Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 392−401

pubs.acs.org/jpr

# Fast Metabolite Identification in Nuclear Magnetic Resonance Metabolomic Studies: Statistical Peak Sorting and Peak Overlap Detection for More Reliable Database Queries

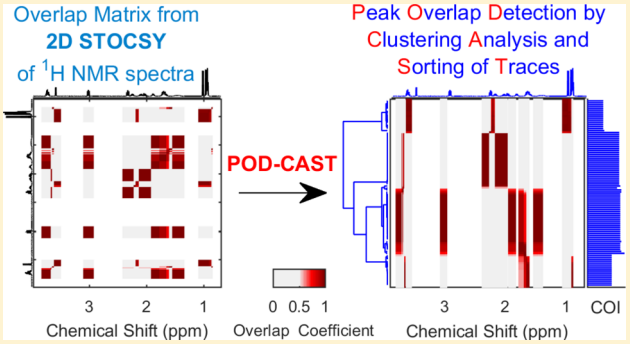
Pablo A. Hoijemberg*,†,‡ and István Pelczer†

†Department of Chemistry, Frick Chemistry Laboratory, Princeton University, Princeton, New Jersey 08544, United States

‡NMR Group, Centro de Investigaciones en Bionanociencias, CIBION-CONICET, Polo Científico Tecnológico, 1425 Ciudad Autónoma de Buenos Aires, Argentina

Supporting Information

**ABSTRACT:** A lot of time is spent by researchers in the identification of metabolites in NMR-based metabolomic studies. The usual metabolite identification starts employing public or commercial databases to match chemical shifts thought to belong to a given compound. Statistical total correlation spectroscopy (STOCSY), in use for more than a decade, speeds the process by finding statistical correlations among peaks, being able to create a better peak list as input for the database query. However, the (normally not automated) analysis becomes challenging due to the intrinsic issue of peak overlap, where correlations of more than one compound appear in the STOCSY trace. Here we present a fully automated methodology that analyzes all STOCSY traces at once (every peak is chosen as driver peak) and overcomes the peak overlap obstacle. Peak overlap detection by clustering analysis and sorting of traces (POD-CAST) first creates an overlap matrix from the STOCSY traces, then clusters the overlap traces based on their similarity and finally calculates a cumulative overlap index (COI) to account for both strong and intermediate correlations. This information is gathered in one plot to help the user identify the groups of peaks that would belong to a single molecule and perform a more reliable database query. The simultaneous examination of all traces reduces the time of analysis, compared to viewing STOCSY traces by pairs or small groups, and condenses the redundant information in the 2D STOCSY matrix into bands containing similar traces. The COI helps in the detection of overlapping peaks, which can be added to the peak list from another cross-correlated band. POD-CAST overcomes the generally overlooked and underestimated presence of overlapping peaks and it detects them to include them in the search of all compounds contributing to the peak overlap, enabling the user to accelerate the metabolite identification process with more successful database queries and searching all tentative compounds in the sample set.




**KEYWORDS:** *NMR, STOCSY, metabolomics, metabolite, identification, database, correlation matrix, overlap, complex mixture, clustering*

The identification of metabolites is one of the main goals in a metabolomics study, for example, in the search of candidate biomarkers for a given disease.[1] Once a sample set is analyzed, the relevant metabolites to a given class separation are sought to be identified with the purpose of understanding the systems biology of the subject of study, identifying the pathways affected by a disease, etc. In the future, the scientific community should be able to tackle the disease by designing a drug targeting the affected pathways, or using that knowledge of the system for disease diagnostics, prognostics, or disease evolution.[2−5]

NMR has been used for over two decades as one of the preferred analytical platforms for metabolomics studies. It allows the detection of metabolites in the micromolar range, and above, as well as presenting the great advantage of being intrinsically quantitative. It is also highly reproducible, robust, nondestructive, and needs little or no sample preparation.

STOCSY, standing for statistical total correlation spectroscopy,[6] was published more than a decade ago and its use helps in the identification metabolites as well as for pathway connectivity and biological information recovery. Without employing any correlation analysis, the process of identification of metabolites would be even much slower, relying only on the information contained in the set of 2D correlation spectra (both homo- and heteronuclear), normally acquired in addition to the standard 1D $^1$H NMR spectra. After selecting a driver peak, the STOCSY analysis shows peaks that belong to a given metabolite (high correlation between them and the selected

**Received:** August 31, 2017
**Published:** November 14, 2017





DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

driver peak), and a list of these peaks can be used as input for a query in any of the publicly available databases such as the Human Metabolome Database used in the present work (HMDB)[7] or others.[8]

A related approach aiming to the generation of peak lists for database search was developed by Raftery and co-workers and named RANSY, standing for ratio analysis NMR spectroscopy.[9] Instead of the correlation among peaks from the same compound, as in STOCSY, the latter is based on the constant ratio among peaks from the spectrum of a given molecule, which depends for each peak on the number of magnetically nonequivalent spins and is constant across all spectra in the set (regardless of the concentration in each sample). Further confirmation with 2D correlation spectra is crucial as well as final confirmation with spiking experiments. It also helps in identifying pathway connectivity, as those peaks for other metabolites can present weaker positive correlations to the driver peak or even negative correlations to it.

The average usage of STOCSY involves the generation of at least one pseudo NMR spectrum, or STOCSY trace, for each metabolite of interest, with the aim of identifying correlations from all peaks in the spectral set to a chosen driver peak. The procedure is usually performed in a trace by trace basis, making it a tedious but necessary work. Several variations or adaptations of STOCSY where proposed years after its initial publication, but the objectives of those are mainly focused on data preprocessing and the assessment of pathway connectivity.[10]

Correlation of peaks to a driver peak can be originated not only by structural (same molecule) correlation, but also by pathway connectivities. iSTOCSY (iterative STOCSY)[11] is an automated algorithm that aims to distinguish one from the other, once a driver peak is chosen, and was devised to reveal pathway connectivities. Overlapping of peaks, which is common in $^{1}$H spectra of biological fluids, could lead to masking of some peaks by others with more correlation or bigger intensities, thus showing a reduced correlation in the STOCSY trace. STORM,[12] another algorithm from the group at Imperial College London, aims to solve the overlap peak issue by selecting a subset of spectra that do not have peak overlap in a given region of interest, thus improving the obtained correlations (now without contribution from the overlapping resonances). Together with RED-STORM,[13] its expanded version for bidimensional spectra, these are the only algorithms that work with selected subsets of spectra. Nonetheless, STORM relies on the absence of overlap in a given subset of spectra, which does not always occur, or at least not for all resonances with peak overlap in the whole range of frequencies.

Grouping resonances within clusters is needed to generate peak lists for the database query. Edison and co-workers reported a simple approach to generate peak lists that works nicely with $^{13}$C NMR data[14] but would certainly be harder to apply in $^{1}$H NMR spectra. CLASSY,[15] another development from the group at Imperial College London, which aims at improving biological information recovery, creates clusters based on a binary connectivity, linking chemical shifts from driver peaks that possess a similar number connectivity number, referred to as nodes. However, it does not specify precisely how it distinguishes among metabolites with similar number of nodes, and moreover how it deals with overlapping resonances. Finally, R-STOCSY[16] creates clusters of consecutive resonances that stand out based on a covariance/correlation ratio landscape, although this can be peaks on a multiplet as well as biologically linked metabolites. Improvements to the latter include OR-STOCSY,[17] which allows for better recovery of biological information by applying a filter from OPLS-DA output, and the application of statistical recoupling of variables to obtain a 2D pseudospectrum that can link clusters without the need for vicinity.[18] This linking, as before, includes both structural and biological connectivities.

