Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2013

Impact Factor: 5.64 · DOI: 10.1021/ac403110u · Source: PubMed

7 AUTHORS, INCLUDING:

Miguel A Rodríguez
Universitat Rovira i Virgili
33 PUBLICATIONS **313** CITATIONS
SEE PROFILE

Raül Tortosa
VHIR Vall d'Hebron Research Institute
39 PUBLICATIONS **177** CITATIONS
SEE PROFILE

Antonio Julià
VHIR Vall d'Hebron Research Institute
57 PUBLICATIONS **982** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis

Arnald Alonso,[†,‡] Miguel A. Rodríguez,[§,∥] Maria Vinaixa,[§,∥] Raül Tortosa,[†] Xavier Correig,[§,∥] Antonio Julià,*[,†] and Sara Marsal[†]

[†]Rheumatology Research Group, Vall d'Hebron Hospital Research Institute, Barcelona, Spain

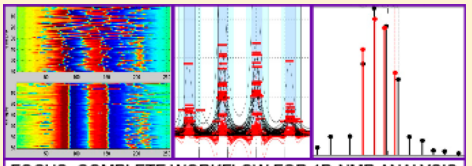
[‡]Department of ESAII, Polytechnical University of Catalonia, Barcelona, Spain

[§]Centre for Omic Sciences, COS-DEEEA-URV-IISPV, Reus, Spain

[∥]Metabolomics Platform, CIBERDEM (CIBER de Diabetes y Enfermedades Metabólicas Asociadas), Reus, Spain

Supporting Information

ABSTRACT: One-dimensional ^{1}H NMR represents one of the most commonly used analytical techniques in metabolomic studies. The increase in the number of samples analyzed as well as the technical improvements involving instrumentation and spectral acquisition demand increasingly accurate and efficient high-throughput data processing workflows. We present FOCUS, an integrated and innovative methodology that provides a complete data analysis workflow for one-dimensional NMR-based metabolomics. This tool will allow users to easily obtain a NMR peak feature matrix ready for chemometric analysis as well as metabolite identification scores for each peak that greatly simplify the biological interpretation of the results. The algorithm development has been focused on solving the critical difficulties that appear at each data processing step and that can dramatically affect the quality of the results. As well as method integration, simplicity has been one of the main objectives in FOCUS development, requiring very little user input to perform accurate peak alignment, peak picking, and metabolite identification. The new spectral alignment algorithm, RUNAS, allows peak alignment with no need of a reference spectrum, and therefore, it reduces the bias introduced by other alignment approaches. Spectral alignment has been tested against previous methodologies obtaining substantial improvements in the case of moderate or highly unaligned spectra. Metabolite identification has also been significantly improved, using the positional and correlation peak patterns in contrast to a reference metabolite panel. Furthermore, the complete workflow has been tested using NMR data sets from 60 human urine samples and 120 aqueous liver extracts, reaching a successful identification of 42 metabolites from the two data sets. The open-source software implementation of this methodology is available at http://www.urr.cat/FOCUS.



FOCUS: COMPLETE WORKFLOW FOR 1D-NMR ANALYSIS

In the past decade metabolomics has experienced an exponential growth thanks to the development and refinement of the analytical techniques used to obtain data from the metabolome.[1] At the same time, advanced bioinformatic methods have emerged in order to deal with the complexity and the high dimensionality of the data generated by these techniques. The recent advances in metabolomics, together with other omic approaches like genomics, transcriptomics, and proteomics, are increasingly leading to an integrated knowledge of systems biology and to the consequent understanding of the biologic regulatory processes that underlie metabolism and disease.[2−5]

Nuclear magnetic resonance (NMR) and chromatography-coupled mass spectrometry (MS) are the analytical techniques that have contributed to the growth of the metabolomic science. From the different NMR techniques, the one-dimensional (1D) proton spectrum (^{1}H NMR) has been the most commonly used technique in NMR-based metabolomics studies.[3] ^{1}H NMR is characterized by a simple and fast acquisition process,[6−8] providing high repeatability spectra that cover a wide range of metabolites.[7,9] However, there are several technical challenges that still need to be solved in order to make the most of this powerful analytical method. These technical challenges affect several of the data processing steps that are required prior to the use of any statistical analysis method. There is actually an intense research activity on solving these technical difficulties, especially on those that more critically affect the quality of the final data set.[6,10−17] However, while several useful methodologies have been developed for each of the processing steps, there is still a lack of a single tool that efficiently integrates the complete NMR data processing workflow.

One first challenge on the data processing workflow is how to deal with the significant biases in peak positions introduced by the sample chemical environment. Chemical variables like ionic strength, pH, and protein content differences between samples can produce changes in peak positions of certain metabolites.[9,18] This means that, even in well-addressed studies

Received: September 27, 2013
Accepted: December 19, 2013
Published: December 19, 2013

with uniform sample processing and randomized designs,[19] spectral data analysis will require the application of complex alignment tools in order to minimize the impact of confounding factors.[20] Multiple alignment methods[14,21−24] have been proposed to correct this variability in chemical shifts of NMR spectra and to guarantee peak correspondence across all the analyzed spectra. However, most current methods are still based on a reference spectrum against which each sample spectrum is aligned. This can be an important source of bias since this reference is not always representative of all the sample spectral diversity.[11] Importantly, methods based on the correlation function[14,24] to compute the alignment correction shifts can introduce other types of errors mainly derived from the presence of very large peaks that account for almost all the weight of the correlation function (despite the presence of lower peaks, where alignment is neglected).

In addition to peak position biases, another major challenge for NMR metabolomics is automatic metabolite identification. MetaboHunter[16] represents an important advance and one of the most common approaches for metabolite identification. In this method, metabolite compounds are matched against a reference spectrum peak list according to the peak positions. However, this approach can lead to a high number of false positives since it does not use intensity and correlation measurements to refine metabolite matching. Another commonly used approach based on the valid cluster concept[13,25] is an improvement with respect to MetaboHunter because it also uses peak intensities and intersample intensity correlation thresholds to group peaks. However, in this method, the scores used to match reference spectra to the data set peaks can lead to suboptimal binary identification when the lowest reference intensity peaks are not found in the sample spectra.

Here, we present FOCUS, a complete workflow for processing large spectral NMR data sets which covers spectral alignment, peak detection, peak quantification, and metabolite identification. FOCUS output provides an accurate matrix of sample peak features ready for chemometric analysis, as well as a metabolite identification scoring system that greatly facilitates peak data interpretation. Furthermore, FOCUS generates html-based result reports containing exhaustive information of each analyzed peak such as quality assessment, intensity correlation patterns, and metabolite assignations. This integrated tool is suited for the current large sample metabolomic studies and it incorporates several methodological advances both on peak alignment and metabolite identification. FOCUS alignment is based in the RUNAS (Recursive UNreferenced Alignment of Spectra) algorithm, which efficiently aligns NMR peaks while avoiding the use of a reference spectrum. With regards to metabolite identification, FOCUS provides a peak-based procedure that significantly speeds up this common and time-consuming task. For each peak, FOCUS generates a list of identification scores for each processed NMR peak, giving the researcher a powerful tool to identify the underlying metabolite.

We demonstrate the improvements of FOCUS in spectral alignment and metabolite identification with respect to other existing methods using both simulated and real (i.e., human urine and liver extract) NMR data sets. With FOCUS, researchers performing moderate to large scale NMR studies will have a complete and powerful tool to make the most of their metabolomic data.

