** *82,* 203–210

Bidirectional Correlation of NMR and Capillary Electrophoresis Fingerprints: A New Approach to Investigating *Schistosoma mansoni* Infection in a Mouse Model

I. Garcia-Perez,[†] A. Couto Alves,[‡] S. Angulo,[†] J. V. Li,[‡] J. Utzinger,[||] T. M. D. Ebbels,[‡] C. Legido-Quigley,[‡,§] J. K. Nicholson,[‡] E. Holmes,[‡] and C. Barbas*[,†]

Faculty of Pharmacy, San Pablo-CEU, Campus Monteprincipe, Boadilla del Monte, 28668 Madrid, Spain, Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College, South Kensington, London, SW7 2AZ, United Kingdom, Department of Public Health and Epidemiology, Swiss Tropical Institute, P. O. Box, CH-4002, Basel, Switzerland, and PSRD, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, United Kingdom

We demonstrate the statistical integration of nuclear magnetic resonance (NMR) spectroscopy and capillary electrophoresis (CE) data in order to describe a pathological state caused by *Schistosoma mansoni* infection in a mouse model based on urinary metabolite profiles. Urine samples from mice 53 days post infection with *S. mansoni* and matched controls were analyzed via NMR spectroscopy and CE. The two sets of metabolic profiles were first processed and analyzed independently and were subsequently integrated using statistical correlation methods in order to facilitate cross assignment of metabolites. Using this approach, metabolites such as 3-ureidopropionate, *p*-cresol glucuronide, phenylacetylglycine, indoxyl sulfate, isocitrate, and trimethylamine were identified as differentiating between infected and control animals. These correlation analyses facilitated structural elucidation using the identification power of one technique to enhance and validate the other, but also highlighted the enhanced ability to detect functional correlations between metabolites, thereby providing potential for achieving deeper mechanistic insight into the biological process.

One of the strategies employed by metabolic phenotyping is metabolic fingerprinting, which involves rapid, high-throughput global analysis to discriminate between samples of different biological states or origins.[1,2] Various metabolic profiling strategies have been applied to the investigation of diseases via the generation of characteristic metabolic fingerprints, which can be used either for disease diagnosis/prognosis or for uncovering changes in pathways, or potential biomarkers underlying mechanisms of disease.[3−5] These technologies have been used to characterize pathologies as diverse as heart disease, neurodegeneration, diabetes, intestinal disorders, cancer, and infectious diseases[6−8] based on the chemical composition of biofluids or tissues. Two of the most commonly used metabolic profiling technologies are proton nuclear magnetic resonance (^{1}H NMR) spectroscopy and mass spectrometry (MS) based methods, with or without prior chromatographic separation.[9−12] However, other analytical platforms such as capillary electrophoresis (CE), high-performance liquid chromatography (HPLC), and infrared (IR) spectroscopy have also been successfully applied to discriminating between healthy and disease states.[3,13−15] Each analytical technology or platform demonstrates different strengths and weaknesses with regard to the ability to differentiate between different biological conditions, depending upon factors such as sensitivity, ease of structural identification of discriminatory

* To whom correspondence should be addressed. Tel: +34 91 37 24 711. E-mail: cbarbas@ceu.es.

[†] San Pablo-CEU.

[‡] Imperial College London.

[||] Swiss Tropical Institute.

[§] King's College London.

(1) Gavaghan, L.; Holmes, E.; Lenz, E. M.; Wilson, I. D.; Nicholson, J. K. *FEBS Lett.* **2000**, *484*, 169–174.

(2) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157–1161.

(3) Ellis, D. I.; Goodacre, R. *Analyst* **2006**, *131*, 875–885.

(4) Garcia-Perez, I.; Vallejo, M.; Garcia, A.; Legido-Quigley, C.; Barbas, C. *J. Chromatogr., A* **2008**, *1204*, 130–139.

(5) Papaspyridonos, K.; Garcia-Perez, I.; Angulo, S.; Domann, P. J.; Vilca-Melendez, H.; Heaton, N.; Murphy, G. M.; Holmes, E.; Barbas, C.; Legido-Quigley, C. *J. Sep. Sci.* **2008**, *31*, 3058–3064.

(6) Williams, R.; Lenz, E. M.; Wilson, A. J.; Granger, J.; Wilson, I. D.; Major, H.; Stumpf, C.; Plumb, R. *Mol. BioSyst.* **2006**, *2*, 174–183.

(7) Duarte, I. F.; Legido-Quigley, C.; Parker, D. A.; Swann, J. R.; Spraul, M.; Braumann, U.; Gil, A. M.; Holmes, E.; Nicholson, J. K.; Murphy, G. M.; Vilca-Melendez, H.; Heaton, N.; Lindon, J. C. *Mol. BioSyst.* **2009**, *5*, 180–190.

(8) Chan, E. C.; Koh, P. K.; Mal, M.; Cheah, P. Y.; Eu, K. W.; Backshall, A.; Cavill, R.; Nicholson, J. K.; Keun, H. C. *J. Proteome Res.* **2009**, *8*, 352–361.

(9) Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protoc.* **2007**, *2*, 2692–2703.

(10) Bollard, M. E.; Contel, N.; Ebbels, T.; Smith, L.; Beckonert, O.; Cantor, G.; Lehman-McKeeman, L. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K.; Keun, H. C. *J. Proteome Res.* **2009**.

(11) Trump, S.; Laudi, S.; Unruh, N.; Goelz, R.; Leibfritz, D. *Magma* **2006**, *19*, 305–312.

(12) Ma, C.; Wang, H.; Lu, X.; Xu, G.; Liu, B. *J. Chromatogr., A* **2008**, *1186*, 412–419.

(13) Garcia-Perez, I.; Whitfield, P.; Bartlett, A.; Angulo, S.; Legido-Quigley, C.; Hanna-Brown, M.; Barbas, C. *Electrophoresis* **2008**, *29*, 3201–3206.

(14) Michopoulos, F.; Lai, L.; Gika, H.; Theodoridis, G.; Wilson, I. *J. Proteome Res.* **2009**, *8*, 2114−2121.

(15) Plumb, R. S.; Granger, J. H.; Stumpf, C. L.; Johnson, K. A.; Smith, B. W.; Gaulitz, S.; Wilson, I. D.; Castro-Perez, J. *Analyst* **2005**, *130*, 844–849.

