Statistical correlations between NMR spectroscopy and direct infusion FT-ICR mass spectrometry aid annotation of unknowns in metabolomics

Jie Hao,† Manuel Liebeke,† Ulf Sommer,§ Mark R. Viant,§ Jacob G. Bundy,†† Timothy M.D. Ebbeles.†*†

†Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, UK.
§NERC Biomolecular Analysis Facility – Metabolomics Node (NBAF-B), School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.

KEYWORDS. Metabolomics, correlation, statistical heterospectroscopy, chemometrics, metabolite, mass spectrometry, NMR spectroscopy.

ABSTRACT: NMR spectroscopy and mass spectrometry are the two major analytical platforms for metabolomics, and both generate substantial data with hundreds to thousands of observed peaks for a single sample. Many of these are unknown, and peak assignment is generally complex and time-consuming. Statistical correlations between data types have proven useful in expediting this process, for example in prioritizing candidate assignments. However, this approach has not been formally assessed for the comparison of direct-infusion mass spectrometry (DIMS) and NMR data. Here, we present a systematic analysis of a sample set (tissue extracts), and the utility of a simple correlation threshold to aid metabolite identification. The correlations were surprisingly successful in linking structurally related signals, with 15 of 26 NMR-detectable metabolites having their highest correlation to a cognate MS ion. However, we found that the distribution of the correlations was highly dependent on the nature of the MS ion, such as the adduct type. This approach should help to alleviate this important bottleneck where both 1D NMR and DIMS datasets have been collected.

Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are widely used technologies in metabolomics, allowing a vast range of metabolites to be assayed in a broad variety of sample types. Direct infusion mass spectrometry (DIMS) produces information rich data sets in minimal time, without prior chromatographic separation. This makes it valuable for applications requiring high throughput such as epidemiology and environmental and toxicological screening. NMR and DIMS both produce hundreds to thousands of peaks characterising a single sample, although, for both techniques, the number of actual metabolites is lower (as individual metabolites may give rise to multiple peaks – because many metabolites have magnetically non-equivalent 1H nuclei, for NMR, or because many metabolites may have both adduct and isotopologue peaks, for DIMS). However, in common with other untargeted metabolomics techniques, the chemical identities of the vast majority of these are unknown a priori.

Identification of unknowns is one of the largest challenges in metabolomics. While individual unknown peaks can, of course, be assigned using classical analytical approaches, the process is time consuming, costly, and success is unpredictable. However, statistical approaches have been shown to help greatly in the identification process, for example, by highlighting signals which are highly correlated to each other and therefore might arise from the same molecule. This has been formalized for NMR studies as “statistical total correlation spectroscopy” (STOCSY) to the point where it (or closely related approaches) is now routinely used to help assign peaks. A similar approach has also been implemented for LC-MS data, e.g. the CAMERA package uses statistical correlation to link peaks which could be structurally related, and RAMSY explicitly considers between-peak associations. The next step is to statistically link signals across multiple platforms for identification, and this has indeed been done, and described as “statistical heterospectroscopy”. However, to date, there have been few attempts to link DIMS and NMR data in this way. Marshall et al. used multiblock partial least squares modelling of parallel 1D 1H NMR and DIMS datasets of the same samples to help identify compounds contributing to the separation of two treatment groups, which utilizes the same principle. Their study, however, focussed on developing chemometric methods for identifying metabolic biomarkers, and neither systematically evaluated the performance of the approach across all observable metabolites, nor explicitly examined pairwise correlation as a tool. Verberk et al. and Southam et al. used cross-platform comparison with NMR data to support the putative annotations in DIMS datasets, but again did not systematically evaluate the value of this approach. Recently, Bingol et al. have published two approaches to link NMR and DIMS data sets. Their approach does not use statistical correlation but uses one platform to validate annotations suggested by the other platform.