THEORY

FOCUS processes raw 1D-NMR spectral data sets in an integrated and unsupervised manner. As depicted in Figure 1,

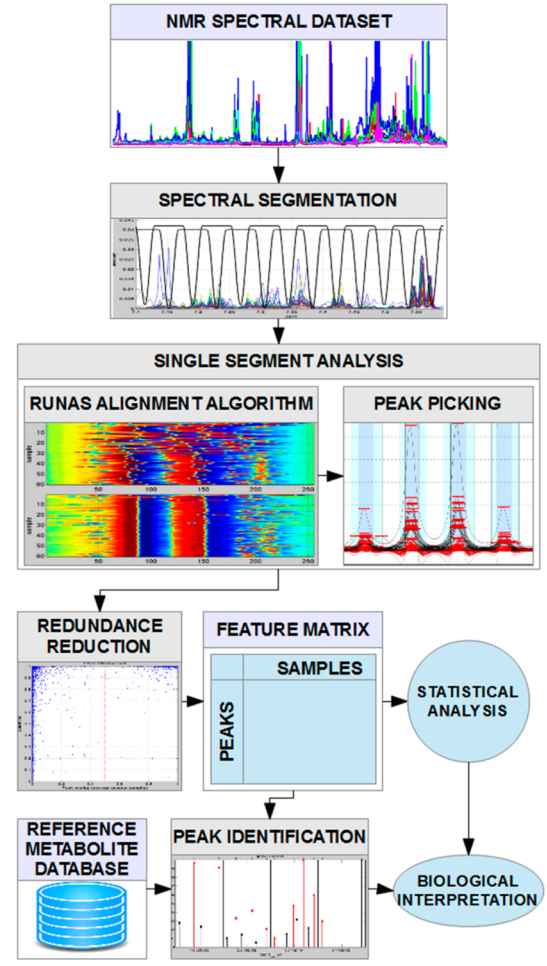


Figure 1. FOCUS workflow schema. This figure shows the FOCUS workflow analysis for processing one-dimensional NMR metabolomics spectra. It starts by computing the informative points of the spectral data set and then splitting the whole spectra in overlapping segments. For each segment, alignment and peak picking (i.e., peak detection and quantification) are independently applied. Once the segment analysis has finished, redundant peaks are removed, and a feature matrix containing peak measurements for each sample is obtained. Statistical analysis can then be performed on this matrix, and after the peak identification procedure, the results from statistical analysis can be interpreted.

FOCUS analysis pipeline consists of four main steps, namely, (i) spectral segmentation, (ii) spectral alignment, (iii) peak detection, and (iv) metabolite identification. Detailed information on each processing step can be found in the Supporting Information.

Spectral Segmentation. Raw 1D-NMR spectra are automatically divided into equally sized overlapping segments (i.e., 50% overlap) wide enough to span one or multiple peaks (see Figure S1 for further details). Segment overlapping guarantees that peaks falling into one segment end will be correctly analyzed on the consecutive segment. Furthermore, this approach also guarantees that each peak will be analyzed twice within different neighborhoods keeping the best result for

dx.doi.org/10.1021/ac403110u | Anal. Chem. 2014, 86, 1160−1169

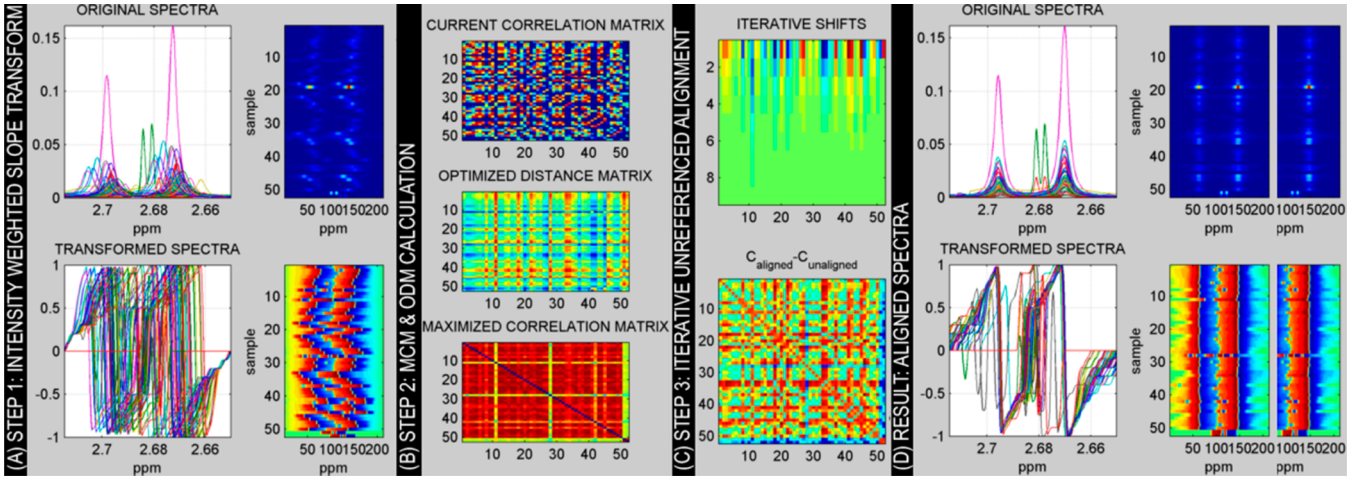


Figure 2. FOCUS spectral alignment. This figure shows the FOCUS procedure for spectral alignment. (A) Shows the spectral signal transformation (intensity weighted slope transform) and how signals and peaks are equalized across samples by using this transformation. (B) Shows the resulting MDM and ODM matrices computed before recursive alignment. (C) Shows the matrix of applied shifts (rows representing iterations and columns applied shifts on each sample) and the difference between the final and the initial spectral correlation matrices. (D) Shows the aligned spectra after applying FOCUS. It can be observed that each signal is only aligned against those signals achieving a high degree of spectral correlation.

further analysis. Subsequent data processing steps concerning alignment and peak picking are performed at the segment level.

Spectral Alignment. The FOCUS alignment method, RUNAS (Recursive UNreferenced Alignment of Spectra), performs spectral ^{1}H NMR alignment using the cross-correlation function between spectra as the alignment maximization function. In contrast to most of the currently available methodologies, RUNAS does not rely on the definition of a reference spectrum. Furthermore, the alignment procedure is based on a spectral transformation (i.e., intensity weight slope transform) that enhances peak shapes and reduces the alignment bias produced by the presence of peaks with highly unpaired intensities. RUNAS consists of several processing steps that are outlined below.