10.1021/ac901728w
Published on Web 12/04/2009

molecules, sample preparation, analytical stability, analytical reproducibility, cost of equipment and reagents, efficiency, and ease of operation. In general, given that all data sets measured to date only probe subsets of the total metabolome, the combined use of two or more analytical technologies provides a more robust strategy for classifying disease and for identifying and validating candidate biomarkers than the use of a single platform, providing the data can be integrated and coanalyzed effectively. To date, there has been relatively little systematic assessment of multiplatform diagnostic strategies, although some statistical tools such as Statistical HeterospectroscopY (SHY),[16] bidirectional partial least square (PLS) with orthogonal filtration (O2-PLS),[17] Intra-Intermetabolite spectroscopy,[18] and Pearsons correlation methods[19−21] have been specifically developed for investigating correlation between data generated across different analytical platforms. Here, we develop and apply a strategy for statistically coanalyzing a combination of NMR spectroscopy and CE to characterize an infection of *Schistosoma mansoni* in the mouse host. This disease model has been selected on the basis that two previous independent studies have shown that characteristic metabolite fingerprinting of *S. mansoni* infection in mouse urine can be obtained using these techniques.[13,22,23] With regard to CE, an analytical strategy, combining two polarities and CE modes, has been employed, demonstrating improved coverage of metabolite classes over the use of single polarity data sets, which permits maximum information from the samples to be obtained.

Schistosomiasis is a chronic and debilitating helminthic parasitic disease with a high associated morbidity.[24,25] Chronic infection results in periportal fibrosis of the liver, localized inflammation, calcification of the bladder, and other pathological sequelae.[26] Since the control and treatment largely relies on a single therapeutic, the anthelmintic drug, praziquantel, the development of resistance would impose a severe problem.[26] Hence, there is an immediate need to achieve a better understanding of the pathological mechanisms associated with this parasite and to focus on identifying potential drug targets and more sensitive methods of diagnosis.

Both NMR and CE are able to measure biological samples directly, without any sample pretreatment, but each technique has inherent strengths and drawbacks. NMR has excellent molecular structural identification capability, which CE with UV detection lacks, whereas CE is an inexpensive tool and is easy to implement in any laboratory even in developing countries where schistosomiasis is endemic. For this reason, integration of data generated by both techniques can be used to not only corroborate the results from the other analytical platform but also harness the complementarity of the two techniques to improve biomarker identification, which may lead to practical widespread implementation of metabolic profiling. In the current work, a new multivariate graphical statistical approach based on the linear autocorrelation matrix (ACM) study will be used to characterize the in vivo metabolic pathway perturbations of a model of *S. mansoni* infection in mice based on the correlation between CE and NMR data. Additionally, bidirectional analysis of NMR and CE data has also been implemented in order to provide an unbiased, sensitive approach to biomarker extraction and identification with respect to assessing the potential for generating novel pathway connectivities.

MATERIALS AND METHODS

Animal Husbandry and Sample Collection. The animal experiments were carried out at the Swiss Tropical Institute (Basel, Switzerland), adhering to local and national guidelines of animal welfare. Details of the experimental design, sample collection and NMR data acquisition have been described in detail elsewhere.[23] In brief, a 15 female NMRI strain mice, aged ∼3 weeks, were purchased from RCC (Itingen, Switzerland). Mice were individually marked and were randomly housed into 3 cages of 5 mice each. Animals were acclimatized for 2 weeks; hence, at the onset of the experiment, mice were ∼5 weeks old, weighing between 17.3 and 27.3 g. Mice were kept under environmentally stable conditions in the animal facility (temperature, 22 °C; relative humidity, 60−70%; light/dark cycle, 12/12 h), fed on special pellets (PAB45−NAFAG 9009, Provimi Kliba; Kaiseraugst, Switzerland), and watered ad libitum. Ten of the mice were each infected intraperitoneally with 80 *S. mansoni* cercariae. The remaining mice were not infected and, hence, served as a control group.

Samples were collected 53 days after infection in order to ensure characterization of the patent stage of infection.

Samples from noninfected control mice were collected at the same time points. Urine samples were collected into Petri dishes by gently rubbing the abdomen of the mice, and the collected samples were immediately transferred into separate Eppendorf tubes and maintained at −40 °C pending analysis.

Sample Preparation for CE-UV Analysis. For each sample, an aliquot of 20 μL of urine was mixed with 20 μL of purified water, centrifuged, and injected directly into the CE apparatus. All samples were initially run with positive polarity and then with negative polarity, using the same sample vials.

Chemicals and Instrumentation Used for CE Analysis. Sodium borate and sodium dodecyl sulfate (SDS) were purchased from Sigma (Poole, UK); sulphated β-cyclodextrin (analytical grade, Sβ_CD) was purchased from Aldrich, and sodium hydroxide and hydrochloric acid were purchased from BDH (Poole, UK). Reverse osmozed deionized water (Milli-Q Synthesis from Millipore, Bedford, MA) was used for standard solution and electrolyte preparations. Chemical standards used for biomarker identification were obtained from Sigma.

CE experiments were carried out on a P/ACE MDQ system (Beckman Instruments, Fullerton, CA) equipped with diode array UV−vis absorbance detection (190−600 nm), a temperature-controlled (liquid cooled) capillary compartment, and an autosampler. Electrophoretic data were acquired and analyzed with 32

(16) Crockford, D. J.; Holmes, E.; Lindon, J. C.; Plumb, R. S.; Zirah, S.; Bruce, S. J.; Rainville, P.; Stumpf, C. L.; Nicholson, J. K. *Anal. Chem.* **2006**, *78*, 363–371.

(17) Rantalainen, M.; Cloarec, O.; Beckonert, O.; Wilson, I. D.; Jackson, D.; Tonge, R.; Rowlinson, R.; Rayner, S.; Nickson, J.; Wilkinson, R. W.; Mills, J. D.; Trygg, J.; Nicholson, J. K.; Holmes, E. *J. Proteome Res.* **2006**, *5*, 2642–2655.

(18) Moco, S. I. A.; Forshed, J.; Vos, R. C. H. d.; Bino, R. J.; Vervoort, J. J. M. *Metabolomics* **2008**, *4*, 202–215.

(19) Zhuang, H.; Savage, E. M. *Poult. Sci.* **2009**, *88*, 214–220.

(20) Rybak, M. E.; Parker, D. L.; Pfeiffer, C. M. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2008**, *861*, 145–150.

(21) Bonadio, F.; Margot, P.; Delemont, O.; Esseiva, P. *Forensic Sci. Int.* **2008**, *182*, 52–56.

(22) Wang, Y.; Holmes, E.; Nicholson, J. K.; Cloarec, O.; Chollet, J.; Tanner, M.; Singer, B. H.; Utzinger, J. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12676–12681.

(23) Li, J. V.; Holmes, E.; Saric, J.; Keiser, J.; Dirnhofer, S.; Utzinger, J.; Wang, Y. *J. Parasitol.* **2009**, *39*, 547–558.