Here, we set out to examine the statistical relationship between NMR and DIMS data obtained in parallel on a set of >200 samples from an environmental metabolomics study. Our aim was to evaluate the usefulness of statistical correlations to aiding identification of unknowns in the NMR data. Since many of the molecules observed in the DIMS data will be below the NMR detection limit, we took an NMR-led approach and looked for DIMS ions which could correspond to a set of 26 metabolites identified in the NMR data. We examined the distributions of NMR-DIMS correlations which link signals of the same metabolite, versus those of different metabolites. We gauged the extent to which DIMS signals with high correlations to the NMR data are useful in reducing the number of possible molecular identities. As expected, we find that correlations between signals from the same metabolite have a distribution very different to the other, unrelated correlations, and that this is strongly influenced by the type of DIMS adduct considered. We demonstrate that using these correlations can greatly inform identification, with the most highly correlated ion being structurally related to the known NMR signal in the majority of cases.

METHODS
Samples and Analysis by NMR and DIMS. We used samples from an environmental physiology study of the earthworm *Lumbricus rubellus* consisting of four separate experiments where worms were exposed to environmental stresses. The stressors were temperature, soil moisture, pH and food restriction. In total 205 samples were collected for parallel NMR and DIMS analysis. We used a previously published protocol to prepare earthworm tissue powder and extract the metabolites. 1 Briefly, frozen powdered tissue (~250 mg) was extracted with 2 ml of ice-cold acetone:methanol:water mixture (2:2:1 vol./vol.), the extracts centrifuged (2 min, 16,000 g), and the supernatants aliquoted for DIMS and NMR analyses. For the NMR measurement, an extra solid-phase extraction was used to remove the abundant compound 2-hexyl-5-ethyl-furan-3-sulfonic acid (HEFS) from the samples. 1D 1H NMR data were acquired on a Bruker 600 MHz spectrometer using standard approaches described elsewhere. 1 Similarly, DIMS data were acquired with standard approaches, using an Advion Triversa Nanomate nanoelectrospray source and Thermo Fisher Scientific LTQ FT Ultra Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS) in positive ion mode only, as the negative ion spectra suffered from an over-abundance of HEFS ions. Data were acquired as described elsewhere in ten selected ion monitoring (SIM) windows from m/z 70 to 1000, using two technical replicates per sample. 1

Data processing. The NMR spectra were phased, baseline corrected and calibrated using INMR v3 (MestreLab Research, Santiago do Compostela, Spain). Selected metabolites were semi-quantified using a binning approach. Bin boundaries were positioned manually to take account of small shifts in peak positions across the sample set, and highly overlapped peaks were avoided. Relative concentrations were estimated by integrating the intensity in each bin. The data were normalized by probabilistic quotient (PQN). 2 DIMS spectra were processed as described elsewhere 3, including several filtering steps. SIM windows were stitched using an internal calibration list of 106 known m/z values and a signal to noise filter (S/N ≥ 3.5), followed by a technical replicate filter (2 out of 2) and a 75 % sample filter, resulting in a combined peak list and intensity matrix. This matrix was further processed using PQN normalization, and k nearest neighbor zero filling (k=5). The final data consisted of 125 NMR bins and 2514 DIMS peaks for the 205 samples.