First, the intensity weight slope transform (IWST) is applied to all the input segment spectra. Briefly, this transformation extracts the sign of the spectrum derivative at each point and weights the resulting signal by its corresponding discretized intensity percentile across all the spectrum points. This transformation results in multiple advantages from a spectral alignment standpoint (see Figure 2A and Figure S2 for further details). After this first step, the RUNAS algorithm proceeds to calculate the optimal distance and the maximal correlation for each pair of samples. The optimal distance is defined as the shift that needs to be applied to one sample spectrum segment in order to maximize its correlation with respect to another sample spectrum. This optimal distance is stored in the optimized distance matrix (ODM). The correlation value between the two spectral segments at this point is called maximal correlation and is stored in the maximized correlation matrix (MCM). All these calculations are based on the fast Fourier transform (FFT)[24] and are computed only once. Figure 2B shows ODM and MCM matrices for an example segment set of spectra. Finally, RUNAS performs recursive alignment by iteratively shifting each sample spectrum in order to maximize its spectral correlation with respect to all the other samples. These shifts are calculated as the averaged optimal distances of each spectrum weighted by their corresponding maximal correlation, as detailed in eq 1:

$$\delta_x^i = \frac{\sum_{y=1}^{N_S} I(\mathrm{MCM}_{xy} \geq C_T)\cdot \mathrm{MCM}_{xy}\cdot \mathrm{ODM}_{xy}^i}{2\cdot \sum_{y=1}^{N_S} I(\mathrm{MCM}_{xy} \geq C_T)\cdot \mathrm{MCM}_{xy}}, \forall\, x \in [1, N_S] \quad (1)$$

where x is the spectrum being shifted, i is the algorithm iteration, MCM is the maximized correlation matrix, ODM the optimal distance matrix, and $I(c)$ the indicator function whose value is zero unless the comparison c is true, in which case its value is set to 1. C_T refers to a correlation threshold, so spectra that do not reach a minimal correlation with the analyzed spectrum are not taken into account in its shift computation. This threshold provides a way to automatically align spectral segments even in those cases where the samples follow more than one peak pattern. At the end of each iteration, the MDM is updated to take into account the applied shifts, and the previous step is repeated until convergence. This usually occurs after 10−20 iterations (see Figure 2C and Figure S3). Figure 2D shows how the example segment set of spectra is correctly aligned without using a reference spectrum and how groups of different spectra are independently aligned.

Peak Detection. FOCUS peak detection approach is based on the computation of a consensus peak signal (CPS) which estimates at each spectral point the frequency of samples having a peak region spanning the considered point. For each sample spectrum, the zero-crossings of the filtered spectrum (i.e., second derivative Gaussian filter) are used to delimitate peak regions on each sample. The CPS is then built by computing at each spectral point the number of samples that present a peak region spanning the considered point and scaled with respect to the number of samples. The CPS is then an estimation of the peak frequency across the samples at each spectral point. In this way, this method guarantees that all the samples will have the same contribution to the definition of peak regions. Once the CPS has been computed, it is filtered using the same previous filter, and the resulting signal zero-crossings delimitate the global peak regions (see Figure S4A for further details). Using CPS to delimitate peak regions also allows taking into account the residual misalignment that may remain after peak alignment because the residual variability on the single-spectrum peak

dx.doi.org/10.1021/ac403110u | Anal. Chem. 2014, 86, 1160−1169

regions will contribute to broaden the corresponding peak on the CPS (see Figure S4B).

Since input spectra are divided into consecutive overlapping segments, redundant peaks can be generated. In order to remove this redundancy, a peak reduction method based on the sample-averaged overlapping within peak pairs extracted from consecutive windows is applied. Peak pairs having large sample-averaged overlapping correspond to redundant peaks, and the method keeps only those peaks with larger peak shape correlations.

At this point, FOCUS provides a data matrix of per-sample peak measurements that is ready for subsequent chemometric analysis. The user can select different types of per-sample peak features, namely, (i) peak areas, (ii) peak maxima, and (iii) peak increments (difference between the peak maximum and minimum value at the peak region limits). Furthermore, three quality scores are provided for each peak, namely, (i) sample-averaged peak shape correlation (i.e., this can be used to evaluate the peak shape correspondence between samples), (ii) correlation between areas and increments (i.e., this can be useful for detecting background interferences or overlapping peaks which might be deconvoluted), and (iii) the median peak intensity percentile with respect to all the detected peaks (i.e., this quality score provides information about the relative concentration level of the metabolite that generates each peak).

Reference-Based Metabolite Identification. FOCUS reference-based metabolite identification is based on the fact that closer NMR peaks in one pure compound spectrum arise from the same proton, and they show less variation across the frequency axis than peaks generated from different protons. Consequently, given a library of known metabolites (i.e., reference spectra), FOCUS will start by clustering the reference peaks of each metabolite proton (see Figure 3A). In this way, each proton cluster will group together peaks with close positions in the spectrum (i.e., distance between peaks < $d_{cluster}$). After this clustering step, a peak-oriented identification step is performed, where the set of target metabolites for each data set peak is limited to those having a close reference peak (i.e., distance between data set peak and reference peak < $t_{cluster}$, see Figure 3B). For each data set peak and a corresponding target metabolite, FOCUS proceeds as follows. (1) Intracluster score: This step consists of identifying the metabolite peaks of the same proton. For this objective, FOCUS tries to identify correlated data set peaks (i.e., intensity correlation at the sample level) at the expected positions where the other peaks from the same proton should be located (see Figure 3C for further details). If a peak from the same proton is not identified, a zero-intensity peak is assigned to it. Finally, the correlation between the intensity levels of the correlated data set peaks and their respective reference peaks is computed in order to obtain the intracluster matching score. (2) Intercluster score: This step consists of identifying the metabolite peaks of the other reference metabolite protons (see Figure 3D). The intercluster matching score is computed as the number of protons of the target metabolite where correlated data set peaks have been found, scaled by the total number of protons. In this calculation, the weight of each proton depends on the intensity of its reference peaks (i.e., protons having relevant peaks will have more importance). (3) Penalization score: If correlated data set peaks are found outside the windows defined around the peaks of the metabolite protons, a penalization score is computed as $Sp = 1 - wN$, where N is the number of correlated peaks outside the windows and w a user defined weighting factor that establishes the degree of penalization.