Karat software (P/ACE MDQ instrument). Separations achieved with normal polarity and micellar electrokinetic chromatography (NP, MEKC) were performed in a fused silica capillary (75 μm internal diameter), 57 cm total, 50 cm effective length (Composite Metal Services, Hallow, Worcester, UK). New capillaries were conditioned for 30 min at 25 °C with 0.1 M NaOH, followed by 0.1 M NaOH for 20 min, deionized water for 10 min, and running buffer (BGE1) for 10 min. Before each analysis, the capillaries used were washed with 0.1 M HCl and deionized water for 1 min and then 2 min with the run buffer. The running buffer (BGE1) comprised 25 mM sodium borate, 75 mM SDS, and 6.25 mM sulphated β-cyclodextrin. The pH was adjusted to 9.50 with 1 M NaOH, after addition of SDS and cyclodextrin. Buffer solutions were filtered through a 45 μm filter before use. The capillaries were maintained at 20 °C, with an 18 kV applied voltage and 5 s hydrodynamic injection. The current obtained in these conditions was around 150 μA.

Separations achieved with reverse polarity (RP) and capillary zone electrophoresis (CZE) were performed using a capillary (Beckman Coulter, Buckinghamshire, England) coated with polyacrylamide and 57 cm in total length and 50 μm in internal diameter. On first use, the capillaries were conditioned by a pressure flush of 0.1 M HCl (1 min), BGE2 (10 min), and an electrokinetic flush of electrolyte with 0.5 kV/cm (10 min). Between runs, the capillaries were flushed under pressure with deionized water (2 min) and BGE2 (2 min). All experiments were performed at 25 °C using a separation potential of −20 kV. Samples were injected at the cathode, with 0.5 psi (3447 Pa) pressure applied for 10 s. Resolved sample components were detected at the anode. Electrophoretic buffer BGE2 was prepared with 0.2 M phosphoric acid, adjusted to pH = 6.20 with NaOH, and 10% (v/v) methanol was added. (The current observed under these conditions was 70 μA.) Data were collected at a frequency of 4 Hz.

Preprocessing of CE Data. Multialignment of the electropherograms was performed with an in-house program developed in Matlab 7.0 (San Pablo-CEU University), using the Correlation Optimized Warping (COW) method previously described.[27]

Chemicals and Sample Preparation for ^{1}H NMR Spectroscopic Analysis. All chemicals used for NMR analysis were purchased from Sigma, Aldrich (Poole, UK).

Each urine sample (~30 μL) was prepared by addition of 25 μL of phosphate buffer (D_2O/H_2O = 9:1, v/v, 0.01% of sodium 3-(trimethylsilyl) propionate-2,2,3,3-d4 [TSP], pH = 7.4) in order to reduce any pH-induced signal shifts. D_2O was used as a field frequency lock, and TSP was used as an internal spectra reference. Approximately 55 μL of mixture was transferred into a micro-NMR tube with diameter of 1.7 mm for NMR analysis.

NMR Spectroscopic Analysis of Samples. ^{1}H NMR spectra were acquired using a Bruker DRX 600 MHz spectrometer (Rheinstetten, Germany) with a 5 mm TXI probe operating at 600.13 MHz. A standard 1-dimensional (1-D) NMR pulse sequence [recycle delay (RD)-90°-t_1-90°-t_m-90°-acquire free induction decay (FID)] was employed for the acquisition of all spectra. The water peak was suppressed by selective irradiation during the RD of 2 s and mixing time (t_m) of 100 ms; t_1 was fixed to 3 μs. The 90° pulse length was adjusted to approximately 10 μs. A total of 256 scans were recorded into 32 k data points with a spectral width of 20 ppm. An exponential function corresponding to a line broadening of 0.3 Hz was applied to each FID prior to the Fourier transformation.

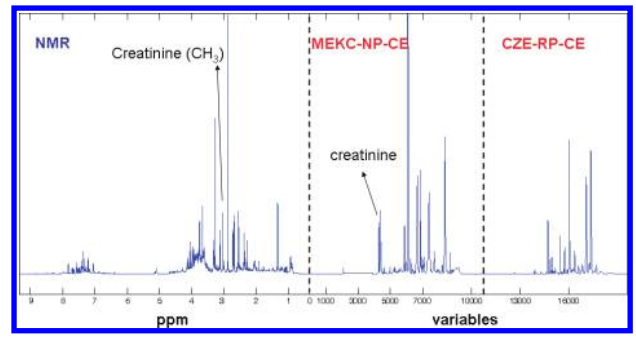

Figure 1. Reference fingerprints for NMR and CE profiles in both MEKC mode in normal polarity and CZE mode in reverse polarity. Creatinine signal on the NMR spectrum and the peak on the CE electropherogram are marked.

Preprocessing of NMR Data. The resulting ^{1}H NMR urine spectra were digitalized over the range of δ 0.35−10.0, imported into Matlab, automatically phased and baseline-corrected, and then normalized to the CH_2 creatinine resonance using a Matlab script developed in-house (Imperial College London). The regions containing the water resonance (δ 4.71−5.05) and the urea resonance (δ 5.17−6.36) were removed. Creatinine was quantified by fitting a Lorentzian function to the CH_2 resonance (4.05 ppm) and integrating it with boundary conditions set to $x_0 \pm 0.03$ ppm (x_0 is the part per million value of the peak apex). The baseline was included in the peak estimation and subtracted before integration. The NMR spectra were then aligned using Matlab software[28] before data analysis.

CE and NMR Data Analysis. Correlation analysis was performed between the CE and the NMR data using two approaches in order to establish single linear correlation. Initially, the raw data after baseline correction and multialignment from both data sets were used. Data from NMR and CE were combined in a single matrix, and a program, developed in Matlab based on the Pearson correlation coefficient, was applied to identify and plot correlations among variables (ACM). First, each set of data was independently normalized with the corresponding creatinine value which appears in both NMR and MEKC-NP-CE, as can be seen in Figure 1. Furthermore, both sets were scaled to the highest peak in the profile to give equal weight to all signals. The Pearson correlation coefficient (r) is a measurement of the strength and direction of a linear relationship between two random variables X and Y. It calculates values ranging from −1 (perfect negative correlation) and 1 (perfect positive correlation). Data were arranged in a matrix, X, with m variables (columns) and n individuals (rows), and a new correlation matrix was obtained, C, which had elements with the correlation coefficients obtained after comparing all the variables.