Assignment of known metabolites in NMR and DIMS data. An initial assignment of the DIMS data was made using MI-Pack 4. For the NMR data, 26 metabolites, associated with 50 bins, could be confidently assigned based on previous studies of earthworm extracts 5. Given the high mass accuracy of the FT-ICR MS instrument (<1 ppm with internal calibration 5), we aimed to identify the corresponding ions in the DIMS data by mass alone, and therefore searched for matching m/z entries in using a mass window of 1 ppm. To do this, we first computed the neutral mass of each NMR identified metabolite and then used a list of possible adduct forms to match observed m/z ratios to the known metabolites. We described adduct forms as [M+B+C–H]+. Here, A specifies the nature of the principal ion, either M or M+1 (first 13C isotope). B is the number of adding HEFS molecules and C the number and identity of adding metal ions or protons. nH indicates the number of protons removed to ensure a single positive charge. We excluded M+2 and heavier isotopologues, dimers and multiple charge states, and allowed up to 3 adducting metal ions. We also allowed up to 3 adducting HEFS molecules, each of which could account for one further metal ion. We allowed addition of a single 19K ion, but only if the more abundant 18K adduct was also observed. Finally, we removed all matches where the number of metal ions was not consistent with the number of exchangeable sites on the molecule. These rules allowed for a wide range of adduct complexity. For example, 21 ions annotated to [M+Na]+ were observed in the whole dataset, compared to 1 ion annotated to [M+1+2HEFS+2Na+K–2H]+. The details of the matching adducts are given in Supporting Information Table S1 and summarized in Figures S1 and S2.

Correlation analysis. Using the set of 26 known metabolites, we calculated Pearson correlation coefficients between all 125 NMR bin intensities and all 2514 DIMS features (314,250 pairs). We then divided these into structural pairs (NMR and DIMS features assigned to the same metabolite, 427 pairs) and non-structural pairs (all others, 313,823 pairs). See Table S1 for a list of structural correlations for each metabolite. Note that we expect false negatives in the non-structural pairs, i.e., ions of metabolites giving rise to NMR signals (perhaps below the limit of detection) but not annotated by the above process. However the number of such errors is expected to be very low in comparison to the number of true non-structural pairs. When more than one adduct matched an observed ion within the 1 ppm search window, all matches were considered and contributed separately to the correlation distribution. We note that this approach is conservative, in that multiple and false matches will tend to dilute the correlation signal. The distributions of structural and non-structural correlations were examined to determine if structural pairs were more highly correlated than non-structural pairs. Receiver Operator Characteristic (ROC) analysis was used to determine whether a simple correlation threshold could distinguish structural correlations from non-structural ones. For the ROC analysis, positives correspond to correlations above the threshold from NMR and DIMS signals of the same metabolite (true positives), or different metabolites (false positives). Negatives are correlations below the threshold from different metabolites (true negatives) or the same metabolite (false negatives). For each ROC curve we computed the area under the curve (AUC) statistic as a measure of the discrimination power of the correlation threshold.

RESULTS AND DISCUSSION

We analysed a set of samples taken from an environmental metabolomics study, representing 205 spectra (both 1D NMR and DIMS) of tissue extracts of *Lumbricus rubellus* earthworms. We make no attempt here at biological interpretation of these spectra, but use them only to investigate the statistical correlation of the NMR and DIMS data. We assigned 26 metabolites, corresponding to 50 bins in the NMR spectra, while 2514 peaks were found in the DIMS data. We computed Pearson correlations between all the NMR intensities of all 26 identified metabolites and the full set of DIMS intensities. To examine the difference between correlations derived from the same molecule in each technique (structural correlations) versus those derived from different molecules (non-structural correlations) we sought to find DIMS peaks corresponding to the 26 NMR-identified metabolites.

![Figure 1](image-url)
NMR data, and a further 6 annotated by 64 with concentrations greater than 10 relatively high adduct complexity allowed T structurally related DIMS signals that correlation might be informative in linking unknown NMR signals with difference between the structural and non-structural distributions suggests that correlation might be informative in linking unknown NMR signals with structurally related DIMS signals.