This work presents a simple methodology that shortens the time employed for the identification of metabolites, not limiting it to those of biological interest to the classification (as a product of the statistical analysis on the acquired spectra), but rather to the whole set of compounds detected in the NMR spectra. It is based on statistical analysis over the 2D matrix, which contains traces that are similar to one another, in this case the 2D STOCSY matrix. Peak overlap detection by clustering analysis and sorting of traces (POD-CAST) clusters the traces that look similar, while it helps dealing with those peaks that present overlap. Its output is composed of a visually appealing figure with easy to read compressed data as well as a list of sorted driver peaks for easier search in databases. The graphical output can be adjusted by varying an overlap threshold value, to differentiate structural correlation from decreased correlations due to overlap (although it cannot eliminate strong correlations due to pathway connectivity, intrinsic to STOCSY). Peak picking for driver peaks can be applied almost automatically or through a list of manually picked peaks. The advantages of POD-CAST were compared to the DemixC approach,[19,20] currently one of the best available methodologies for assessing complex mixtures (designed to work on TOCSY experiment data, but easily adaptable to data from STOCSY analysis). POD-CAST can be also applied over data sets where similar traces are to be identified, for example, TOCSY spectra.

## METHODS

### Simulated Spectra Set

Raw free induction decay files corresponding to 1D NOESY with presaturation $^{1}$H NMR experiments were downloaded from the Human Metabolome Database[7] for glutamic acid, leucine, lysine, and valine (HMDB ID numbers: HMDB00148, HMDB00687, HMDB00182, and HMDB00883, respectively). All four spectra were individually phase and baseline corrected using Mnova (v. 10.0) software (Mestrelab Research S.L., Santiago de Compostela, Spain). The four spectra were normalized to an equal (arbitrary) area for the integration of the $\alpha$ protons between 3.57 and 3.77 ppm, yielding an equimolar relation to one another. Twenty spectra were created by arithmetic combinations of the processed spectra using randomly generated coefficients ranging from 0.8 to 1.2 for each amino acid spectrum to form a set of 20 spectra, with their compositions independent to one another. The spectra were superimposed, several regions for peaks arising from small impurities were selected as "blind" to set the intensities within them to zero (0.80−0.85, 1.05−1.35, 2.85−2.95, and 3.35−3.40 ppm), and then the region from 0.7 to 3.9 ppm was exported for statistical analysis.

### Sample Preparation of "Artificial Mixture Set"

Twenty samples were prepared containing variable volumes of stock solutions in $D_2O$ (D, 99.9%, Cambridge Isotopes Laboratories, Inc., Andover, MA, USA) with ∼500 $\mu$M sodium azide (>99.0%, Sigma-Aldrich) of 11 compounds: arginine, ascorbic acid, choline chloride, citrulline, folic acid, glutamic


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

acid, histidine, ornithine hydrochloride, pantothenic acid (calcium salt), taurine, and tryptophan (all reagents were >99.0%, Sigma-Aldrich). The stock solutions were prepared as follows: a mass between 0.17 and 0.25 g was weighed, and the solid was added to 1.0 mL of the sodium azide solution in a conical microtube. After vortexing for 10 s to help dissolution (there was no effort made to aid dissolution further in cases of slow dissolution), the microtubes were centrifuged at 3200 rpm for 5 min on an Adams Compact II Centrifuge (Becton Dickinson and Company, Sparks, MD, USA). The supernatants were collected with glass Pasteur pipettes, with the exception of the solutions of choline chloride and ornithine hydrochloride, as both had all the solid dissolved after vortexing. The next step involved the addition of random volumes (between 35 and 50 $\mu$L) from each of the stock solutions into 20 microtubes to generate 20 samples that were composition independent. The final step consisted of completing 600 $\mu$L on every sample using the sodium azide solution. From each of the 20 samples, 50 $\mu$L was diluted to 550 $\mu$L using the sodium azide $D_2O$ solution and transferred into UP5−7 5 mm OD NMR tubes (New Era Enterprises, Inc., Vineland, NJ, USA) for the acquisition of $^1$H NMR spectra.

## NMR Data Acquisition

The 1D $^1$H NMR spectral data on all diluted samples and 2D experiments (COSY, HMBC and HSQC) on one concentrated sample were collected at 295 K on a Bruker Avance-III spectrometer equipped with an Ultrashield magnet at 500 MHz using a TCI ($^1$H/$^{13}$C/$^{15}$N//$^2$H) CryoProbe optimized for $^1$H, increasing sensitivity relative to traditional room temperature probes,[21] under the control of TopSpin software (v. 3.2), all products of Bruker-Biospin (Billerica, MA, USA). 1D $^1$H NMR spectra were acquired using a home-refined version of the excitation sculpting method incorporated with the water signal suppression.[22] A $^1$H−$^1$H 2D TOCSY experiment of one of the concentrated samples was performed on a Bruker Avance-III HD spectrometer equipped with an Ultrashield magnet at 800 MHz using a QCI ($^1$H/$^{19}$F−$^{13}$C/$^{15}$N//$^2$H) CryoProbe.

## NMR Data Processing

The 1D $^1$H NMR spectra were processed using Mnova, starting with a one-time zero filling and a 1 Hz Gaussian function apodization, followed by phasing and baseline corrections. The spectra were referenced to the choline methyls proton resonance at $\delta$ 3.19 ppm (singlet).[7] The spectra contained over 200 peaks each, and despite possible small pH differences among the samples, there was great consistency in the position of the peaks throughout the set. The peak shifts variations were corrected to within 0.001−0.002 ppm by local peak alignment, performed manually in Mnova for all spectra of the data set simultaneously. Spectra used for the analysis had the empty regions on the extremes stripped off and the intensities around 4.78 ppm, the water peak, were suppressed to zero, before being exported for statistical analysis. The spectra obtained were clean, free from residual artifacts and identified contaminants.

## Statistical Analysis

The STOCSY analysis[6] is normally performed by choosing a driver peak, whose correlations are shown in a color coded trace over a pseudo 1D NMR spectrum. Each point on the trace (its shape) is defined by the covariance of the given chemical shift to the driver peak in the whole spectral sample set, while color coding is used to show the magnitude of the correlation coefficient (in some cases the absolute value of it is shown instead). This correlation arises from the collinearity of the driver peak chemical shift to the whole range of the spectra, variables, over the spectra from all the samples, observations.

POD-CAST analysis was performed using a script on Matlab R2014b (Mathworks, Natick, MA) over STOCSY traces. STOCSY analysis was performed on the data sets imported from MS Excel using a script on Matlab. This script, kindly provided by researchers at Imperial College London, was refined to produce the 1D STOCSY traces corresponding to each driver peak (given a peak list or selecting all peaks above a chosen threshold), and a 2D matrix containing the correlation information from every trace, that is to say, the information found on the color coding of the STOCSY traces.