Second, to characterize the metabolic consequences of a *S. mansoni* infection and to maximize separation between *S. mansoni*-infected and noninfected mice samples, (PLS-DA)[29] models were

(24) Magnussen, P. *Acta Trop.* **2003**, *86*, 243–254.
(25) Grysseels, B.; Polman, K.; Clerinx, J.; Kestens, L. *Lancet* **2006**, *368*, 1106–1118.
(26) Doenhoft, M. J.; Cioli, D.; Utizinger, J. *Curr. Opin. Infect. Dis.* **2008**, *21*, 659–667.

constructed using unit variance scaled and mean-centered NMR and/or CE data as the descriptor matrix and class information (i.e., infected or control) as a binary response variable. In order to make meaningful comparisons, the number of PLS components was constrained to be smaller than one-third of the sample size. The number of latent components, h, that maximizes predictivity was estimated to be $h = 5$ for all data sets: NMR, NP-CE, RP-CE, and the combined data set. The model predictivity was assessed by the Q^2 parameter,[30] indicating how well the model predicts new data using cross validation.

Statistical total correlation spectroscopy (STOCSY) (^{1}H NMR-NMR),[31,32] bidimensional Statistical CE-CE correlation (S-NP, MEKC-RP, CZE CE), and correlation between NMR and CE (CE-NMR- SHY) methods were applied to the data. Correlations greater than a threshold θ of 0.8 for self-correlation (^{1}H NMR-NMR) and 0.3 for heteroanalytical correlation, corresponding to high sensitivity and a high probability of detecting structural correlations,[33] were plotted. Peaks found to be significantly different between infected and noninfected samples in the PLS-DA models were identified in order to elucidate the structure of metabolites and to establish possible pathway associations with other metabolites. Correlations between NMR and CE data, driven from selected identified differentiating metabolite signals in the NMR or CE profiles, were further analyzed to aid peak identification.

RESULTS AND DISCUSSION

Both NMR[22,23] and CE[13] have been shown independently to be suitable metabolite profiling tools with respect to characterizing a *S. mansoni* infection in the mouse. Typical profiles from each technique are illustrated in Figure 1, after scaling to creatinine, each showing a data-rich metabolic fingerprint. Multivariate statistical analysis applied to each of these analytical data sets independently showed good separation between infected and control animals. Following the generation of principal component analysis (PCA) score plots, which showed discrimination between control and infected animals (data not shown), PLS-DA was used to achieve optimal infection-related differentiation and to identify candidate biomarkers of the infection (Figure 2). Here, we show the enhanced quality of a PLS-DA model built using the combined data derived from ^{1}H NMR spectroscopy and from normal and reverse polarity CE over PLS-DA models generated from individual analytical methods (Figure 2) with the predictivity (Q^2) of the model increasing from 0.82 (NMR), 0.76 (NP-CE), and 0.59 (RP-CE) to 0.88 for the combined data set.

However, the real value of data integration across multiple analytical platforms is not only to provide a more robust description of the metabolic response to a biological challenge but also to harness the strengths of both technologies to improve biomarker identification and to cross validate results from the other analytical platform. We use correlation analyses to mine and interpret the spectra generated by each platform in order to establish (i) structural correlations, where signals deriving from the same chemical compound correlate across the profiles generated via two different techniques and (ii) metabolic correlations, or functional correlations, whereby signals corresponding to compounds which share a metabolic pathway or demonstrate a shared response to disease covary, either within or between platforms.

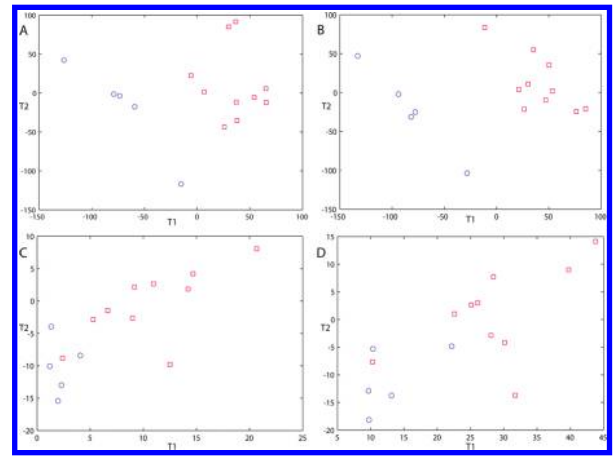

Figure 2. Scores plot of the first two latent components of the PLS-DA models of spectra/electropherograms of urine collected from mice with *S. mansoni* infections (□) and from uninfected control mice (○). Panel B shows a better class separation on the combined data set of ^{1}H NMR and {MEKC-NP, CZE-RP}-CE data than data sets constructed from only ^{1}H NMR (A), MEKC-NP-CE (C), or CZE-RP-CE (D).

Previously, we designed and applied a data treatment software based on Pearson's algorithm in order to find relationships between fingerprints obtained as sum of electropherograms (two CE modes).[27] The autocorrelation matrix (ACM) was applied to CE urine fingerprints from mice infected with *S. mansoni*, showing that one of the diagnostic metabolites for the infection, phenylacetylglycine (PAG), was positively correlated to four metabolites in the same pathway (phenylalanine, phenylacetate, 2-hydroxyphenylacetate, and 4-hydroxyphenylacetate), all with p values <0.05.[27] Likewise, we have applied correlation analyses to both ^{1}H NMR data (statistical total correlation spectroscopy or STOCSY) or between ^{1}H and other nuclei (heterocorrelation) in order to identify correlations from signals deriving from the same molecule or pathway-related molecules.[31,34] The same principle has also been used to cross correlate data from different analytical platforms, NMR and LC-MS data (SHY), with a view to achieving enhanced recovery of biomarkers of hydrazine toxicity.

Here, we extend this concept and apply statistical correlation to establish pairwise correlation structures between the NMR data and the normal and reverse polarity CE data, respectively. Analysis based on the data obtained from the individual analytical platforms resulted in the identification of infection-related modulations for a series of metabolites. As reported in previous studies, the analysis based on an individual data set for both NMR and CE identified decreased concentrations of urinary hippurate and citrate and increased concentrations of PAG, to be associated with

(27) Angulo, S.; Garcia-Perez, I.; Legido-Quigley, C.; Barbas, C. *Electrophoresis* **2009**, *30*, 1221–1227.
(28) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 56–66.
(29) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2007**, *20*, 341–351.
(30) Eriksson, L. J. E.; Kettaneh-Wold, N.; Wold, S. *Multi-and Megavariate Data Analysis and Applications*; Umetrics: Umea, 2001.
(31) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, *J. Anal. Chem.* **2005**, *77*, 1282–1289.