To explore the likelihood of false mass matches, given the relatively high adduct complexity allowed, we carried out a "null test". We selected all high concentration metabolites in normal human urine, as they would be relatively unlikely to be observed in worm extracts. From an initial list of 64 with concentrations greater than 10 μM, we removed 14 observed in our NMR data, and a further 6 annotated by Mi-Pack, leaving a test set of 44 metabolites. Our validation consisted of two stages: firstly, were there any potential matching ions; and secondly, what was the correlation distribution of any matches? For the first stage, three quarters (33) had no match to any DIMS peak. Five of the remaining 11 metabolites we would expect to be present in earthworm extracts6, but were presumably below NMR detection limits in our current study. This left 6 metabolites as potential false matches. For the second stage, we found that the correlations of these 6 to NMR were low and strongly resembled the non-structural correlation distribution (median 0.021, 95% interval -0.238 to 0.485, Figure S3). We conclude that the number of false positive matches of empirical formulas to the DIMS peaks is relatively low.

Using the set of known metabolites, we set out to evaluate the ability of a simple correlation threshold to discriminate between structural and non-structural NMR-DIMS pairs. We also investigated how the discrimination varied with signal intensity, signal variability (which will be dominated by biological variation) and DIMS adduct complexity. The top left panel in Figure 2 presents the ROC curve for all adduct types, showing that overall there was an excellent discrimination between structural and non-structural correlations (AUC=0.934). A threshold of r=0.415 gave a good balance between true and false positive rates (FPR=0.026, TPR=0.759) across all adducts. The optimal positive predictive value of 79.0% was found at a threshold of r=0.17 (TPR = 0.871, FPR = 0.207).

No significant relationship between structural correlation levels and the NMR signal intensity or variability was found in our data. However, analyzing subsets of specific adducts shows a number of interesting features (Figure 2). The best discrimination was shown for adducts incorporating two metal ions (AUC=0.959), and the related situation where there were three metal ions plus HEFS (i.e. adding a formally neutral HEFS which includes an additional metal ion, AUC=0.973). Both these cases represent molecules with a single exchangeable acidic proton (i.e. one metal ion to replace the acidic proton and one to confer the positive charge). It is clear that not only are metal adducts more common than the ‘standard’ [M+H]⁺ ion in this dataset, but that it is important to consider multiple metal ions for DIMS data. This corresponds with other studies that have also identified the presence of metal ion adducts with n≥2, especially in acidic compounds.

Figure 2: Different adduct types perform differently in discriminating structural from non-structural correlations. ROC curves using a simple correlation threshold are shown for all data combined (top left) and also broken down by adduct complexity (+/- HEFS, number of adding metal ions) and isotopologues. ROC curves are only shown for subsets of size greater than 10.
As expected, isotopologues ([M+1]) were strongly correlated to the NMR signals with an AUC of 0.894. Which correlation patterns are observed when examining individual metabolites in more detail? We compare two examples here by taking each metabolite and ranking all DIMS peaks by their correlation to the NMR signal. When more than one NMR signal was present, the one with the highest mean correlation was used. We then examine whether the DIMS signals putatively annotated to the metabolite are highly ranked. The earthworm metabolite lombricine has a remarkable 33 matched adducts in the DIMS data (Figure 3). This is because of the relative complexity and number of exchange sites on this metabolite: it is based on serine with an additional guanidinoethyl phosphate group. The adducts include combinations of up to 2 HEFS and 4 metal ions, and there is evidence that these all represent true positive matches: when considered in the context of the correlation rank of all observed DIMS peaks, all 33 fall within the top 51 ranked ions (out of 2514). Out of the 33 adducts, 22 contained at least one formally neutral HEFS molecule. Interestingly, several HEFS adducts are more strongly correlated to the NMHEFS molecule. Which metabolite are highly ranked.

The distribution of overall correlations is interesting, but many metabolomics practitioners will want to know how useful this would be in helping to characterize unknowns in a real-world situation. We therefore present three examples (with decreasing confidence of identification from the NMR data alone) where the correlations genuinely proved of practical value. Firstly we consider Ne,Nε,Nε-trimethyllysine (lysine-betaine HMDB01325), recently assigned in NMR spectra of earthworm extracts. (Figure 3) The only detectable NMR peak in typical complex mixtures is the singlet from the trimethylammonium group at δ = 3.12 ppm, which contains little structural information and lies very close in chemical shift to other singlets in the 1D (and 2D HSQC) NMR spectra. It is thus difficult to assign this metabolite in NMR metabolic studies without supporting evidence. Confidence in the assignment for these 1D NMR spectra is enhanced by the fact that the three highest-ranked correlations are to peaks for the [M+H]+ ion, the [M+1+H]+ ion and the [M+Na]+ ion for trimethyllysine.