The first step was to generate a square matrix containing information related only to the driver peaks, as there is no interest in correlations of the driver peaks to the noise areas. To that purpose, the overlap matrix **O** was obtained in which each overlap coefficient $O_{ij}$ is the inner product of the correlation traces $i$ and $j$. Since $O_{ij}$ would have the maximum overlap value for traces where $i = j$, each trace was then normalized to that value, which in the end renders matrix **O** with all diagonal values equal to 1. Later on, a binary connectivity matrix **B** was determined from matrix **O**, such that elements $B_{ij}$ are equal to 0 if $O_{ij}$ is equal or smaller than a chosen threshold, and equal to 1 if $O_{ij}$ is greater than it (the default threshold is 0.5). Then a cumulative overlap index (COI) was obtained for each row $i$ of matrix **B**, calculated by adding all $j$ elements in row $i$. Matrix **B** shares some similarity to the connectivity matrix obtained for CLASSY,[15] while the cumulative overlap index is related to the importance index calculated in DemixC.[19,20]

The presence of negative correlations has a wealth of information in STOCSY analysis, as it is normally linked to biochemically meaningful correlations.[6] Note that any $O_{ij}$ element that had a negative value is also converted to zero in matrix **B** for the calculation of the COI values. In the results of the present work, the negative correlations are purely coincidental (and are thus not shown in any overlap matrix plot), as the STOCSY matrices derived from the analysis over a set of 20 simulated spectra, in the first case, and from 20 artificial mixtures of independent compositions, in the second case.

The following step was the application of a hierarchical clustering analysis[23] (HCA) to matrix **O** using a Euclidean distance measurement for the metrics with the Ward type of linkage, which rendered appropriate results. The type of linkage serves the purpose of visualization by creating a dendrogram, in which the closest traces appear together in branched leaves, the height of which is a measurement of the distance between the traces. The longer the distance, the more dissimilar are the traces. Leaves form then branches and all these are linked based on the linkage type. The resulting dendrogram has the chemical shifts of the driver peaks as labels for the leaves.

The last step consists in sorting the traces of matrix **O** by using the order of the leaves in the dendrogram in the vertical dimension of the matrix to produce a vertically sorted matrix, $\mathbf{O}_v$. This matrix would have dissimilar traces far from one another, and traces whose distances are smaller among them close together, forming bands. These bands, plotted in a color coded fashion with different shades of red above the chosen threshold for the construction of matrix **B**, would be as wide as the number of peaks of the compound comprising it (if there is no peak overlap at all, overlap cases are discussed below). To


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

enhance visualization, the dendrogram is placed next to matrix $\mathbf{O}_v$, showing a match between branches in the dendrogram and bands in the sorted matrix, and an average spectrum is placed on top of the sorted matrix for guidance. Traces with overlapping peaks will share overlap (weaker correlation than for pure compounds) with peaks from other bands.

Matrix $\mathbf{O}_v$ can be sorted even further, in the horizontal dimension with the same order given by the dendrogram leaves, rendering a diagonal block type of matrix, $\mathbf{O}_{vh}$. Any overlap detected in matrix $\mathbf{O}_v$ will appear as an off-block rectangle, which can appear contiguous to the diagonal block (for overlapping peaks between neighbor bands), or off-diagonal for overlapping peaks in non-neighboring bands, as it will be seen below. The dendrogram is displayed next to matrix $\mathbf{O}_{vh}$, now in both axes.

The COI is added as a label for each driver peak trace, and as a bar plot on the other side of matrices $\mathbf{O}_v$ or $\mathbf{O}_{vh}$, and it is an indication of the number of peaks with appreciable or substantial connectivity to the driver peak, as it does not distinguish between $O_{ij}$ values just above the threshold or close to unity. As it is discussed below, overlap regions are easily detected when a trace contains differing COI numbers from the most frequent values within the band, or when the COI exceeds the leaf count for the band. These overlapping traces tend to appear in the contacts between bands, but may also appear in other bands, oftentimes due to the existence of overlap with more than two compounds for the given driver peak. The default threshold value of 0.5 for the COI creation from matrix $\mathbf{B}$ can be increased gradually, to cause the disappearance of intermediate overlap coefficient values. This will lead to a COI distribution where overlapping contribution among peaks is lost, the COI values will be reduced and most peaks adding to the COI will be nonoverlapping peaks.

### Database Query

The sorted list of chemical shifts was split into groups, each one composed of the driver peaks of a given band. All lists were used separately as input for the 1D $^1$H NMR search in the Human Metabolome Database (peak tolerance was set to 0.02 ppm). The hit list provided by the database was analyzed and the spectra of the candidate metabolites was compared to the correlation trace for matching (either the 1D STOCSY trace or the correlation trace in the 2D matrix). In cases where neighbor bands showed a high degree of overlap, with a few traces identified by a COI value higher than the most frequent for each neighbor band, the extra driver peaks were also incorporated to the other group to check for improvement in the matching index on the query. Once the assignment of the group of peaks to the given candidate was defined, through 2D NMR correlation experiments for additional confirmation, all (or most) driver peaks were assigned to any given compound (or more than one compound if there was overlap).

## RESULTS AND DISCUSSION

Figure 1A shows the stacked spectra from HMDB of all four compounds in the simulated spectra set, glutamic acid, leucine, lysine, and valine. Valine and leucine have isopropyl groups and thus methyl groups. The number of methylenes varies among the four amino acids: no methylene in valine, one in leucine, three in glutamic acid, and four in lysine. The structural similarity present has a consequence, there is strong overlap of peaks from methylene and methyne groups on the spectra of these four compounds (the methyl groups have peaks that barely overlap around 0.965 ppm). The overlapping of peaks is an issue itself when performing a STOCSY analysis. The correlation is mixed if the chemical shift of the driver peak corresponds to peaks from two or more compounds, and the resulting STOCSY trace normally has contributions from correlations to the peaks of all the compounds involved, unless the contributions to the overlapping peak are uneven and the correlations corresponding to the smaller contributor are more obscured.

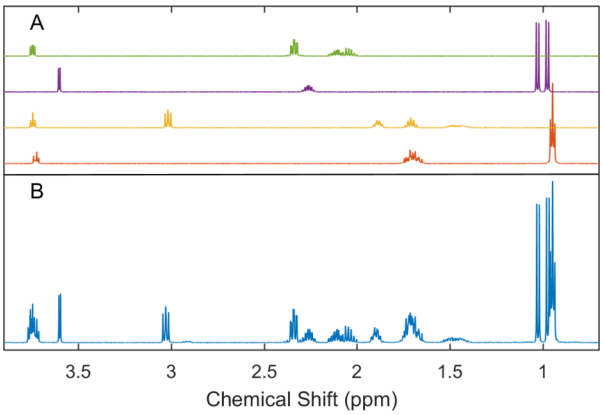



**Figure 1.** (A) Stacked $^1$H NMR spectra from the Human Metabolome Database for glutamic acid, valine, lysine, and leucine, from top to bottom, normalized to an equal integral value for the $\alpha$ $^1$H resonance. (B) Average spectrum of the 20 simulated $^1$H NMR spectra from linear combination of spectra from glutamic acid, valine, leucine, and lysine with randomly assigned coefficients between 0.8 and 1.2 for each.

Figure 1B presents the average of the 20 simulated spectra for the "mixtures" of glutamic acid, leucine, lysine, and valine. This simulated spectra set has the advantage that all the peaks are aligned within it. Although it is not what normally happens with real mixtures, it serves the good purpose of showing the results of POD-CAST on a simple data set. The overlaps of peaks mentioned above are in fact useful to show the potential of POD-CAST toward their quick detection.

Figure S1 in the Supporting Information shows the stacked STOCSY traces for driver peaks chosen at: 0.948, 1.021, 2.310, 3.590, and 3.718 ppm, A, B, C, D, and E, respectively. It can be clearly seen that traces A and E match as well as traces B and D (trace C does not match to either). It is obvious that STOCSY produces redundant information, as its basis is the correlation of one peak to another or others, and the correlation is reciprocal. The key is to take advantage of this redundancy in a way that the outcome results in a faster metabolite identification process. The first step toward automation resides in selecting a complete set of driver peaks. Although it could be arguable in metabolomics that there is no need to identify all the metabolites in a given sample set and the focus should be put only on those metabolites that end up being significantly different between two cohorts, there is an additional benefit of performing the process with all the peaks on the spectrum. Regardless of the possible identification of all the metabolites, accounting for all the peaks will provide a good estimation of the total amount of metabolites being detected.

A 2D STOCSY matrix was obtained using all the peaks (94) from an average spectrum of the whole data set as driver peaks. This rectangular matrix was reduced to the square matrix $\mathbf{O}$, seen in Figure 2, showing only cross correlation values in color coding that represent the normalized inner product between


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

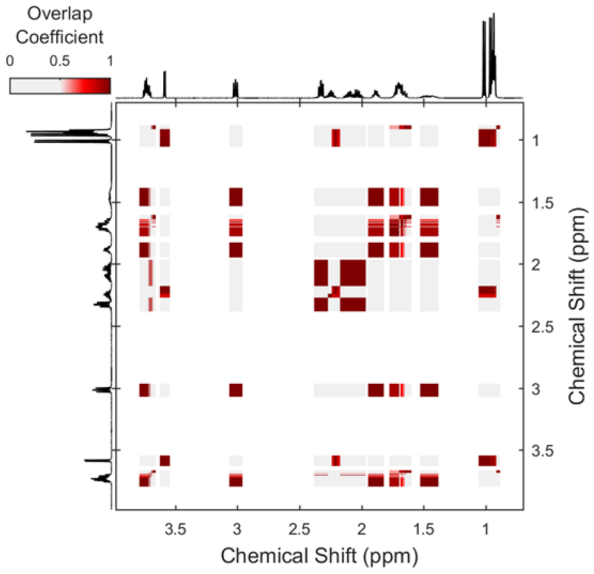



**Figure 2.** Overlap matrix **O** resulting from the normalized inner of all pairwise combinations of the correlation traces for every driver peak, obtained after performing a STOCSY analysis over the simulated spectra set. Average spectrum is overlaid on each axis. White background corresponds to regions without (driver) peaks.

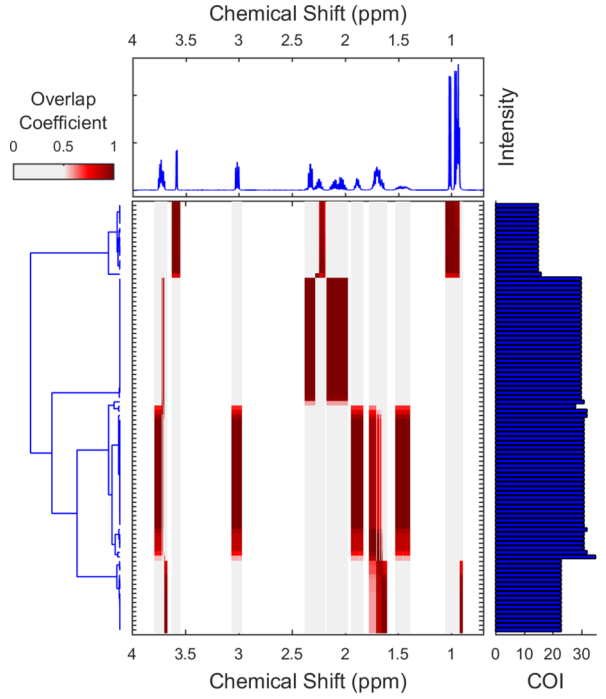



**Figure 3.** POD-CAST on simulated spectra set. (Top) Average spectrum. (Left) Dendrogram obtained by HCA over the traces from the overlap matrix of the 2D STOCSY traces using automatic peak picking. (Center) Overlap matrix **O** (Figure 2) sorted by the order of the driver peaks obtained after HCA and dendrogram generation. (Right) Cumulative overlap index indicating the number of peaks with overlap coefficient values above the chosen threshold.

the traces of the two driver peaks. The comparison of the traces for the driver peaks chosen in Figure S1 will show the information redundancy discussed. Moreover, it can be easy to identify several other traces that would match any of the ones chosen above. Despite being a square matrix of reduced size, compared to a full size 2D STOCSY matrix, the visualization on the actual chemical shift scale helps linking the peaks to one another through the cross peaks, as in any typical homonuclear 2D correlation experiment.

The rest of the steps in POD-CAST were followed to produce Figure 3 (an expanded version showing the labels is included in the Supporting Information in Figure S2). The traces in matrix **O** were clustered to produce the dendrogram depicted in Figure 3 (left side) and a sorted list of chemical shifts as labels for the leaves, yielding branches that group these driver peaks (as detailed in Figure S2, left side). The reordered list was then used to sort matrix **O** on the vertical dimension, creating matrix $\mathbf{O}_v$ and grouping the resembling traces into bands (Figure 3 and Figure S2, center), while conserving the horizontal scale for the traces to be comparable to the average $^1$H NMR spectrum for the simulated set (Figure 3 and Figure S2, top). The bar plots on the right side of Figure 3 and Figure S2 show the COI value for each trace. As expected, there are only four clearly identifiable bands (matching the branches in the dendrogram), each one ideally belonging to the driver peaks of each one of the amino acids.

The last step was the search by peak list in the "1D NMR search" section of the Human Metabolome Database[7] using the groups of peaks from the four bands found (peak groups are listed in Table S1). Table S2 presents the number of peaks for each band, the compound name of all matching hits and their hit rank, with the addition of the first nonmatching hit, and the Jaccard Index (JI) and Match Ratio (MR) values for the searches. The MR is a ratio, expressed as a fraction, of the number of matching peaks over the total number of peaks involved in the query: matching, nonmatching left in the database peak list and nonmatching left on the submitted peak list. The JI is essentially the calculation of the MR ratio. A perfect search will result in a JI of 1.0, while submission of a peak list with missing (incomplete) or extra (expanded) peaks will result in a reduced JI value. The total number of peaks on the database differs slightly for several hits on a given compound, as for example for lysine the hit list has both the D- and L-isomers, and there usually appears spectral data from more than one database (HMDB[7] and BMRB[24]) for a given metabolite.

All four searches have shown a marked difference on the JI values from the matching hits and the best nonmatching hit, with a greater number of matching peaks for the matching hits. In three out of the four queries the best nonmatching hit is a compound with far more peaks on its spectrum than for the best hit. The best matching case was for L-valine, with all peaks matching and a JI of 1.0. The worst matching case was one of the L-leucine spectra with a JI of 0.632. These four cases were strong matching situations, although the search can have in some cases good matching for other molecules rather than the one really producing the NMR signals in the mixture. This is related to the diversity of the compounds, the population density of compounds with similar peaks, and the subset of peaks used in the query as picked from POD-CAST (it can happen that not all the peaks of the molecule were picked due to overlap with other peaks, incomplete peak picking or worse resolution of multiplets than in the databases).

The overlapping of peaks has always been presented in STOCSY as an issue,[6,15,19,20] as the aspect of a STOCSY trace that combines correlations to, at least, two compounds makes identification and assignment more cumbersome.[25] One of the main problems of the overlap regions is that the correlation



DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

trace for a given driver peak can show a mixture of correlations to two or more compounds that share the driver peak, or correlations to peaks of only one of those compounds. Moreover, it does not need to be a perfect apex to apex overlap, but rather any part of a peak can cause conflict with peaks from other compounds. The outcome would depend on the relative weights of the different molecules, directly related to their concentrations, at the chemical shift of the driver peak across the sample set. In some cases of overlap, if the driver peak belongs only to molecule A, molecule B can be masking a correlation to another peak of molecule A when the difference in concentration is important. This last problem cannot be overcome by POD-CAST applied on STOCSY data, but in a regular NMR-based metabolomics project it could be detected by inspection of the set of 2D NMR spectra for coupled resonances, namely COSY and TOCSY, or other experiments.

However, for the cases where the intensities of the overlapping peaks are similar, POD-CAST helps in the detection of these overlapping regions, both at first glance of the traces and analyzing the COI (*vide infra*), which allowed us to overcome the overlap problem, detect the overlapping peaks and improve the search results. A detailed visual inspection on the boundaries between bands II and III, and also between bands III and IV, shows overlap traces that clearly merge (weaker) correlations from both bands. Having detected that, the searches were performed once again for bands II and IV incorporating to each peak list the chemical shifts of the overlapping traces from band III, as shown in Table S1. In both cases, there was an improvement on the JI found for the hits: for band II it increased to 0.875, and for the matches of band IV it grew to 0.920 and 0.696 (first and second matching hits, respectively). Note that each of the sets of peaks incorporated to other bands already show a singularity in the dendrogram, being separated from the branch containing most of the leaves.

From the 7 peaks added to the original band IV list, 6 peaks belong to a singular overlapping region for lysine and leucine peaks, between 1.6 and 1.8 ppm, as shown in Figure S3 The quintuplet of lysine centered at 1.710 ppm (spectrum A) overlaps with the multiplet of leucine centered at about 1.690 ppm (spectrum B). The average spectrum for the whole spectra set (spectrum C) shows clearly that some peaks from both spectra overlap. In fact, instead of 18 peaks (13 from leucine plus 5 from lysine), the derived average spectrum has 14 peaks detected, meaning that 4 out of the 5 peaks from lysine are overlapping with peaks from leucine. The STOCSY trace with driver peak at 1.740 ppm (data not shown) clearly shows 5 peaks in the region with high correlation values, the quintuplet. This means that band III has more peaks than expected between 1.6 and 1.8 ppm, as a result of the quintuplet center at 1.710 ppm passing undetected in the peak detection. The peaks at 1.707 and 1.713 ppm show the contribution from the former due to their proximity. Removal of one of those two peaks around 1.710 ppm and the peak at 3.727 ppm (derived from a similar analysis) from the peak list in the search for band III further increases the JI to 0.816, not because of more matching peaks, but due to a reduction in the number of nonmatching peaks.

Similar conclusions can be obtained employing the information in the COI bar plot with a fairly simple interpretation. The COI value for each trace reflects the number of peaks that have a normalized overlap coefficient above a given threshold. The default value of 0.5 is a conservative (empirical) value that allows to capture intermediate and high correlations (i.e., overlapping peaks and peaks from the same molecule). This leads to three main scenarios for the analysis of the COI values: (a) the case of a compound without overlapping peaks will show COI values throughout the band traces that nicely match the number of peaks in the dendrogram branch, as happens for example in band I on both the simulated (*vide supra*) and the artificial (*vide infra*) spectra sets, (b) if the COI values throughout the band are even, but greater than the total number of leaves in the dendrogram branch, this will point to a band whose peak list should be expanded with overlapping peaks from other bands (for instance, band IV from the simulated spectra set in Figure 3 has an evenly distributed value of 23 for the COI on its 16 peaks, and its peak list was expanded with the 7 lowest peaks from band III for an improved query output), and (c) a band with a frequent COI value and higher COI values for some of its traces (it can be a smaller branch on its own if the overlap is between multiplets) will reveal a band containing all of its peaks, including some that would overlap with other bands (as will appear below for the artificial mixture set).

There is an added adjustment that can be made at this stage with POD-CAST: tuning the threshold value for the creation of matrix **B**, which then leads to the COI. If this threshold is increased for the simulated spectra set to 0.875, as shown in Figure S4, the contribution from overlapping peaks on the COI is lost and only the high correlation elements remain. Clearly, each band block has an almost even COI value, reduced from that in Figure S2 by the number of overlapping peaks, 5 in this case for lysine and leucine. Even more, these five peaks from the multiplet form a block of their own, with COI of 5. This adjustment of the threshold is then a valuable tool for assessing the traces corresponding to overlapping peaks in a quick manner.

Sorting matrix $\mathbf{O}_v$ additionally in the horizontal dimension provides a diagonal matrix similar to that of the global hierarchical clustering in CLASSY.[15] Again, the focus in POD-CAST is to be able to identify compounds, while dealing with overlap regions. Figure S5 (center) shows matrix $\mathbf{O}_{vh}$ for the simulated spectra set. The sections of the bands with an even COI in Figure 3 will render high overlap square blocks through the diagonal in matrix $\mathbf{O}_{vh}$. The overlap regions are again easy to identify, appearing as rectangles with weaker overlap. In the simulated spectra set, only the regions connecting the lowest three squares show these rectangles, equivalent to traces between bands in Figure 3. These plots where each little square represents a driver peak are ideal to visualize the concept behind the COI.

Despite the simplicity of this simulated spectra set, there is an enormous benefit on the application of POD-CAST to any given metabolomics set. The choice of the simulated spectra set as a proof of concept had the objective of minimizing possible drawbacks appearing in real sets, of which the most important could be peak (mis)alignment, bigger variations in concentration (dynamic range), peak overlaps, and peak picking. As it was already discussed elsewhere,[26] peak alignment is crucial for obtaining good quality correlation values, and hence STOCSY traces. The same applies when the whole 2D STOCSY correlation matrix is sought. The STOCSY traces might also be affected if a compound is only present in small concentrations and in a few samples from the whole data set (instead of being present in a big percentage of the samples), and even more if one or some of those peaks suffer from overlap with any other peak or more than one. This is a worst


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

case scenario, but anything in between this and the simulated set has different degrees of weaknesses when performing a STOCSY analysis on it.

Finally, the choice of driver peaks is crucial for the present analysis. As mentioned above, the introduction of all (or most) of the peaks on the spectra set allows a good estimation of the number of compounds detected in the samples. However, some small peaks on the average spectrum picked as driver peaks could be "noise", if coming from molecules at low concentrations, while others can be $^{13}C$ satellite peaks coming from molecules at high concentrations. Even though the satellite peaks would appear in the reference spectra in the databases, they will not appear on the tables of assigned peaks, and hence are not "searchable". These satellite peaks are indeed expected to show the same correlations as the principal peaks they are related to, and show together in the same bands in matrix $\mathbf{O}_v$, if included, but their inclusion in the database queries will likely reduce the overall JI by adding nonmatching peaks. The peaks from "noise" might be showing correlations to any set of peaks, or poor correlations in general, and add confusion to the database search process. This is the reason why it is imperative to analyze the peaks and overlap traces next to the average spectrum used for peak picking, to be able to keep these "polluting peaks" out of the list of driver peaks to be assigned.

The artificial mixture set, whose average spectrum is shown in Figure S6, presents at least 200 peaks if peak picking is performed just above a given threshold over the noise level. To focus on the application of POD-CAST and its advantages, the following analysis will be first reduced to the major components of the artificial mixture, using as driver peaks those with intensities above that of the downfield peak at 8.55 ppm. Matrix $\mathbf{O}$ for this reduced set is shown in Figure S7, and Figure 4 shows matrix $\mathbf{O}_v$ for this reduced set (an expanded version showing the labels is included in the Supporting Information in Figure S8).

At first glance, it appears that only five bands are present in the sorted matrix, both analyzing the dendrogram and the matrix itself. However, an inspection of matrix $\mathbf{O}_{vh}$ for the artificial mixture set, shown in Figure S9 reveals a disconnection between the second upper block and the lower blocks. This in fact can be identified as a small band, also detected in the dendrogram and with singular COI values compared to the neighbor blocks. In consequence, not only it will be analyzed as a band itself, from a total of six, but also peaks from overlap with its neighbor bands will be introduced into its peak list. Matrix $\mathbf{O}_{vh}$ also shows pairwise overlap between the lowest three bands: bands IV and V, bands IV and VI, and bands V and VI show rectangles of intermediate overlap values. These support the selection of additional peaks, which are missing to them due to overlap, from band IV to be added on bands V and IV.

The database search was then pursued using the peak lists of the six bands, as listed in Table S3, incorporating now directly overlapping driver peaks. This addition was also supported by the observation of both the dendrogram branches and matrix $\mathbf{O}_v$, and by analyzing the COI values as described above. Table S4 shows the results for the database query showing the hits and the first nonmatching compound, together with the corresponding JI and MR values, and includes the peaks from six compounds: choline chloride, pantothenic acid, taurine, arginine, citrulline, and ornithine. Matching hits for this set range from a perfect match, JI of 1.0, for choline and taurine, to

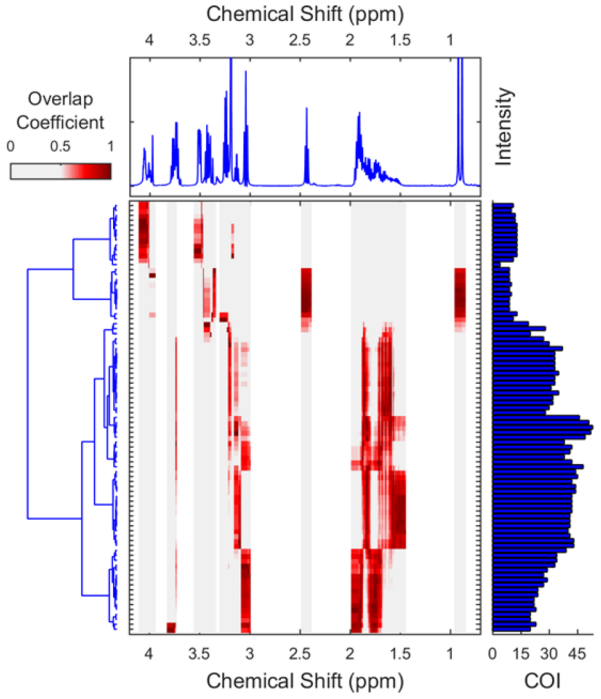



**Figure 4.** POD-CAST on artificial mixture set: (Top) Average spectrum. (Left) Dendrogram obtained by HCA over the traces from the overlap matrix of the 2D STOCSY traces using manual peak picking, peak list reduced to the most abundant compounds. (Center) Overlap matrix $\mathbf{O}$ (Figure S7) sorted by the order of the driver peaks obtained after HCA and dendrogram generation. (Right) Cumulative overlap index indicating the number of peaks with overlap coefficient values above the chosen threshold.

a lower JI value of 0.583 for citrulline (although the third matching hit for arginine is slightly lower).

Of the six molecules in the reduced set of high intensity peaks, three amino acids have a rather small structural difference, as shown in their chemical structures in Chart S1. Arginine, citrulline, and ornithine structures differ on the groups bound to the nitrogen atom in the end of the aliphatic chain of three methylenes, with arginine being a guanidine derivative, citrulline a urea derivative, and ornithine having the free amino group. Their $^1H$ 1D NMR spectra have two triplets downfield from 3.0 ppm for the $\alpha$ and $\delta$ protons, and two complex multiplets between 1.5 and 2.0 ppm for the $\beta$ and $\gamma$ methylenes. The overlap of the six multiplets creates the continuous signal overlap region between 1.5 and 2.0 ppm, seen in the top of Figure 4. The peaks from the three amino acids are clearly identified as well as the overlapping peaks for citrulline and ornithine within the arginine band.

The peak list for the artificial mixture set was expanded to include low intensity peaks, mostly aromatics and those between 2 and 3 ppm. A POD-CAST analysis with this list of peaks provided the results seen in Figure S10, where the bands are annotated for the corresponding compounds. The main observations drawn from the figure were: (1) no histidine peaks are present; (2) folic acid peaks do not present overlap in the upfield region of the spectra; (3) only the aromatic peaks from tryptophan are detected, the aliphatic ones are overlapped with high intensity peaks and there is no correlation observed with them; (4) only the doublet at 4.535 ppm is seen for ascorbic acid, the remainder of the correlations are obscured by more intense peaks and the two peaks appear within the


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

arginine band; (5) glutamic acid peaks present a strong overlap (most of its peaks) with folic acid (an amide formed from glutamic acid amino group and pteroyc acid), and its peaks were split in two bands; (6) taurine peaks were separated compared to Figure 4, as the bands that have overlap (arginine and pantothenic acid) have also other bands in between (glutamic acid and tryptophan).

POD-CAST was also successfully tested in sample sets from biological origin, for example the 1D $^{1}$H NMR spectra from the published study on falcons affected by aspergillosis.[27] Figure S11 shows the sorted overlap matrix from the spectra from plasma samples obtained from falcons, with the putative annotation for each band. The main band consists of peaks from the wide lipid resonances. Although it is better seen in a magnification (not shown), traces on the top section of the band correspond to resonances associated with lipids linked to HDL, while those in the lower section of the band to the VLDL type. The second major band is that for glucose (both forms together), without the anomeric doublet peaks, that due to the overlap are contained within the lipids band. The third band in width, in the bottom of the figure, corresponds to peaks from 3-hydroxy-butyrate, although some peaks are missing from the multiplet at 4.14 ppm due to overlap with lactate. These overlapping lactate peaks are in a neighbor band, but the rest of the lactate six peaks are split into other two bands, one peak almost adjacent to the lipids band (the one upfield on the 1.33 ppm doublet) and a pair of nonoverlapping peaks in between bands of leucine and pyruvate. Other amino acids and small size metabolites completes the list: glycine, citrate (one clean doublet only, the other overlaps with lipids), alanine (the doublet only, overlapping with lipids), valine, urea, creatinine, phenylalanine/tyrosine/formate (overlap also due to their small intensity), betaine, and creatine.

It is worth reinforcing that POD-CAST does not seek to replace STOCSY, rather in this case it is being applied to STOCSY data to boost its performance. For example, the implementation of STOCSY in the last case would imply the analysis of over 115 individual STOCSY traces, to be grouped based on their similarity and peaks with high correlation put into peak lists for database queries. Before even beginning to look at these traces, a POD-CAST analysis reduces the number to about 20 sets of 1D STOCSY pseudospectra that are expected to contain similar correlation profiles, as revealed in Figure S11. This would roughly imply a 5-fold reduction on the time of analysis, if not more. In addition, the visual identification of traces from driver peaks with overlapping resonances, and their decluttering into the appropriate subsets, is cumbersome to perform manually. Using POD-CAST the identification of traces with mixed contributions is more direct, by looking at the COI profile, the overlap traces and the doubly sorted overlap matrix looking for cross-peaks off the diagonal square blocks.

After obtaining the best hits possible in the database query and comparing the spectra of the hits with representative STOCSY traces (chemical shifts, integration, multiplicity and couplings), the compound identification should proceed as it is standard in any NMR metabolomics study. The assignment should go from Level 2 of confidence, "putatively annotated compound", to Level 1, "identified compound", by using two orthogonal analytical techniques to the analysis of the metabolite of interest and to a reference standard.[28] For this purpose, the chemical shifts from the bands are used in conjunction with the data obtained from the collected 2D correlation experiments, namely TOCSY, COSY, HSQC, and HMBC in this work for the artificial mixture set.

POD-CAST shares some similarities with approaches like CLASSY and DemixC. In CLASSY, the main focus is put on the biological information recovery instead of the identification of compounds. This identification step is a crucial step in CLASSY, as it needs to create local clusters that would then be "related by global hierarchical clustering". While the authors easily explain how to obtain local clusters (having no overlapping peaks), they mention the problem for identifying overlapping peaks or peaks in crowded regions. For example, arginine and ornithine are located in one big cluster in their study (even though it is worth recognizing the intrinsic biological correlation between them), while both metabolites are easily identified with POD-CAST (even with citrulline contributing to the crowded region between 1.5 and 2.0 ppm). Still, while the claim is that the interpretation of the biological perturbations is not hindered by splitting sets of structurally related peaks, it is easy to realize that for identification purposes any database query with a limited peak list will produce inferior results.

DemixC is proposed for the analysis of mixtures and identification of the spin systems analyzing the TOCSY spectrum of the mixture, with the ulterior identification of the mixture components through database search. DemixC will create an overlap matrix from the inner products of the correlation traces, reduce its diagonal peak to the intensity of the second highest peak, and normalize the adjusted overlap matrix. Then an importance index will be calculated for each trace, as a cumulative measurement of the overlap, and a clustering algorithm applied over the vector of importance indexes. This is "equivalent" to counting peaks of overlap for each trace and ordering the traces by that number, instead of strictly comparing the chemical shifts of those overlaps, as in POD-CAST. This cumulative overlap contains low, intermediate, and high overlap values, so that traces from compounds with no overlap will have an importance index similar to the number of peaks it has, but if it had overlapping peaks the importance index will be higher due to the added intermediate overlap values.

For example, if a mixture of two compounds had a molecule with only three doublets, not overlapping with any other peak in the spectrum, and a second molecule had another set of three doublets, also nonoverlapping, a crude analysis with DemixC will count 6 peaks for each trace in the importance index of each one of the 12 traces, and the traces would form a mixed cluster of 12 peaks with an importance index value of 6, instead of 12. A database query for these 12 peaks will most likely produce faulty hits. Instead, POD-CAST will readily identify that there are two sets, composed of six peaks each with even COI values of 6 for all traces, by clustering the different overlap traces. As a result, the two subsets would be expected to produce successfully in database queries. In the reported DemixC examples, there is no analysis on the significance of the importance index value, and no discrimination between spectra subsets with equal number of peaks or importance index.

For cases where there is overlap, DemixC has a strategy for choosing a representative trace for the database query. DemixC will select for each cluster the trace with the smallest importance index value, so that "the likelihood is maximized that the selected traces reflect individual components free of spurious contributions from other spin systems".[20,29] This solution clearly leads to incomplete peak lists, as overlapping


DOI: 10.1021/acs.jproteome.7b00617
J. Proteome Res. 2018, 17, 392−401

peaks present in other clusters are being neglected in the diminished peak lists. This, like the split sets proposed in CLASSY for crowded regions, will lead to poorer matching scores, while POD-CAST is expected to perform better by including also the overlapping peaks in the peak list.

## CONCLUSION

A simple and fast methodology was developed for accelerating and improving database queries for the identification of compounds in metabolomics studies. POD-CAST was able to quickly and efficiently cluster all the peaks selected as driver peaks for STOCSY traces from $^{1}$H NMR spectra in two model spectra sets, one simulated (lacking misalignment issues), and the other from experimental spectra from artificial mixtures, and also in a sample set from biological origin. POD-CAST was also able to detect peak overlap regions and identify peaks missing in peak lists to improve the search scores. This issue of overlap in correlation, in these cases in STOCSY, has always been described as problematic, hindering identification, and was normally avoided, ignored, or neglected (yielding incomplete peak lists for the database query and inferior results).

POD-CAST bases its peak overlap detection on the weaker correlation observed for compounds with overlapping peaks when these peaks are used as driver peaks for STOCSY, compared to using nonoverlapping peaks from the same compounds, or peaks from compounds without peak overlap, as driver peaks. In addition, POD-CAST has an adjustable parameter, the threshold used to generate (from a binary matrix) the COI, which measures the cumulative overlap and can account for different strengths of overlap to bypass cases where the COI is more or less even throughout a band.

Other approaches use indexes that are equivalent to counting peaks, like the importance index in DemixC or the sum of the overlap trace elements in CLASSY, which derives from the inner product of a binary connectivity matrix and is used to create local clusters for identification. On the basis of only these **P** vectors, slightly different in each approach, they will likely fail to differentiate traces from spectra with the same number of peaks but at different chemical shifts. In similar cases, POD-CAST will successfully cluster overlap traces from all driver peaks based on the distances among the traces, and similar traces will be clustered together in bands, regardless of the number of peaks within it (compared to numbers from other bands).

Visualization is key for quick interpretation of the data. POD-CAST presents a plot that includes all the necessary information to assess the presence of bands of similar traces and simultaneously account for overlapping peaks absent from some bands, based on their cross correlation and the COI values for each trace. Doubly sorted overlap matrices are also useful for the identification of overlapping peaks. Database queries are expected to provide high quality output, although this depends on the spectra chemical shift population distribution and the number of similar spectra within the database.

While DemixC is proposed for the analysis of TOCSY spectra, it is also suggested to be useful for metabolomics, which is completely understandable as the combination of TOCSY traces for the different spin systems within a molecule should resemble the STOCSY trace for that molecule. Likewise, POD-CAST could be used for analyzing traces from TOCSY spectra, and this will be exemplified in a separate work.

Current efforts are also concentrated in the application of POD-CAST in other type of traces, as well as its extension to $^{13}$C NMR spectra. The use of $^{13}$C NMR spectra has recently been highlighted as a promising alternative to avoid the overlap issue in STOCSY, due to the greater peak spreading between 20 to 200 ppm, and was supported by the development of a specialized probe with enhanced $^{13}$C sensitivity.[14]

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00617.

> Stacked selected STOCSY traces from simulated spectra set; POD-CAST on simulated spectra set; chemical shifts for all driver peaks in each band for simulated spectra set; database search in HMDB for simulated spectra set; overlap region 1.6−1.8 ppm from simulated spectra set; POD-CAST on simulated spectra set, higher threshold; POD-CAST on simulated spectra set, block diagonal matrix; average spectrum for artificial mixture set; overlap matrix **O** for artificial mixture set; POD-CAST on artificial mixture set; POD-CAST on artificial mixture set, block diagonal matrix; chemical shifts for all driver peaks in each band for artificial mixture set; database search in HMDB for artificial mixture set; chemical structures for amino acids arginine, citrulline, and ornithine; POD-CAST on artificial mixture set, expanded peak list; POD-CAST on falcon plasma set (PDF)

## AUTHOR INFORMATION


### Corresponding Author

*E-mail: pablo.hoijemberg@cibion.conicet.gov.ar.

### ORCID

Pablo A. Hoijemberg: 0000-0002-4870-3309


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

P.A.H. thanks Guillermo Moyna and Andrés Zelcer for valuable comments. P.A.H. thanks Julien Wist for helpful discussions. P.A.H. thanks the postdoctoral fellowship from Fundación Argentina de Nanotecnología sponsored by Centro de Investigaciones en Bionanociencias "Elizabeth Jares Erijman" (CIBION-CONICET). The authors thank Prof. Jeremy K. Nicholson and Dr. Kirill A. Veselkov (Imperial College, London, UK) for kindly sharing their MATLAB scripts to generate STOCSY plots. This work is dedicated to the memory of Bence.

## REFERENCES


(1) Nicholson, J. K.; Lindon, J. C. Systems biology: Metabonomics. *Nature* **2008**, *455* (7216), 1054−1056.

(2) Beger, R. D. A review of applications of metabolomics in cancer. *Metabolites* **2013**, *3* (3), 552−574.

(3) Armitage, E. G.; Barbas, C. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *J. Pharm. Biomed. Anal.* **2014**, *87*, 1−11.

(4) Emwas, A. H. M.; Salek, R. M.; Griffin, J. L.; Merzaban, J. NMR-based metabolomics in human disease diagnosis: Applications,




DOI: 10.1021/acs.jproteome.7b00617
*J. Proteome Res.* 2018, 17, 392−401

limitations, and recommendations. *Metabolomics* **2013**, *9* (5), 1048−1072.
(5) Geiszler, P. C.; Auer, D. P.; Daykin, C. A. The journey from metabolic profiling to biomarkers: The potential of NMR spectroscopy based metabolomics in neurodegenerative disease research. *Curr. Metabolomics* **2013**, *1* (2), 160−179.
(6) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; et al. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal. Chem.* **2005**, *77* (5), 1282−1289.
(7) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2012**, *41* (D1), D801−D807.
(8) Ellinger, J. J.; Chylla, R. A.; Ulrich, E. L.; Markley, J. L. Databases and Software for NMR-Based Metabolomics TL - 1. *Curr. Metabolomics* **2012**, *1* (1), 28−40.
(9) Wei, S. W.; Zhang, J.; Liu, L. Y.; Ye, T.; Gowda, G. a N.; Tayyari, F.; Raftery, D. Ratio Analysis Nuclear Magnetic Resonance Spectroscopy for Selective Metabolite Identification in Complex Samples. *Anal. Chem.* **2011**, *83* (20), 7616−7623.
(10) Robinette, S. L.; Lindon, J. C.; Nicholson, J. K. Statistical spectroscopic tools for biomarker discovery and systems medicine. *Anal. Chem.* **2013**, *85* (11), 5297−5303.
(11) Sands, C. J.; Coen, M.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Data-driven approach for metabolite relationship recovery in biological \n 1H NMR data sets using iterative statistical total correlation spectroscopy. *Anal. Chem.* **2011**, *83* (6), 2075−2082.
(12) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M. D.; Nicholson, J. K. Subset optimization by reference matching (STORM): An optimized statistical approach for recovery of metabolic biomarker structural information from 1H NMR spectra of biofluids. *Anal. Chem.* **2012**, *84* (24), 10694−10701.
(13) Posma, J. M.; Garcia-Perez, I.; Heaton, J. C.; Burdisso, P.; Mathers, J. C.; Draper, J.; Lewis, M.; Lindon, J. C.; Frost, G.; Holmes, E.; et al. Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Anal. Chem.* **2017**, *89* (6), 3300−3309.
(14) Clendinen, C. S.; Lee-McMullen, B.; Williams, C. M.; Stupp, G. S.; Vandenborne, K.; Hahn, D. A.; Walter, G. A.; Edison, A. S. 13C NMR metabolomics: Applications at natural abundance. *Anal. Chem.* **2014**, *86* (18), 9242−9250.
(15) Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K. Cluster analysis statistical spectroscopy using nuclear magnetic resonance generated metabolic data sets from perturbed biological systems. *Anal. Chem.* **2009**, *81* (16), 6581−6589.
(16) Blaise, B. J.; Shintu, L.; Elena, B.; Emsley, L.; Dumas, M. E.; Toulhoat, P. Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. *Anal. Chem.* **2009**, *81* (15), 6242−6251.
(17) Blaise, B. J.; Navratil, V.; Emsley, L.; Toulhoat, P. Orthogonal filtered recoupled-STOCSY to extract metabolic networks associated with minor perturbations from NMR spectroscopy. *J. Proteome Res.* **2011**, *10* (9), 4342−4348.
(18) Blaise, B. J.; Navratil, V.; Domange, C.; Shintu, L.; Dumas, M. E.; Elena-Herrmann, B.; Emsley, L.; Toulhoat, P. Two-dimensional statistical recoupling for the identification of perturbed metabolic networks from NMR spectroscopy. *J. Proteome Res.* **2010**, *9* (9), 4513−4520.
(19) Zhang, F.; Brüschweiler, R. Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angew. Chem., Int. Ed.* **2007**, *46* (15), 2639−2642.
(20) Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Brüschweiler, R. Strategy for automated analysis of dynamic metabolic mixtures by NMR. Application to an insect venom. *Anal. Chem.* **2007**, *79* (20), 7748−7752.
(21) Kovacs, H.; Moskau, D.; Spraul, M. Cryogenically cooled probes - A leap in NMR technology. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46* (2−3), 131−155.
(22) Hwang, T. L.; Shaka, A. J. Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J. Magn. Reson., Ser. A* **1995**, *112*, 275−279.
(23) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: a review. *ACM Comput. Surv.* **1999**, *31* (3), 264−323.
(24) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; et al. BioMagResBank. *Nucleic Acids Res.* **2007**, *36* (Database), D402−D408.
(25) Couto Alves, A.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets. *Anal. Chem.* **2009**, *81* (6), 2075−2084.
(26) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.* **2009**, *81* (1), 56−66.
(27) Pappalardo, L.; Hoijemberg, P. A.; Pelczer, I.; Bailey, T. A. NMR-Metabolomics Study on Falcons Affected by Aspergillosis. *Curr. Metabolomics* **2014**, *2*, 155−161.
(28) Viant, M. R.; Kurland, I. J.; Jones, M. R.; Dunn, W. B. How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* **2017**, *36*, 64−69.
(29) Robinette, S. L.; Zhang, F.; Brüschweiler-Li, L.; Brüschweiler, R. Web server based complex mixture analysis by NMR. *Anal. Chem.* **2008**, *80* (10), 3606−3611.


DOI: 10.1021/acs.jproteome.7b00617
*J. Proteome Res.* 2018, 17, 392−401