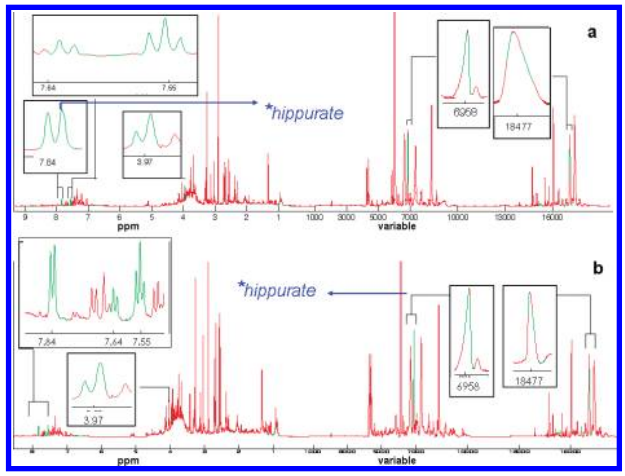

Figure 3. Correlation coded profiles using a significance criterion of $p = 0.01$ after selecting the following as the driver peak: (A) the hippurate signal at $\delta = 7.84$ in NMR showing correlation with the other hippurate signals in the NMR spectrum and with the hippurate signals in MEKC normal polarity mode and the hippurate in the CZE reverse polarity mode, clearly colored in green and (B) hippurate in the MEKC normal polarity mode, which appears in green and highly correlated to hippurate in CZE reverse polarity mode and to hippurate in NMR data.

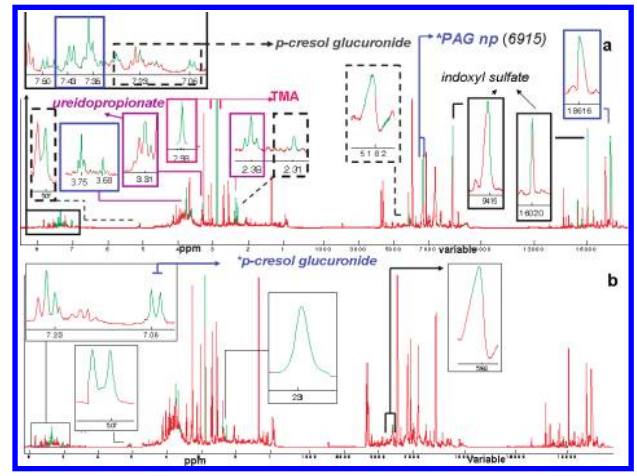

Figure 4. (a) Correlation structure obtained after selecting PAG in MEKC normal polarity mode and applying a $p = 0.005$ significance criterion (as indicated by green coding). Signals highly correlated to the MEKC peak for PAG include signals deriving from PAG (highlighted in blue) in the NMR spectra ($\delta = 7.43$ (m), $\delta = 7.35$ (m), $\delta = 3.75$ (d), $\delta = 3.68$(s)) and in CZE reverse polarity mode (var. 18616) to TMA (highlighted in pink, $\delta = 2.89$ (s)). Other signals highly correlated to the peak driver are indoxyl sulfate (highlighted in black in NP,MEKC-CE var.= 9415 $r = 0.87$ and in RP,CZE-CE var.= 16020 $r = 0.82$), 3-ureidopropionate (highlighted in purple, $\delta = 3.31$ (t) $r = 0.82$ and $\delta = 2.38$ (t) $r = 0.79$), and *p*-cresol glucuronide (highlighted by a dashed line $\delta = 7.23$ (d), $r = 0.78$ ($\delta = 7.06$ (d) $r = 0.83$, $\delta = 5.07$ (d) $r = 0.79$ and $\delta = 2.31$ (s), $r = 0.84$, and in NP,MEKC-CE var. = 5182 $r = 0.89$). (b) Correlation structure obtained after selecting *p*-cresol glucuronide in the NMR spectra ($\delta = 7.06$ (d)) and applying a $p = 0.009$ significant criterion. Strong correlation were found for the signals $\delta = 7.23$ (d)), $\delta = 5.07$ (d), and $\delta = 2.31$ (s) in the NMR spectra and for the signal at variable 5182 in the NP, MEKC-CE mode.

S. mansoni infection. Additionally, NMR analysis reported infection-related changes in trimethylamine (TMA), *p*-cresol glucuronide, trimethylamine, taurine, 2-oxoisocaproate, acetate, and 2-oxoisovalerate,[18] whereas CE uniquely identified urate, phenylalanine, indoxyl sulfate, isocitrate, and urea.[13]

In order to show proof of concept for cross assignment of signals across the analytical platforms, a signal corresponding to hippurate was chosen in the NMR matrix from which to drive a univariate Pearson correlation matrix encompassing all NMR and CE data. This signal was selected, as it was present in all three matrices NMR, MEKC-NP, and CZE-RP CE (Figure 3), with numerous signals deriving from the various proton environments apparent in the NMR spectra. Additionally, hippurate demonstrates consistent infection-related modulation. The signal at $\delta = 7.84$ was chosen since there is little ovelap of this signal with other endogenous signals in the NMR urine spectra. Statistically significant correlations ($p < 0.01$) with the other hippurate signals in the NMR spectrum and both modes of CE (MEKC normal polarity mode variable 6958 and in CZE reverse polarity mode variable 18477) were established (denoted in green in Figure 3A). As expected, the same result was obtained when the process was reversed and the correlation was driven from the hippurate signal in CE in MEKC normal polarity mode (Figure 3B), demonstrating that correlations were easily identified in both directions.

The next step was to investigate the capacity for detecting wider correlation structures such as those between molecules related by pathway or shared response to the infection. The variable with the highest weight for discriminating infection in the CE matrix, MEKC normal polarity mode (variable 6915) corresponding to phenylacetylglycine (PAG), was selected to generate a correlation matrix using a correlation cutoff level with $p < 0.005$. Signals that covaried across the data set are colored in green (Figure 4a) and signify a high correlation ($p < 0.005$) with this PAG signal. These signals included corresponding NMR signals deriving from PAG (multiplet $\delta = 7.43$ ($r = 0.96$), multiplet $\delta = 7.35$ ($r = 0.93$), multiplet $\delta = 3.75$ ($r = 0.90$), and singlet $\delta = 3.68$ ($r = 0.53$), highlighted in blue). The last one gave a lower correlation value due to signal overlap. Correlation between the driver PAG signal and PAG signals in both CE modes was also evident, but the correlation coefficient corresponding to the PAG signal in the CE reverse polarity mode (var. = 18616) was slightly lower ($r = 0.75$) due to the fact that PAG coelutes with urate in this mode. In addition to self-correlation, signals from other molecules also showed significant correlations. This included indoxylsulphate (highlighted in black: NP,MEKC-CE var.= 9415 $r = 0.87$; RP,CZE-CE var.= 16020 $r = 0.82$; and NMR $\delta = 7.71$ (d) $r = 0.79$, $\delta = 7.50$ (d) $r = 0.$ 7.36), 3-ureidopropionate (highlighted in purple: $\delta = 3.31$ (t)) $r = 0.82$ and $\delta = 2.38$ (t) $r = 0.79$), trimethylamine (TMA) (highlighted in pink: $\delta = 2.89$ (s) $r = 0.63$), *p*-cresol glucuronide (highlighted by a dashed line: $\delta = 7.23$ (d), $r = 0.78$ ($\delta = 7.06$ (d) $r = 0.83$, $\delta = 5.07$ (d) $r = 0.79$ and $\delta = 2.31$ (s), $r = 0.84$, NP,MEKC-CE var. = 5182 $r = 0.89$). This correlation structure suggests biological connectivities underpinned by gut microbial cometabolism. PAG is a gut microbiota-associated metabolite,[35] while TMA can be derived from dietary choline via microbial processing.[36] Likewise, urinary hippurate typically results from glycine conjugation to benzoic

