Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology

Steven L. Robinette*, Ramadan Ajredini#, Hasan Rasheed#, Abdulrahman Zeinomar#, Frank C. Schroeder‡, Aaron T. Dossey§ and Arthur S. Edison†*

# Department of Biochemistry and Molecular Biology and National High Magnetic Field Laboratory†, University of Florida, Gainesville, Florida 32611. ‡ Boyce Thompson Institute, Cornell University, Ithaca, New York 14853, USA.

§ Currently at the United States Department of Agriculture, Agricultural Research Service

*To whom correspondence should be addressed: aedison@ufl.edu

! Currently at the Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London, UK SW7 2AZ.
Supplemental Methods

The experiments using human urine were done with samples collected at Imperial College following all appropriate ethical guidelines and standards. Human urine samples (n=10) were collected, filtered with a 20 µm nylon mesh filter, and stored at -30 °C. Urine samples were prepared by centrifugation at 1500 rpm for 10 minutes and 400 µL of urine was combined with 100 µL of NMR phosphate buffer (pH 7.4) consisting of 1 mM TSP, 3 mM NaN₃, in D₂O. TOCSY spectra were acquired on a Bruker Avance II with cryoprobe operating at 800.32 MHz using a DIPSI pulse sequence with excitation sculpting presaturation. TOCSY spectra were collected at a resolution of 2048 x 256 complex data points and zero-filled to 4096 x 1024. Two-dimensional Fourier transform, phasing, and baseline correction were executed in Topspin and the real data imported into Matlab for HATS analysis.
Table 1S:

| Ppa | Thr | Ala | Pro | Malate | Glucose | Glu | β-Ala | 2OG | Asp | Asn | NC-Glu | AIB |
|-----|-----|-----|-----|--------|---------|-----|-------|-----|-----|-----|--------|-----|
| Pre | Sucrose | Trehalose | OPE | Bet | Unk1 | Unk2 |
| Both | Lactate | | | | | |

Abbreviations: Ppa, Pristionchus Pacificus; Pre, Panagrellus Redivivus, Thr, Threonine; Ala, Alanine; Pro, Proline; Glu, Glutamate; β-Ala, β-Alanine; 2OG, 2-oxoglutarate; Asp, Aspartate; Asn, Asparagine; NC-Glu, N-Carbamylglutamate; AIB, DL-amino-isobutyrate; OPE, O-phosphoethanolamine; Bet, Betaine; Unk1, Unknown 1; Unk2, Unknown 2;
A) Overlay of the aromatic region from two unaligned human urine TOCSY spectra, (B) segments identified by HATS alignment algorithm, and (C) overlay of same two urine spectra after alignment. While the chemical shift corrections are small, the segmentation procedure allows alignments to be made in dense spectral regions with near crosspeak overlap. Regions of crosspeak chemical shift variation corrected by alignment are indicated with black arrows.
PLS-DA pattern recognition of *P. redivivus* (red: Pre) vs. *P. pacificus* (blue: Ppa) flow-through. $R^2 = 1$, $Q^2 = 0.80$ with 5-fold cross validation. Results implicate the same compounds separating the species as PCA, though with slightly different weightings.
Principal Component scores for 50% MeOH TOCSY spectra. This figure demonstrates the clear separation of *P. redivivus* (red: Pre) vs. *P. pacificus* (blue: Ppa) spectra along PC 1, allowing the PC1 loadings to be interpreted as the chemical differences between the 50% MeOH fractions of the species as was done in Figure 5.
PCA using 50% MeOH 1D $^1$H spectra. While 1D spectra are more commonly used for metabolomic analysis and can give rise to equally good class discrimination, for applications in which novel or unusual compounds are expected or in cases of significant spectral overlap, the increased dispersion and additional structural information provided by 2D correlation spectra such as TOCSY can aid in the interpretation and assignment of PCA loadings. Here, highly degenerate methylene resonances in the ascaroside side chains and other lipid derivatives result in congested 1D spectra in the 1-1.5 ppm region with difficult to interpret loadings. Additionally, the lack of coupling information from these side chain resonances to protons near functional groups complicates discrimination between ascaroside species and other polar lipid derivatives.
Figure 5S

Duplication of crosspeak lineshapes in PCA loadings due to mis-alignment. Arrows indicate a crosspeak present in both Ppa and Pre spectra with positional variability. This variability results in the apparent duplication of this peak in PCA loadings that is resolved by alignment.
Impact of alignment on PCA loading coefficients. The positional variability of the sucrose, glucose, and trehalose crosspeaks shown in Figure 2 results in the attenuation of loading coefficients, shown here by the colors of the crosspeak contours and the corresponding colormap. These lower loading coefficients indicate that there is less contribution of these crosspeaks to the unaligned PCA model.