Secondly, we used the approach to assign a metabolite not previously described in earthworms. A singlet was observed in the NMR spectra at δ 1.485 ppm, which was most highly correlated to a peak at m/z=496.36473. This could not be matched to any putative metabolite by mass alone, and so we also investigated lower-ranked correlations (Table S2). This generated some hits to metabolites that could trivially be ruled out on the basis of their expected NMR spectra (emodastine; and the three structural isomers 6-acetamido-3-aminohexanoate, glycyl-leucine, and N-acetyl-lysine), and (ranked 14th) 2-aminoisobutyrate (HMDB01906, or 2-methylalanine; calculated m/z=104.07065 for the [M+H]+ ion, observed m/z = 104.07060). This last compound would indeed be expected to produce a singlet at approximately the observed chemical shift, and we confirmed the assignment both by spiking an authentic standard and by acquiring a heteronuclear 2D spectrum (data not shown).

For the third example, we assigned a putative empirical formula to what is probably a novel metabolite. The NMR spectra show an ABX spin system at 5.18, 2.81, and 2.60 ppm, probably corresponding to the fragment R₂R₂CHCH₂R₃, where the methylene is near a chiral centre. (Note that this spin system was represented by 5 bins, as bin boundaries were chosen to minimize overlap from other resonances.) Correlating all of these resonances to the DIMS data yielded a number of high-ranking peaks, including the highest-ranked correlation for all peaks, which could all be assigned to the neutral mass 263.06412 (Table S3, supporting information); this corresponds to the probable empirical formula C₁₉H₂₃N₂O₆ (mass error between 0.02 ppm for [M+H]+ to 0.14 ppm for [M+Na]+ or [M+K]+). The only database match that we could find for this formula is the metabolite ascorbic acid. This is unlikely to be correct because a) we would expect to see other peaks in the NMR spectra, which are not observed, and b) it is a plant metabolite (while ascorbic acid, or a metabolite thereof, might be dietary in origin, the simpler explanation is that it is a different compound). No authentic standard is commercially available for confirmation.

We cannot say for certain if this is the correct empirical formula for this metabolite, but we think it is likely: our confidence is increased both by the observation of multiple adduct matches in the DIMS data corresponding to the same neutral mass, and by these correlations being observed for all NMR peaks. Hence, although in this case we have not completed a full metabolite assignment to a level 1 identification (according to the MSI reporting standards), it is clear how this approach could be of great value in helping future assignment efforts from complex mixtures, without necessarily needing purification of individual compounds.

However, it is not sufficient to rely on examples taken for a few individual metabolites. How does this approach fare if we now compare systematically the best-case correlation for all of the metabolites we could identify by NMR? In particular, we were keen to see in how many cases the ion with the highest-ranked correlation matched to the correct metabolite, as this would be of the most practical use for true unknowns. For 15 out of the 26 NMR-identified metabolites, the highest-ranked correlation was to an ion matching the correct adduct (Figure 4). Two other metabolites had the correct adduct within the top 10 correlations. Some of the exceptions, i.e. adducts with relatively poor correlations, included known isomers (e.g. succinate is a smaller and simpler molecule, and may be more representative of commonly observed metabolites than lombricine. The top 7 most highly ranked correlations all match to succinate adducts, while a more complex adduct appears lower in the ranking ([M+HEFS]+2K-H]+, rank 52). An examination of the individual sample intensities for this adduct reveals that there are missing values in several samples, which could account for the low correlation between the NMR and DIMS signal intensities. Several alanine adducts also correlate well with the succinate NMR signal, and rank closely to, but below, the top-ranked succinate adducts. The latter represents an example of a biological rather than structural correlation, easily understood as alanine and succinate are both increased as a result of anaerobic metabolism in invertebrates (similar lactate for vertebrates).
leucine and isoleucine), or compounds whose signals overlap with other metabolites in the NMR spectra (e.g. threonine) and thus would be expected to show lower correlations. We consider this to be surprisingly good performance – especially given that we are comparing two “one-dimensional” techniques, with substantial overlap occurring in the NMR spectra, and the anticipated ion suppression and other matrix effects in the DIMS data.