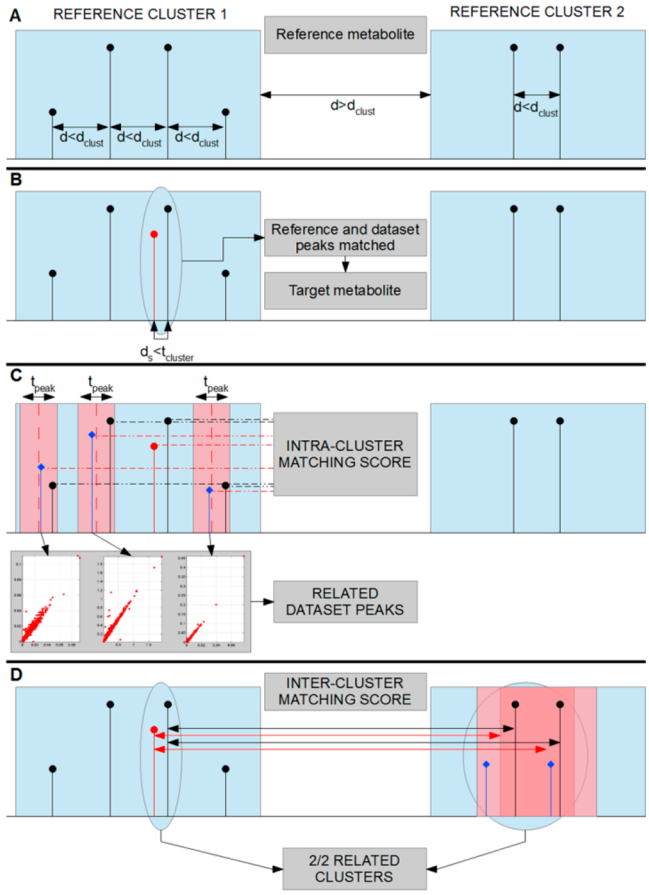


Figure 3. Metabolite identification algorithm. This figure shows the processing steps for metabolite identification. (A) The peaks of each reference metabolite spectrum are grouped in clusters (i.e., putative protons) if their distance does not exceed the maximum intracluster distance $d_{cluster}$. (B) The reference metabolite of subfigure A is a target metabolite for the data set peak (red line) because one of its peaks (black lines) is close to the data set peak (red line). (C) Intracluster matching is performed by identifying correlated data set peaks (i.e., blue lines) at the positions where the other reference proton peaks should be found (a tolerance window t_{peak} is defined around each expected position). Scatter plots show the intensities of the correlated peaks (blue peaks) against the intensity of the peak being identified (red peak). Finally, the intracluster score is defined as the correlation of the mean intensity levels of the data set peaks (blue and red peaks) versus the library-defined intensities of the target metabolite peaks (black peaks). (D) Intercluster matching is defined by the number of protons having at least a data set related peak. Search windows are defined by the reference cluster distances and expanded by a tolerance factor.

Following this procedure, FOCUS obtains, for each detected peak, a list of candidate metabolite identifications. This candidate list can be sorted by the corresponding identification scores (i.e., average of the intracluster, intercluster, and penalization scores) in order to identify the metabolite that better represents this data set peak. If a reference metabolite has only one peak, the scoring is performed by averaging the penalization score with a singlet score, which is proportional to the closeness of the reference and data set peaks. After this identification step, FOCUS also creates a putative annotated

dx.doi.org/10.1021/ac403110u | Anal. Chem. 2014, 86, 1160−1169

feature matrix, linking each peak with its top-scored metabolite identification.

Results Report Generation. In addition to the per-sample peak feature matrix, FOCUS generates an exhaustive and user-friendly results report which greatly facilitates the exploration of the obtained results. This web-based report is created locally (i.e., no need of accessing an external server) and uses JavaScript plugins to improve user navigation along the results. For each analysis, a summary report is created with the list of detected peaks characterized by their positions, their quality scores, and the best metabolite identification. A set of group identifiers is also provided, where each peak group refers to peaks that have been grouped due to their high-intensity correlations. Besides this summary report, each peak contains a link to another webpage that contains detailed information such as correlated peaks and all the metabolites that can correspond to the peak sorted by their identification scores. For further details, see Figure S5 and the example report available on the methodology webpage (http://www.urr.cat/FOCUS).

Software. The FOCUS methodology presented here has been developed using Matlab (i.e., data processing) and Python (i.e., automatic report generation). The Matlab package and a demo example can be downloaded from http://www.urr.cat/FOCUS. A Python script for creating a web-based report of the results generated by FOCUS and an example report can also be accessed from this web address.

EXPERIMENTAL SECTION

Liver Extract Data Set. Liver extract ^{1}H NMR measurements of 120 samples were conducted as previously described.[26,27] Briefly, for each hepatic sample, ∼50 mg of tissue was removed, flash frozen, and mechanically homogenized in 2 mL of H_2O/CH_3CN (1/1). Each homogenate was centrifuged at 5000*g* for 15 min at 4 °C, and the supernatant containing hydrophilic metabolites was subsequently frozen at −80 °C. For NMR measurements, the hydrophilic extracts were reconstituted in 600 μL D_2O containing 0.5 mM trisilylpropionic acid (TSP) and transferred to a 5 mm NMR tube. The 1D Nuclear Overhauser Effect Spectroscopy with a spoil gradient (NOESY) was used to record 1D ^{1}H NMR spectra using a 600.2 MHz frequency Avance III-600 Bruker spectrometer (Bruker, Germany) equipped with an inverse TCI 5 mm cryoprobe. A total of 256 transients were collected across 12 kHz spectral width at 300 K into 64 k data points, and exponential line broadening of 0.3 Hz was applied before Fourier transformation. A recycling delay time of 8s was applied between scans to ensure correct quantification. The frequency spectra were phased, baseline corrected, and then calibrated (TSP, δ = 0.0 ppm) using TopSpin software (version 2.1, Bruker, Germany).

Human Urine Data Set. Control individuals were recruited as part of the Immune-Mediated Inflammatory Disease Consortium (IMIDC)[28] repository. Urine samples were obtained from 60 healthy individuals attending to blood bank donations at university hospitals from different regions in Spain in collaboration with the Spanish National DNA Bank. Urine was collected into 15 mL universal containers with clorhidric acid. All samples were processed 24 h later and stored at −80 °C until analysis. After the samples were thawed, 400 μL of each HCl-preserved urine sample was aliquoted, and we added 200 μL of buffer phosphate (1.5 mM Na_2HPO_4/NaH_2PO_4 in D_2O with 0.62 mM of TSP, pH = 6.8) to each sample. The final solution was transferred to a 5 mm NMR tube for subsequent ^{1}H NMR acquisition. We applied the same parameters described above for spectral acquisition and processing. However, in this case, the recycling time was set to 7 s.

Metabolite Databases. When running FOCUS on the ^{1}H NMR liver extract and human urine data sets, peaks were annotated by using in-house metabolite databases (see Table S1). These databases contain the peak positions and relative intensities of the metabolite reference spectrum measured at pH = 7.5.

The database used for metabolite identification in the liver extract data set contains 31 reference metabolite spectra and is based on a list of previously manually identified metabolites.[27] The database used in the human urine data set analysis was composed of 47 reference metabolite spectra that are commonly found within this biofluid[6,29] either as endogenous/exogenous metabolites or as potential sample-handling contaminants.

Alignment Performance Evaluation. We have tested FOCUS alignment performance and compared it to two of the most commonly used alignment methods: Icoshift[14] and Correlation Optimized Warping (COW).[24] The Icoshift algorithm is based on segment-shifting by maximizing spectral correlation against a reference signal. Like FOCUS, Icoshift uses the FFT in order to speed up calculation, as proposed by Wong et al.[24] The reference spectrum is commonly computed as the average spectrum of all the sample data, and the segments in which the whole spectra are divided can be specified by the user or otherwise determined to be regularly spaced. COW is a warping alignment method based on dynamic programming that stretches or compresses segment sections in order to better match the signals to be aligned. This method is also based on a target spectrum against which sample spectra are aligned. The input data for this method consists of the segment length and the maximum warping range in the specified segment length.

In order to evaluate and compare these alignment methods, we first used simulated spectral data sets under a large range of parametric scenarios. These data sets were characterized by the presence of two and three peaks per sample, using the Lorentzian distribution[9] as the basis function for building simulated peaks (see Supporting Information and Figure S6 for details) and the different parametric scenarios have been taken into account by jointly modifying the following parameters: (a) distance between peaks, (b) standard deviation of the applied shifts to unalign sample spectra, (c) scale parameter of the Lorentzian distribution, (d) peak missingness, (e) sample size, and (f) intensity ratio between peaks. Evaluation was performed on 973 parametric scenarios for each data set considering all the possible permutations. The first metric used to evaluate alignment performance was the distance between the originally aligned spectra and the algorithmically aligned spectra correlation matrices. The second metric was the correlation between the known shifts applied to unalign the data set and the shifts derived from the algorithmic alignment. This metric was only computed for FOCUS and Icoshift because the COW approach does not use alignment shifts.

In order to also test the alignment performance over a real data set, we have also used a human urine NMR data set composed of 60 samples where spectral unalignment due to pH intersample variation is clearly visible. The performance was measured using the averaged spectra correlation[30] on different sample sizes (i.e., 10, 30, and 60 samples). A total of 48 informative segments were selected across the entire spectrum

dx.doi.org/10.1021/ac403110u | *Anal. Chem.* 2014, 86, 1160−1169

to perform this analysis in order to avoid performance evaluation over uninformative segments (see Figure S7).

RESULTS AND DISCUSSION

Alignment Performance Evaluation. Spectral alignment on the simulated spectral data sets showed a significant improvement when using FOCUS in comparison to Icoshift and COW (see Figure 4 and Table 1). Applying FOCUS

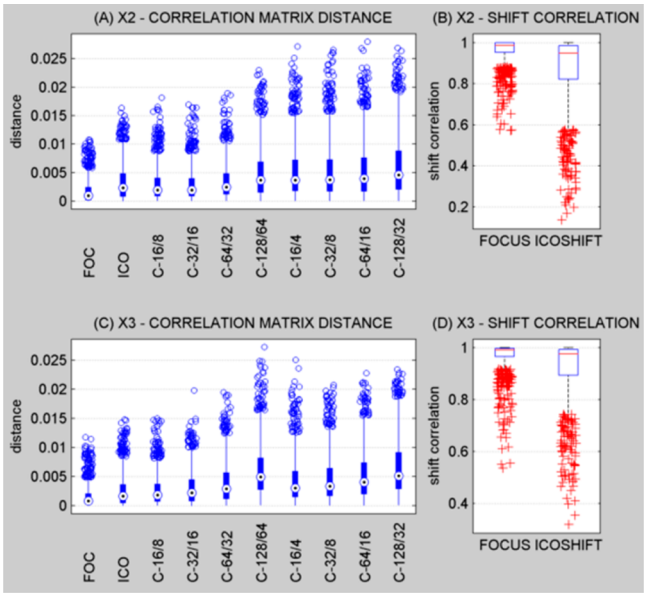


Figure 4. Simulated data set alignment results. This figure shows the alignment performance measured over the simulated data sets. (A) and (C) show the distribution of distances between real and algorithmically aligned spectral correlation matrices across the different parametric scenarios over the doublet (X2) and triplet (X3) data sets. In the same way, (B) and (D) show the distribution of correlation coefficients between the true and the algorithmically computed shifts by FOCUS and Icoshift. The COW method was evaluated using a combination of different parameters (C-[segment length]/[slack][22]).

Table 1. Performance Alignment Results[a]

	simulated data set				human urine data set
algorithm	distance (C)		C (shift)		C (spectra)
	X2	X3	X2	X3	
FOCUS	1.7×10^{-03}	1.6×10^{-03}	0.96	0.97	0.70
Icoshift	3.4×10^{-03}	2.7×10^{-03}	0.87	0.92	0.64
COW	3.0×10^{-03}	2.8×10^{-03}			0.62

[a]This table shows the most important performance measures obtained within the simulated and human urine data sets. Two simulated data sets were evaluated presenting, respectively, two (X2) or three (X3) peaks per sample. Performance measures refer to the distance between the true and the algorithmically aligned correlation matrices as well as the shift correlation between the true applied shifts and the correction shifts applied by the algorithms. Performance results on the human urine data set are based on the averaged spectra correlation.

significantly reduced the correlation matrix distance between the true aligned spectra and the algorithmic aligned spectra when compared with Icoshift (Wilcoxon test; P-value = 1×10^{-41}) and COW (Wilcoxon test; P-value = 8×10^{-38}). FOCUS improvements can also be extended to the shift correlation performance measures (see Table 1), where FOCUS improved the Icoshift alignment results both on the doublet data set (i.e., 10%) and on the triplet data set (i.e., 5%) when evaluating the distances between the expected and the obtained correlation matrices. A more detailed analysis of the results (see Figures S8, S9, and S10) shows that COW performance was more sensitive to sample shifts than FOCUS and Icoshift, as FOCUS is the algorithm with the most stable performance against changes in the degree of spectral unalignment, the distance between peaks, peak width, and sample size. The simulation analysis results showed the robustness of the FOCUS method against high degrees of unalignment, given that the correlation matrix distances only increased in 2.5% (i.e., from 1.63×10^{-3} to 1.67×10^{-3}) when doubling the standard deviation of the applied shifts to unalign the spectra. Instead, Icoshift and COW showed increases of 19.7% (i.e., from 2.71×10^{-3} to 3.24×10^{-3}) and 76.9% (i.e., from 2.11×10^{-3} to 3.74×10^{-3}), respectively. Therefore, the RUNAS alignment algorithm implemented in FOCUS shows clearly a better performance than the previous methodologies, making it particularly suitable to those data sets suffering moderate to high unalignment bias either globally or in specific spectral regions.

When evaluating spectral alignment results on the human urine spectral data set, we also found a significant performance improvement when using FOCUS (see Figure 5A and Table 1). FOCUS, Icoshift, and COW, respectively, showed average spectral correlation improvements of 53.49% (from an initial correlation of 0.46 to 0.70), 39.38% (from 0.46 to 0.64), and 35.64% (from 0.46 to 0.62) with respect to the unaligned spectral data set. Reducing the number of analyzed samples (i.e., from 60 to 30 or 10 samples) produced only moderate performance reduction in the three methods, with consistently better results with the FOCUS aligned data set. With regard to per-sample correlation results, FOCUS improvements were similar in all the spectra (see Figure S11). This confirms that FOCUS performance is superior in any sample size and does not depend on the particular sample spectra. The effects of applying FOCUS on unaligned spectral data sets are clearly visible on those segments in which the peaks are more susceptible to suffer chemical shifts. Figure 5B shows a spectral segment corresponding to hippuric acid peaks that are highly affected by unalignment. In these cases, the average spectrum clearly does not represent the true peak distributions. FOCUS superior performance is not only derived from its correlation gain but also from the increased height-to-width ratio of the resulting peaks on the averaged spectrum and from the removal of peak artifacts introduced by the other alignment methodologies (see Figure 5B).

Automated Analysis of the Human Urine Data set. FOCUS processing workflow was applied to a set of 60 human urine spectra. The unsupervised analysis was performed using the default parameter values: windows with a 50% of overlap and a length of 0.077 ppms (256 spectral data points), minimum peak width was of 0.01 ppms and minimum sample frequency to consider a peak was 10%. The only parameter that needed to be adapted was the minimum intensity increment to consider a peak which depends on the intensity scale of the analyzed spectra. This minimal parameter adaptation reflects one of the FOCUS strengths, which is its capacity to easily adapt to the singular characteristics of different NMR data sets with very few user inputs.

After the sliding window analysis, a set of 390 peaks was obtained which was reduced to 240 peaks (i.e., 38% reduction)

dx.doi.org/10.1021/ac403110u | Anal. Chem. 2014, 86, 1160−1169

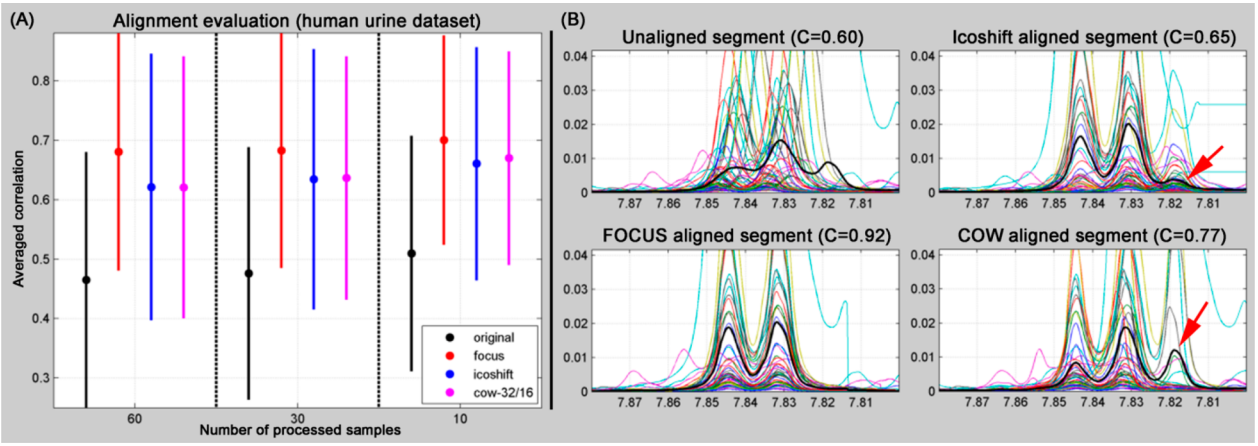


Figure 5. Human urine alignment results. This figure shows the alignment performance measured over a data set of 60 human urine NMR spectra. (A) Shows the averaged spectra correlation before (i.e., original) and after alignment (i.e., FOCUS, Icoshift, and COW). These performance measures were computed on the complete data set (60 samples) but also when the number of analyzed samples was reduced (i.e., 30 and 10). (B) Shows a spectral segment corresponding to hippuric acid peaks and the alignment results for the three algorithms tested. Black lines represent the averaged spectrum. These figures show how unalignment can introduce spurious peaks that are not solved by Icoshift and COW algorithms (red arrows).

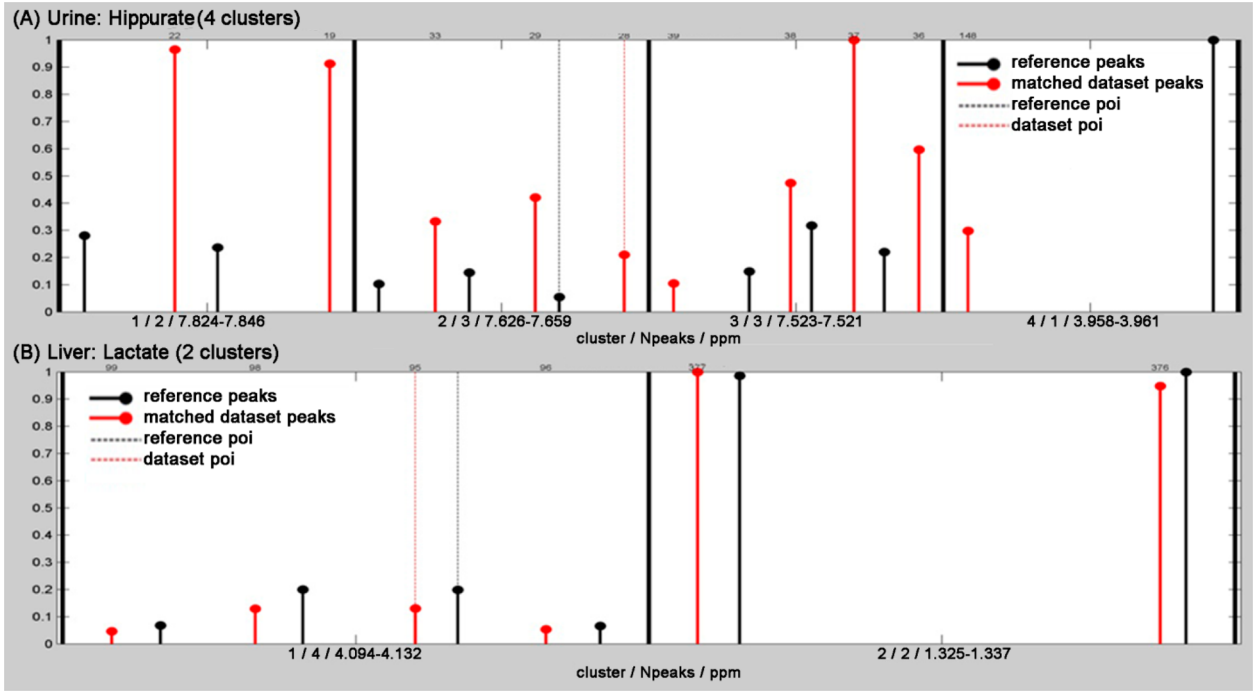


Figure 6. Metabolite identification. This figure shows two examples of successful identification of metabolites on NMR data sets using FOCUS. (A) Shows the identification results of hippurate on the human urine data set. The hippurate reference spectrum is characterized by four clusters (clusters are separated by thick black lines and reference peaks are represented with black lines ended with circles proportional to the reference peak intensities). Data set peaks matched are represented with red lines. (B) Shows the identification result of lactate on the liver extracts data set.

after applying redundancy reduction. The mean intensity correlation between the redundant peaks was 0.948, which confirms the accuracy of the redundancy reduction approach (see Figure S12). At this point, the NMR data processed by FOCUS represents a quality set of per-sample peak measurements that can be directly used to perform statistical analyses.

The correlation analysis identified 188, 168, 142, and 123 peak groups using the default grouping correlation thresholds of 0.95, 0.90, 0.85, and 0.80, respectively. Regardless of the correlation threshold used, 20 peak groups gathered more than two peaks, being clear candidates for the identification step because their peaks were highly correlated between them and uncorrelated with the remaining data set peaks. These correlation patterns between peaks are used by the FOCUS metabolite identification algorithm to assign to each data set peak the most plausible metabolite. Metabolite clusters were determined using the default maximum intracluster distance of 0.05 ppm, and the intensity correlation threshold to consider related data set peaks was set to 0.80. The tolerance window for matching reference and data set peaks and for intercluster peak matching was set to 0.03 ppm, thus allowing for a certain shift of the data set peaks with respect to the reference metabolite

dx.doi.org/10.1021/ac403110u | *Anal. Chem.* 2014, 86, 1160−1169

peaks. Finally, the peak tolerance window for intracluster matching was set to its default value (0.005 ppm) given that the distance between peaks of the same cluster does not depend on pH variations. This procedure obtained a set of 22 correct metabolite identifications that were manually verified and supported by 43 peak−metabolite associations (see Table 2 and Table S2). These identifications demonstrated the good performance of the FOCUS identification procedure. For example, peaks associated to citrate, hippurate, and trigonelline obtained large identification scores (i.e., respectively, 1.00, 1.00, and 0.97) due to the presence of correlated peaks inside each cluster and between the clusters (see the hippurate identification example in Figure 6). Creatinine is also another example of correct identification based only on the intercluster matching since reference clusters are only composed of one peak (see Figure S13). Furthermore, other identifications as lactate also obtained large identification scores (i.e., 0.94) although clusters with lower intensity peaks were not identified because FOCUS prioritizes the identification of clusters with higher intensity peaks (see Figure S13).

Automated Analysis of the Liver Extracts Data set. FOCUS processing workflow was also applied to a set of 120 liver extract NMR spectra. Like in the previous analysis, only the minimum intensity increment was changed with respect to the default values. After alignment and peak detection steps, a set of 413 peaks was obtained, which was reduced to 228 (i.e., 44.8% reduction) after the redundancy reduction step. The averaged intensity correlation between the redundant peaks was 0.995 also confirming the accuracy of the approach (see Figure S11).

The correlation analysis provided 112, 81, 53, and 36 peak groups using the default grouping correlation thresholds (0.95, 0.90, 0.85, and 0.80). The number of groups gathering more than two peaks was 15 at all the correlation thresholds, which certifies them as good candidates for the identification step. Since this data set shows a lower peak position variance across samples, the tolerance window for matching reference and data set peaks and for intercluster peak matching could be reduced down to 0.01 (in comparison to the tolerance parameter used in the human urine data set which was of 0.03). For the same reason, the grouping correlation threshold was increased up to 0.90. The metabolite identification procedure succeeded to correctly identify 20 metabolites on this data set supported by 63 peak−metabolite associations rightly matched (see Table 2). Metabolite compounds like lactate, taurine, tyrosine, and glucose achieved high identification scores due to high intracluster and intercluster matching scores (see lactate example in Figure 6). Lactate identification is a clear example of how FOCUS is able to discriminate identification of partially overlapping metabolites by detecting nonoverlapped multiplets: in this case, both lactate and threonine have an identical doublet in the 1.33 ppm range. Nevertheless, the identification of correlated lactate peaks from another multiplet (4.08 ppm) increased the lactate score in comparison to the threonine score (i.e., threonine-related multiplets were not found). Other identifications, like glutamine, are good examples of the algorithmic improvements achieved by FOCUS intracluster matching. This metabolite has a peak cluster between 2.39 and 2.54 ppm that groups 12 peaks. From these 12 peaks, only 4 peaks have been found in this data set with high-intensity correlations. This low ratio of cluster identified peaks would result in low identification scores in previous methodologies. Instead, FOCUS obtains a high identification score (i.e., 0.84)

Table 2. Metabolite Identification Results[a]

liver extract data set			urine data set		
metabolite	score	ppm	metabolite	score	ppm
creatine	1.00	3.041	citrate	1.00	2.692
glucose	1.00	4.642	creatinine	1.00	3.054
lactate	1.00	4.118	hippurate	1.00	7.846
taurine	1.00	3.437	dimethylamine	0.97	2.720
acetate	0.99	1.920	trigonelline	0.97	4.448
choline	0.97	3.206	choline	0.96	3.209
fumarate	0.97	6.521	creatine	0.96	3.939
alanine	0.95	1.489	alanine	0.95	1.496
leucine	0.95	0.978	lactate	0.94	1.347
isoleucine	0.93	1.010	ethanol	0.93	1.203
tyrosine	0.92	7.205	formate	0.93	8.468
glycerol	0.90	3.647	methanol	0.93	3.367
valine	0.90	1.041	taurine	0.92	3.418
β-hidroxybutyrate	0.88	1.208	acetone	0.88	2.241
glutamine	0.86	2.463	acetylcarnitine	0.86	3.201
creatinine	0.85	3.057	TMAO	0.85	3.288
uridine	0.85	7.871	glycine	0.83	3.577
ADP	0.80	8.233	methylguanidine	0.78	2.832
phenylalanine	0.78	7.432	pyruvate	0.74	2.348
glutamate	0.73	2.360	*cis*-aconitate	0.74	3.107
			acetate	0.68	1.931
			phenylalanine	0.66	7.430

[a]This table shows the successful metabolite identifications on the urine and the liver extracts data sets using FOCUS.

because the reference peaks that have not been identified are characterized by low intensities with respect to the identified peaks also overlapping with glutamate peaks (see Figure S14). The same can be observed on the phenylalanine cluster between 7.32 and 7.44 ppm, where only 4 of 8 peaks were identified but corresponded to the highest intensity reference peaks and showed the same intensity pattern. Importantly, the identification algorithm robustness against peak position shifts can be clearly observed in the lactate and the taurine identifications (see Figure 6 and Figure S14). Lactate peaks on the liver spectra were found with a uniform shift of 0.003 ppm with respect to the reference. The first taurine peak cluster showed a perfect matching between reference and data set peaks, while the second cluster showed a difference of 0.008 ppm.

Although multiple peaks have been unequivocally associated to one metabolite, other peaks may be associated to more than one metabolite with high identification scores. Such is the case of peak 363 at 1.754 ppm (see Table S3), which is located in the overlapping region of lysine and leucine spectra and obtains identification scores of 0.83 and 0.82 for lysine and leucine, respectively. In such cases, the exploration of the correlation patterns of the peaks of interest or the identification of nonoverlapping multiplets can help to unequivocally choose the optimal metabolite (i.e., leucine identification is reached by identifying its triplet on 0.948 ppm). FOCUS provides an interactive tool which will allow the users to easily navigate on the generated results for each peak, significantly facilitating this time-consuming task.

Discussion on General Algorithmic Performance and Analytical Technique Limitations. Our results show a good performance of FOCUS in both synthetic and real data sets, thus making it a suitable solution for untargeted profiling of large-scale metabolomic 1H-NMR spectra studies. FOCUS has

dx.doi.org/10.1021/ac403110u | Anal. Chem. 2014, 86, 1160−1169

been shown to accurately handle spectral artifacts and the most common analytical source of variances of this type of profiling study.

Automated segmentation analysis with overlapping avoids peak detection errors when the sample spectrum peaks are close to the bounds of the analyzed segment because they will be close to the middle of the following segment. Although FOCUS has demonstrated a very good performance handling misaligned spectra, in some rare occasions slight misalignment errors have been observed for specific segments containing peaks with uncorrelated shifting patterns. In such cases, FOCUS provides two complementary methods to reduce the potential impact of this limitation. First, the overlapping segmentation analysis guarantees that each peak will be analyzed twice with different neighborhoods, keeping only the best alignment result. Second, the FOCUS peak detection method based on the CPS sets the peak integration range accounting for the residual shift variability that may have not been corrected by the peak alignment method. In those rare cases where these methods are not sufficient to deal with these problems, the users can choose to set their own limits around the peak of interest to avoid the interferences of the neighboring peaks.

The FOCUS peak detection method has also shown to have a large dynamic range (i.e., 1:1000) on the analyzed data sets. This is due to the use of a frequency threshold: true signal peak positions are correlated, while noise peaks are uncorrelated so they will rarely exceed the frequency threshold. As previously commented, CPS signal has also demonstrated to be robust against residual shifts by broadening the integration area (see Figure S4B).

The identification module in FOCUS attempts to assign metabolites to NMR peaks according to a database of standard spectral references. FOCUS has been able to properly identify most of the common metabolites present in the studied NMR spectra from biological samples. Furthermore, the FOCUS identification report includes several quality control measurements aimed at facilitating the identification of those cases where metabolite identity assignment can be difficult due to the inherent technical limitations of an untargeted analysis (i.e., heavily overlapped signals).

CONCLUSIONS

The results presented here show that FOCUS NMR analysis software addresses the main problems that still affect the processing of high-throughput NMR metabolomic data, FOCUS provides an integrated workflow that performs all the necessary processing steps required to obtain a set of spectral measurements ready for the usual chemometric analyses. Importantly, the FOCUS algorithm for spectral alignment (i.e., RUNAS) has demonstrated a highly significant improvement when dealing with moderate to highly unaligned spectral data sets. This behavior has been achieved by applying an innovative approach where no reference spectra are needed and the raw spectra are mathematically transformed to maximize peak alignment and minimize outlier artifacts. Furthermore, the peak detection method also avoids the bias produced by outlier samples and bases its detection on the peak-sample frequency. Additionally, FOCUS also provides a new and efficient method for metabolite identification, facilitating this time-consuming task.

In order to demonstrate the usefulness of FOCUS, we have analyzed two spectral NMR data sets obtained from 60 human urine samples and 120 liver extracts, as well as an exhaustive spectral data simulation under a large number of parametric scenarios. According to our results, the FOCUS methodology is an optimal data processing workflow for 1D-NMR analysis, and its accuracy and efficiency makes it suitable for the forthcoming large-scale metabolomic studies.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: toni.julia@vhir.org. Fax: +34934034510.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the by the Spanish Ministry of Economy and Competitiveness grants (PI12/01362 and IPT-010000-2010-36) and by the AGAUR FI grant (2013/00974).

REFERENCES

(1) Zhang, A.; Sun, H.; Wang, P.; Han, Y.; Wang, X. *Analyst.* **2012**, *137*, 293−300.
(2) Collino, S.; Martin, F. P.; Rezzi, S. *Br. J. Clin. Pharmacol.* **2013**, *75*, 619−629.
(3) Emwas, A.-H.; Salek, R.; Griffin, J.; Merzaban, J. *Metabolomics* **2013**, 1−25.
(4) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−171.
(5) Goodacre, R. *Metabolomics* **2010**, *6*, 1−2.
(6) Da Silva, L.; Godejohann, M.; Martin, F.-P. J.; Collino, S.; Bürkle, A.; Moreno-Villanueva, M.; Bernhardt, J.; Toussaint, O.; Grubeck-Loebenstein, B.; Gonos, E. S.; Sikora, E.; Grune, T.; Breusing, N.; Franceschi, C.; Hervonen, A.; Spraul, M.; Moco, S. *Anal. Chem.* **2013**, *85*, 5801−5809.
(7) Lauridsen, M.; Hansen, S. H.; Jaroszewski, J. W.; Cornett, C. *Anal. Chem.* **2007**, *79*, 1181−1186.
(8) Wu, H.; Southam, A. D.; Hines, A.; Viant, M. R. *Anal. Biochem.* **2008**, *372*, 204−212.
(9) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430−4442.
(10) Liebeke, M.; Hao, J.; Ebbels, T. M. D.; Bundy, J. G. *Anal. Chem.* **2013**, *85*, 4605−4612.
(11) MacKinnon, N.; Ge, W.; Khan, A. P.; Somashekar, B. S.; Tripathi, P.; Siddiqui, J.; Wei, J. T.; Chinnaiyan, A. M.; Rajendiran, T. M.; Ramamoorthy, A. *Anal. Chem.* **2012**, *84*, 5372−5379.
(12) MacKinnon, N.; Somashekar, B. S.; Tripathi, P.; Ge, W.; Rajendiran, T. M.; Chinnaiyan, A. M.; Ramamoorthy, A. *J. Magn. Reson.* **2013**, *226*, 93−99.
(13) Mercier, P.; Lewis, M.; Chang, D.; Baker, D.; Wishart, D. *J. Biomol. NMR* **2011**, *49*, 307−323.
(14) Savorani, F.; Tomasi, G.; Engelsen, S. B. *J. Magn. Reson.* **2010**, *202*, 190−202.
(15) Smolinska, A.; Blanchet, L.; Buydens, L. M. C.; Wijmenga, S. S. *Anal. Chim. Acta* **2012**, *750*, 82−97.
(16) Tulpan, D.; Leger, S.; Belliveau, L.; Culf, A.; Cuperlovic-Culf, M. *BMC Bioinformatics* **2011**, *12*, 400.
(17) Zhang, S.; Nagana Gowda, G. A.; Ye, T.; Raftery, D. *Analyst* **2010**, *135*, 1490−1498.
(18) Xiao, C.; Hao, F.; Qin, X.; Wang, Y.; Tang, H. *Analyst* **2009**, *134*, 916−25.
(19) Ransohoff, D. F. *Nat. Rev. Cancer.* **2005**, *5*, 142−149.
(20) Goodacre, R.; Broadhurst, D.; Smilde, A.; Kristal, B.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; Ebbels,

T.; Kell, D.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjöström, M.; Trygg, J.; Wulfert, F. *Metabolomics* **2007**, *3*, 231−241.
(21) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. *AIChE J.* **1998**, *44*, 864−875.
(22) Tomasi, G.; van den Berg, F.; Andersson, C. *J. Chemom.* **2004**, *18*, 231−241.
(23) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* **2008**, *81*, 56−66.
(24) Wong, J. W. H.; Durante, C.; Cartwright, H. M. *Anal. Chem.* **2005**, *77*, 5655−5661.
(25) Jacob, D.; Deborde, C.; Moing, A. *Anal. Bioanal. Chem.* **2013**, *405*, 5049−5061.
(26) Samino, S.; Revuelta-Cervantes, J.; Vinaixa, M.; Rodríguez, M. Á.; Valverde, Á. M.; Correig, X. *Biochimie* **2013**, *95*, 808−816.
(27) Vinaixa, M.; Ángel Rodríguez, M.; Rull, A.; Beltrán, R.; Bladé, C.; Brezmes, J.; Cañellas, N.; Joven, J.; Correig, X. *J. Proteome Res.* **2010**, *9*, 2527−2538.
(28) Julià, A.; Domènech, E.; Ricart, E.; Tortosa, R.; García-Sánchez, V.; Gisbert, J. P.; Nos Mateu, P.; Gutiérrez, A.; Gomollón, F.; Mendoza, J. L.; Garcia-Planella, E.; Barreiro-de Acosta, M.; Muñoz, F.; Vera, M.; Saro, C.; Esteve, M.; Andreu, M.; Alonso, A.; López-Lasanta, M.; Codó, L.; Gelpí, J. L.; García-Montero, A. C.; Bertranpetit, J.; Absher, D.; Panés, J.; Marsal, S. *Gut* **2012**, *62*, 1440−1445.
(29) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; Dame, Z. T.; Poelzer, J.; Huynh, J.; Yallou, F. S.; Psychogios, N.; Dong, E.; Bogumil, R.; Roehring, C.; Wishart, D. S. *PLoS ONE* **2013**, *8*, e73076.
(30) Forshed, J.; Torgrip, R. J. O.; Åberg, K. M.; Karlberg, B.; Lindberg, J.; Jacobsson, S. P. *J. Pharm. Biomed. Anal.* **2005**, *38*, 824−832.