(32) Crockford, D. J.; Lindon, J. C.; Cloarec, O.; Plumb, R. S.; Bruce, S. J.; Zirah, S.; Rainville, P.; Stumpf, C. L.; Johnson, K.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2006**, *78*, 4398–4408.
(33) Couto Alves, A.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. *Anal. Chem.* **2009**, *81*, 2075−2084.

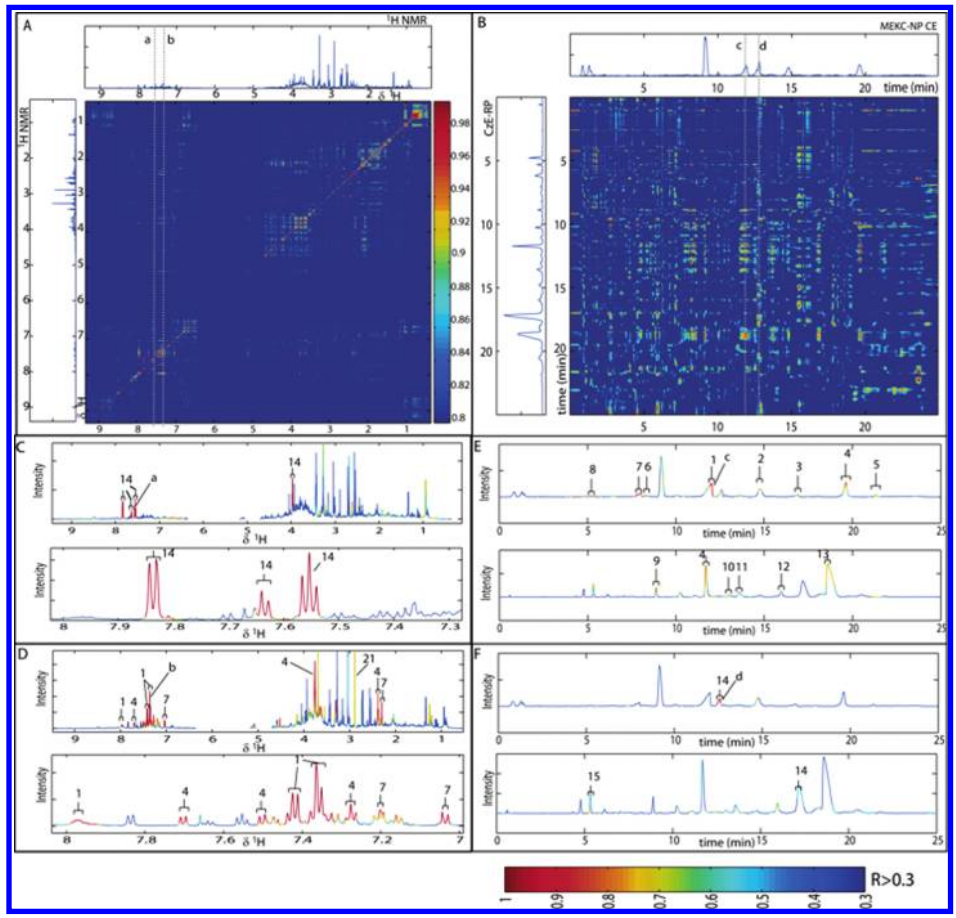

Figure 5. NMR-NMR correlations and CE normal and reverse polarity mode cross correlations depicting structural and biological correlations. (A) 2D STOCSY of the control and infected mice data thresholded at $\theta = 0.8$, corresponding to the whole NMR profile. Trace lines a and b indicate the correlation profile in the 2D plots corresponding to the driver peaks in the 1-D STOCSY plots (C and E). (B) Two-dimensional S-CE[2] of the control and infected mice data thresholded at $\theta = 0.3$, corresponding to the whole CE normal and reverse polarity profile. Trace lines c and d point at the correlation profile on the 2-D plots, corresponding to the driver peaks in 1-D S-CE (D and F). (C) 1-D S-NMR driven from peaks of the NMR spectrum at $\delta = 7.55$, corresponding to the metabolite hippurate depicted by the trace line a. (D) 1-D S-NMR driven from peaks of the NMR spectrum at $\delta = 7.36$, corresponding to the metabolite PAG depicted by the trace line b. (E) 1-D S-CE driven from peaks of the MEKC, NP CE mode at 12.02 min for the region shown in panel A depicted by the trace line c corresponding to the metabolite PAG. (F) 1-D S-CE driven from peaks of the MEKC, NP CE mode at 12.82 min for the region shown in panel A depicted by the trace line d. corresponding to the metabolite hippurate. Key to annotation of peaks in the NMR, MEKC,NP CE, and CZE,RP CE profiles: 1, phenylacetylglycine (PAG); 2, urate; 3-ureidopropionate; 4, indoxyl sulfate; 5, unknown; 6, phenylalanine; 7, *p*-cresol glucuronide; 8, adenine; 9, unknown; 10, phenylacetate; 11, 2-hydroxyphenylacetate; 12, 4-hydroxyphenylacetate; 13, colution of urate and PAG; 14, hippurate; 15, 2-oxoglutarate; 16, fumarate; 17, isocitrate; 18, citrate; 19, unknown; 20, unknown; 21, trimethylamine; 22, creatinine.

acid, and *p*-cresol glucuronide is derived from microbial degradation of tyrosine by species such as *Clostridium difficile*, *C. scatalogens*, and certain *Lactobacillus* species.[37,38]

The strength of the correlation coefficient indicates whether the correlation is likely to be structural or whether signals are more likely to be a product of functional association, e.g., from metabolites in shared pathways or resulting from perturbed functions, with stronger correlation values being suggestive of autocorrelation, i.e., correlation between signals deriving from the same metabolite.

A similar procedure was followed for *p*-cresol glucuronide which was found to be a strong discriminator of infection status in the NMR model.[22] Here, the signal at $\delta = 7.06$ deriving from the infected group was used as the driver signal and highlighted significant correlations with $\delta = 7.23$ (d), $\delta = 5.07$ (d), and $\delta = 2.31$ (s) in the NMR spectra which also belong to *p*-cresol glucuronide (Figure 4b). Strong correlations were also found for the signal at variable 5182 in the NP-CE mode. This signal was confirmed as *p*-cresol glucuronide on spiking with an authentic standard of the compound.

Another example of cross assignment of metabolites was the detection of indoxyl sulfate in the NMR spectra via correlation to the highly discriminatory signal (variable 9415) in the MEKC normal polarity mode CE matrix. Signals for indoxyl sulfate in NMR were $\delta = 7.50$ (d) ($r = 0.88$), $\delta = 7.71$ (d) ($r = 0.82$), $\delta = 7.36$(s) ($r = 0.80$), clearly highlighted in green; other signals such as $\delta = 7.209$ (dd) ($r = 0.75$) and $\delta = 7.28$ (dd) ($r = 0.69$) were only partially green, possibly due to some overlap in that area (figure not shown). This metabolite was not previously described as a discriminatory metabolite signal for *S. mansoni* infection

(34) Keun, H. C.; Athersuch, T. J.; Beckonert, O.; Wang, Y.; Saric, J.; Shockcor, J. P.; Lindon, J. C.; Wilson, I. D.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 1073–1079.
(35) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668–676.

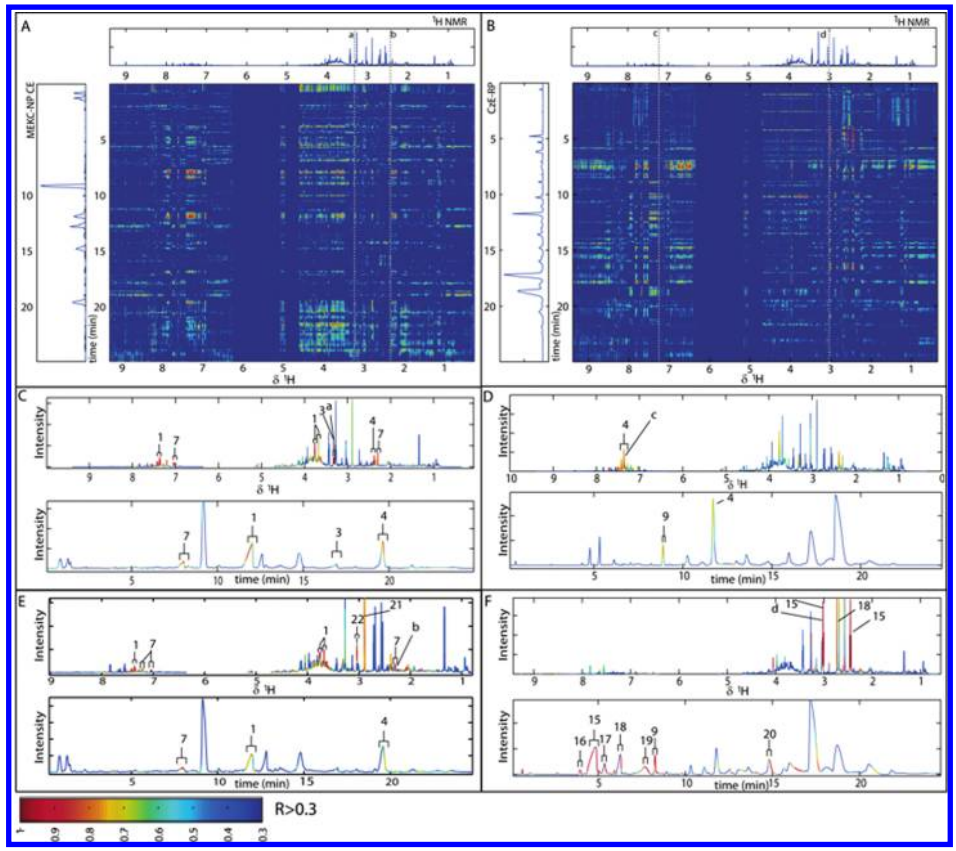

Figure 6. NMR vs. CE normal and reverse polarity mode cross correlations depicting structural and biological correlations. (A) 2-D S-NMR-MEKC, NP CE of the mice data thresholded at $\theta = 0.3$, corresponding to the NMR spectra vs the MEKC,NP CE. (B) 2-D S-NMR-RP,CZE, of the mice data thresholded at $\theta = 0.3$, corresponding to the NMR spectra vs the CE CZE,RP mode. Trace lines a−d point at the correlation profile on the 2-D plots, corresponding to the driver peaks in the S-NMR-MEKC,NP CE and S-NMR-CZE,RP. (C) 1D S-NMR-MEKC,NP CE plots between NMR and CE normal polarity mode profiles driven from peaks of the NMR spectra at $\delta = 2.31$, corresponding to the metabolite *p*-cresol glucuronide; the 1-D S-NMR-CE corresponds to the trace line b on panel A. (D) 1D S-NMR-RP,CZE plots between NMR and CE reverse polarity mode profiles driven from peaks of the NMR spectra at $\delta = 7.50$, corresponding to the metabolite indoxyl sulfate acid; the 1-D S-NMR-CZE,RP CE corresponds to the trace line c on panel B. (E) 1D S-NMR-MEKC,NP CE plots between NMR and CE normal polarity mode driven from peaks of the NMR spectra at $\delta = 3.31$, corresponding to the metabolite 3-ureidopropionate; the 1-D S-NMR-MEKC,NP CE corresponds to the trace line a on panel B. (F) 1D 1-D-S-NMR-CZE,RP CE plots between NMR and CZE,RP CE mode profiles driven from peaks of the NMR spectra at $\delta = 2.45$ ppm, corresponding to the metabolite 2-oxoglutarate (2OG); the 1-D S-NMR-RP,CZE corresponds to the trace line d on panel B. The peaks of the NMR, MECK, NP and CZE profiles were annotated with numbers: 1, PAG; 2, urate; 3, ureidopropionate; 4, indoxyl sulfate; 5, unknown; 6, phenylalanine; 7, *p*-cresol glucuronide; 8, adenine; 9, unknown; 10, phenylacetate; 11, 2-hydroxyphenylacetate; 12, 4-hydroxyphenylacetate; 13, coelution of urate and PAG; 14, hippurate; 15, 2OG; 16, fumarate; 17, isocitrate; 18, citrate; 19, unknown; 20, unknown; 21, TMA; 22, creatinine.

based on the infection models derived from NMR data. High levels of this metabolite were manifested in the urine of infected mice and may relate to perturbed renal function known to be caused by *S. mansoni* infection.[39]

To complement these results, a series of dimensional (2-D) correlation matrices were constructed including NMR-NMR (Figure 5A) and MEKC,NP-CZE,RP CE (Figure 5B), respectively, and the heterocorrelation matrix between NMR and CE (Figure 6A,B). These matrices contained a rich pattern of correlation structures displayed using a color coded hierarchy related to strength of correlation. To illustrate the correlation patterns, exemplar lines were extracted from areas of the matrix containing signals of interest. In the case of the STOCSY NMR plot, two lines corresponding to hippurate (Figure 5A.a) and PAG (Figure 5A.b) are illustrated. As expected, this produces a similar result to the 1D correlation plots. For example, hippurate resulted in a strong correlation with itself (Figure 5C), and PAG correlated with indoxyl sulfate, *p*-cresol glucuronide, 2-oxoglutarate, and TMA (Figure 5D). Although the information content relating to correlation for specific metabolites is similar to that generated by a 1-D correlation diagram, the 2-D plot places the correlation patterns for these metabolites within the global pattern of metabolic connectivities across the whole metabolite profile. Moreover, the 2D plots are not as susceptible to artifacts such as slight signal shifts, e.g., pH-dependent positional variation as the plots have better resolution.

The S-NP,MEKC-RP,CZE CE matrix of the normal versus reverse CE data was plotted, and the resulting correlelogram was investigated to recover additional biological and structural information. Here, the two axes of the plot represent NP and RP modes. Areas of red represent strong correlations while blue indicates

(36) Zeisel, S. H.; Wishnok, J. S.; Blusztajn, J. K. *J. Pharmacol. Exp. Ther.* **1983**, *225*, 320–324.

no correlation. As for Figure 5A, two lines corresponding to hippurate (Figure 5B.d) and PAG (Figure 5B.c) are illustrated. Thus, it is possible to visualize multiple correlations in this 2D plot relating to both self-correlation if a molecule was detected in both CE modes, as well as intermolecular correlations between closely related metabolites (Figure 5B). Since PAG and hippurate were highly discriminant between classes, a 1-D S-CE color coded electropherogram was plotted in order to investigate putative biological correlations with other metabolites. As can be seen in Figure 5E, PAG was highly correlated to urate, 3-ureidopropionate, indoxyl sulfate, *p*-cresol glucuronide, adenine, and phenylalanine (in the MEKC,NP mode) and to phenylacetate, 2-hydroxyphenylacetate, 4-hydroxyphenylacetate, and a coeluting peak of uric acid and PAG (in the CZE-RP mode). All of these metabolites showed relatively higher levels in the urine obtained from *S. mansoni*-infected mice.

As for the 1D plots, hippurate resulted in strong correlation to itself and to 2-oxoglutarate (2OG) (Figure 5F). All of these metabolites showed decreased urinary excretion in the infected mice group. Hippurate is predominantly formed by the conjugation of glycine with benzoate (a mitochondrial metabolite), which occurs in the mitochondria[40] and again points to the impact of the infection on the host microbiome.

In order to identify additional compounds in the CE profiles, a 2-D S-NMR-CE correlation between NMR and each CE mode was plotted independently (Figure 6A,B). Since NMR peaks of *p*-cresol glucuronide were correlated with two unknown peaks in CE, one of them subsequently confirmed as *p*-cresol glucuronide, in the normal polarity mode, it was decided to verify the structural correlations with the 2-D correlation plot, which replicated the original result. The 1-D correlation, driven by the *p*-cresol glucuronide peak at $\delta = 2.3$ in the NMR profile and in the MECK-normal polarity mode, is plotted in Figure 6C, which clearly shows the *p*-cresol glucuronide peak with a much higher correlation with signals from other *p*-cresol glucuronide proton environments and in the MECK,NP mode, than with other metabolite signals. This plot also helped to investigate putative biological correlations with other peaks; *p*-cresol glucuronide correlates to PAG (1) and indoxyl sulfate (4). These metabolites showed an obvious increase in urinary excretion in the infected mice. Once more, structural correlation was exemplified (Figure 6D), where the indoxyl sulfate signal in the NMR spectrum correlates strongly to itself and with an unknown peak in the CZE-RP mode.

The process of selecting key infection discriminatory variables in the NMR spectra has been repeated for other well-known metabolites in order to collect further evidence of the advantage of capitalizing on the exquisite capability of NMR for molecular structural identification working synergistically with CE to provide a relatively cheap tool which can be implemented and operated easily. Strong correlations between 3-ureidopropionate and *p*-cresol glucuronide and PAG and indoxylsulphate are shown in Figure 6E, whereas the correlation between the 2-oxoglutarate signal in the NMR with signals from hippurate, fumarate, citrate, and isocitrate in the CE matrix are illustrated in Figure 6F.

We have demonstrated the extension of statistical heterospectroscopy and presented a strategy for harnessing the molecular structural elucidation power of one analytical method to validate and identify metabolites profiled using another analytical platform. This method has been exemplified for a *S. mansoni* infection in a mouse model whereby new candidate biomarkers such as 3-ureidopropionate were identified and correlated with several indices of gut microbiotal disturbance. This strategy for adopting a multiplatform approach to the characterization of disease should be widely applicable across different analytical platforms and for a range of physiological and pathological conditions.

(37) Yokoyama, M. T.; Carlson, J. R. *Appl. Environ. Microbiol.* **1981**, *41*, 71–76.
(38) Elsden, S. R.; Hilton, M. G.; Waller, J. M. *Arch. Microbiol.* **1976**, *107*, 283–288.
(39) Martinelli, R.; Noblat, A. C.; Brito, E.; Rocha, H. *Kidney Int.* **1989**, *35*, 1227–1233.
(40) Gatley, S. J.; Sherratt, H. S. *Biochem. J.* **1977**, *166*, 39–47.

ACKNOWLEDGMENT

I.G.-P. acknowledges her fellowship and mobility grant from Universidad San Pablo-CEU. The authors acknowledge the funding from Ministry of Science and Technology (MCIT) CTQ2008-03779 J.V. is grateful to the Swiss National Science Foundation (Grant Nos. PPOOB-102883 and PPOOB-119129).

Received for review August 1, 2009. Accepted November 6, 2009.

AC901728W