Our data here had a larger number of adducts than is likely to be common in most studies, because of the presence of the surface-active and high-concentration earthworm metabolite HEFS, leading to more complicated mass spectra than usual. Therefore, we believe this represents a more stringent test than is likely to be the norm.

CONCLUSIONS

Although the approach of correlating metabolomic data across platforms is widely discussed, it has remained surprisingly under-utilised in practice. DIMS data are increasingly used in metabolomics, especially where sample throughput is paramount, and it is particularly important to exploit multiple methods to aid assignment in this context due to lack of additional information about each peak, for example retention time. Further, ion suppression effects might be expected to hamper inter-platform correlations using DIMS data, although such effects are known to be minimized for nanoelectrospray ionization. Hence, it is somewhat surprising that our analysis has shown very little overlap between the distributions of structural and non-structural correlations. It is well accepted that one of the major benefits of NMR spectroscopy for metabolite profiling is its high instrument precision, which enables detection of subtle metabolic changes between samples. Structural correlations. It is well accepted that one of the major benefits of NMR spectroscopy for metabolite profiling is its high instrument precision, which enables detection of subtle metabolic changes between samples. It is therefore encouraging to note the good agreement between the two datasets in our study, suggesting that the nanoelectrospray DIMS approach is performing with comparable effectiveness, at least for the relatively high concentration metabolites. We have demonstrated that the approach can be an effective and real-world tool for aiding compound identification. In particular, in around 60% of the cases we studied, the highest ranking correlation to an NMR signal was from its cognate mass spectral signal. The technique will clearly not work well in all cases, for example where signals are overlapped, or where isomers exist. However, the approach is conceptually and practically simple and we expect it to find routine use in metabolomics in the future.

ASSOCIATED CONTENT

Supporting Information
Figures S1 & S2: distributions of matched adduct types. Figure S3: Distribution of correlations for the null test. Table S1: details of all NMR & DIMS assignments and correlations Table S2: top ranked correlations for metabolite 2-methylalanine. Table S3: correlations between NMR and DIMS for putative novel metabolite. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
*t.ebbels@imperial.ac.uk, *j.bundy@imperial.ac.uk.

Present Addresses
‡Max Planck Institute for Marine Microbiology, Department of Symbiosis, Celsius Strasse 1, Bremen, Germany

Author Contributions
The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

This work was funded by the Natural Environment Research Council (NERC), UK (grant number NE/H009973/1). TE and JH acknowledge additional support from the EU COSMOS project (grant agreement 312941). The FT-ICR MS instrument at the NERC Biomolecular Analysis Facility at the University of Birmingham (R8-H10-61) was obtained through the Birmingham Science City Translational Medicine: Experimental Medicine Network of Excellence project, with support from Advantage West Midlands. We are grateful to three anonymous referees whose comments improved the paper.

ASSOCIATED CONTENT

Supporting Information
Figures S1 & S2: distributions of matched adduct types. Figure S3: Distribution of correlations for the null test. Table S1: details of all NMR & DIMS assignments and correlations Table S2: top ranked correlations for metabolite 2-methylalanine. Table S3: correlations between NMR and DIMS for putative novel metabolite. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES


