ACCOUNTS
of chemical research

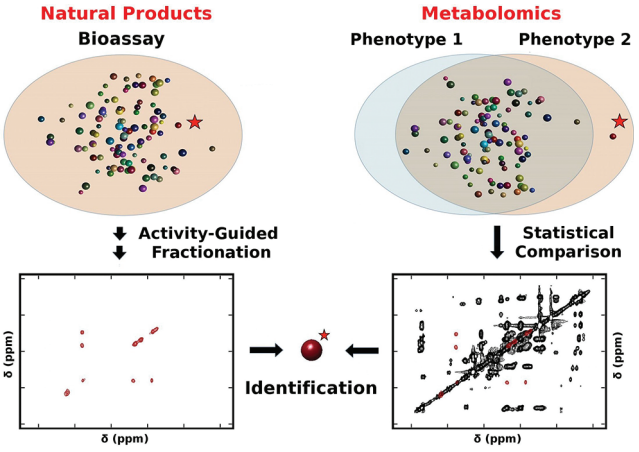
NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin

STEVEN L. ROBINETTE,[†] RAFAEL BRÜSCHWEILER,[‡]
FRANK C. SCHROEDER,[§] AND ARTHUR S. EDISON[*,⊥]

[†]*Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK,* [‡]*Department of Chemistry & Biochemistry and National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32306, United States,* [§]*Boyce Thompson Institute and Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States, and* [⊥]*Department of Biochemistry & Molecular Biology and National High Magnetic Field Laboratory, University of Florida, PO Box 100245, Gainesville, Florida 32610-0245, United States*

RECEIVED ON JUNE 14, 2011

CONSPECTUS



Small molecules are central to biology, mediating critical phenomena such as metabolism, signal transduction, mating attraction, and chemical defense. The traditional categories that define small molecules, such as metabolite, secondary metabolite, pheromone, hormone, and so forth, often overlap, and a single compound can appear under more than one functional heading. Therefore, we favor a unifying term, biogenic small molecules (BSMs), to describe any small molecule from a biological source.

In a similar vein, two major fields of chemical research,natural products chemistry and metabolomics, have as their goal the identification of BSMs, either as a purified active compound (natural products chemistry) or as a biomarker of a particular biological state (metabolomics). Natural products chemistry has a long tradition of sophisticated techniques that allow identification of complex BSMs, but it often fails when dealing with complex mixtures. Metabolomics thrives with mixtures and uses the power of statistical analysis to isolate the proverbial "needle from a haystack", but it is often limited in the identification of active BSMs. We argue that the two fields of natural products chemistry and metabolomics have largely overlapping objectives: the identification of structures and functions of BSMs, which in nature almost inevitably occur as complex mixtures.

Nuclear magnetic resonance (NMR) spectroscopy is a central analytical technique common to most areas of BSM research. In this Account, we highlight several different NMR approaches to mixture analysis that illustrate the commonalities between traditional natural products chemistry and metabolomics. The primary focus here is two-dimensional (2D) NMR; because of space limitations, we do not discuss several other important techniques, including hyphenated methods that combine NMR with mass spectrometry and chromatography.

We first describe the simplest approach of analyzing 2D NMR spectra of unfractionated mixtures to identify BSMs that are unstable to chemical isolation. We then show how the statistical method of covariance can be used to enhance the resolution of 2D NMR spectra and facilitate the semi-automated identification of individual components in a complex mixture. Comparative studies can be used with two or more samples, such as active vs inactive, diseased vs healthy, treated vs untreated, wild type vs mutant, and so on. We present two overall approaches to comparative studies: a simple but powerful method for comparing two 2D NMR spectra and a full statistical approach using multiple samples. The major bottleneck in all of these techniques is the rapid and reliable identification of unknown BSMs; the solution will require all the traditional approaches of both natural products chemistry and metabolomics as well as improved analytical methods, databases, and statistical tools.

Published on the Web 09/02/2011 www.pubs.acs.org/accounts
10.1021/ar2001606

1. The Common Theme: Biogenic Small Molecules

The identification and functional analysis of biogenic small molecules (BSMs) form the primary objectives of both natural products chemistry and metabolomics: natural products chemists are traditionally interested in the identification of new molecular structures with biological activity, whereas metabolomics and the related field of metabonomics have focused on correlating known BSMs with specific biological properties. Despite similar objectives, application of similar techniques, and frequent study of the same organisms, natural products and metabolomics research have remained largely separate fields, lacking regular interaction and exchange of ideas.

BSMs control intracellular processes (metabolism), intercellular processes (nervous system and hormonal regulation), intraspecific communication (e.g., via pheromones), and interspecific interactions (e.g., via defensive compounds); as a result, many drugs are derived from BSMs.[1] Many biogenic small molecules function at multiple levels, making it difficult and often artificial to differentiate between "metabolites", that is, compounds that are part of primary metabolism, and "secondary metabolites", that is, compounds not necessarily required for metabolic functioning. Similarly, functional categorizations such as "hormone", "pheromone", "biosynthetic intermediate", or "catabolite" can be problematic. For example, citric acid cycle intermediates are ligands of GPCRs that function in intercellular signaling.[2] Therefore, for the purpose of this review, we use "biogenic small molecules" (BSMs) as a unifying term to refer to all "metabolites", "secondary metabolites", and "natural products". Small molecules generally have molecular weights less than 1500 Da, but it should be noted that BSMs are further distinguished from larger biomolecules such as proteins and nucleic acids in that they are not strictly derived from a small number of known building blocks. As a result, the molecular structures of BSMs can be highly diverse and irregular, and their identification and characterization can present great analytical challenges.

This separation of natural products research and metabolomics appears ultimately rooted in differences of experimental design. The basic goal of both fields is to take complex mixtures of BSMs and identify a subset of compounds that describe a biological process or possess intrinsic biological activity. Whereas natural products research has relied on activity guided fractionation to isolate simple mixtures or individual compounds, metabolomics researchers have replaced chemical isolation with direct comparative analysis of complex mixtures to statistically identify subsets of BSMs relevant in a specific biological context.

We think that the fields of metabolomics and natural products are essentially "two sides of the same coin" and would both benefit from an increased exchange of expertise and approach. In natural products research, the process of isolating individual components can lead to chemical modification or degradation of the BSMs of interest. Furthermore, activity-guided fractionation necessarily risks losing important biological information that was encoded in the original BSM mixture, because most of the sample is not analyzed. The use of activity-guided fractionation therefore often fails in cases where several BSMs act in synergy. For example, a family of BSMs recently identified in the model organism *Caenorhabditis elegans* act synergistically as a mate-attracting pheromone, but most of this activity is lost when individual components of the pheromone are separated during fractionation.[3]

On the other hand, metabolomics uses statistical analysis of NMR and mass spectra of complex BSM mixtures to detect spectral features that correlate to a phenotype or biological property of interest and their response to stimuli. One of the most pressing challenges in metabolomics is the subsequent chemical identification of the BSMs represented by the detected spectral features. This problem of assigning peaks to BSMs is a major bottleneck in the typical metabolomics workflow. Structure elucidation techniques commonly applied in natural products research, especially integrated use of 2D NMR spectroscopy and mass spectrometry, may hold the key to increasing the number of identifiable compounds in metabolomic analyses.

2. Experimental Techniques That Bridge Natural Products and Metabolomics

In this Account, we focus on several approaches to mixture analysis by NMR. Each uses different sampling strategies, data sets, and computational algorithms, but they all illustrate methods that bridge natural products and metabolomics. We start with the most chemically intuitive approach of simply using 2D NMR spectra to analyze unfractionated mixtures for unstable BSMs. Next, we show that the statistical method of covariance can be used to enhance the resolution of 2D NMR spectra and thus enable *in silico* separation of BSMs in a mixture for database or *ab initio* identification using individual spectra. We then describe a comparative approach in which 2D NMR spectra from two different organisms with different genetic backgrounds are

compared to highlight unique BSMs. The final section highlights some of the powerful methods of statistically extracting specific chemical information from 1D and 2D NMR data of a particular biological state using tools of multivariate analysis. Specific technical aspects of NMR spectroscopy,[4] hyphenated techniques,[5] statistical analysis,[6,7] and overall protocol development[8] have been reviewed elsewhere.

2.1. Direct Observation of Mixtures. Two-dimensional NMR spectroscopy is arguably the most important spectroscopic technique for the elucidation of novel or unexpected structures.[4,9] Two-dimensional NMR spectra are unique in that they provide direct evidence for atom connectivity and spatial arrangements. In contrast to metabolomics, which routinely uses 1D NMR spectra for the characterization of complex, often entirely unfractionated, BSM mixtures,[10] natural product researchers traditionally relied on pure, isolated samples for analysis, whereas definite identification of novel compounds from complex mixtures was usually not pursued or deemed possible. However, recent examples have demonstrated that identification of new compounds via 2D NMR spectroscopy of complex mixtures not only is feasible but frequently offers significant advantages over fractionation-based approaches.

Development of methods for identifying novel structures via 2D NMR analyses of mixtures was motivated by several instances in which traditional activity-guided fractionation failed to reveal the BSMs of interest. In one of the first examples, the poison gland secretion of *Myrmicaria* ants was suspected to contain highly toxic alkaloids; however, chromatographic fractionation revealed only nontoxic monoterpene hydrocarbons in some samples. Subsequent 2D NMR analyses of unfractionated ant venom then revealed the highly unstable heptacyclic alkaloid myrmicarin 430A, representing one of the first examples for natural products based on the oligomeric assembly of several similarly functionalized fatty-acid chains.[11,12] Myrmicarin 430A, whose intriguing structure and biogenesis has spawned several efforts toward its total synthesis,[13] represents one of the first examples of BSMs that were identified without ever having been isolated. Similarly, a 2D NMR spectroscopic screen of a library of spider venom samples revealed that sulfated nucleosides, which despite their structural simplicity had not previously been found in nature, form major venom components in several spider species, including the infamous brown recluse, *Loxosceles reclusa* (Figure 1).[14,15] These arthropod examples showed that high-resolution 2QF-COSY spectra are well suited for detecting and characterizing unknown BSMs from complex

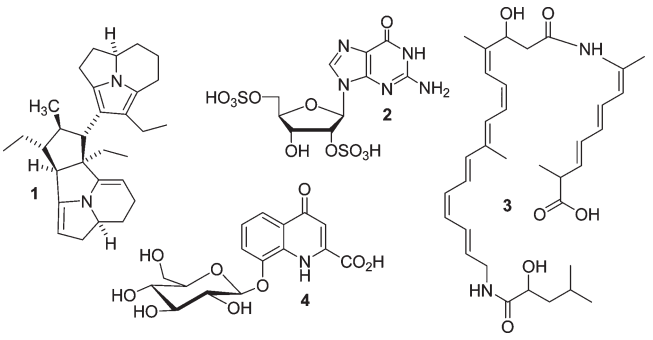


FIGURE 1. Structures of BSMs (biogenic small molecules) identified via NMR-spectroscopic analysis from largely unfractionated metabolite mixtures. Myrmicarin 430A (**1**) and bacillaene (**3**) represent members of a small but growing class of natural products that have never been isolated in pure form.[11,12,15,35] The sulfated nucleoside **2** was identified from spider venom.[15] The xanthurenic acid derivative **4** is a natriuretic identified from partially fractionated human urine samples.[49]

mixtures because they provide detailed structural information and often permit interpretation of overlapping signals, one primary challenge for using NMR spectroscopy for mixtures.

2.2. Statistical Methods for Identification of BSMs from Single Biological Mixtures. The popularity of 1D ^{1}H NMR in metabolomics and metabonomics applications is primarily founded on the simplicity and efficiency of data collection. Natural product research, on the other hand, has adopted from early on the powerful repertoire of 2D NMR techniques. Due to the narrow NMR linewidths of small molecules, 1D ^{1}H NMR spectra of mixtures can provide some rather specific information about the mixture components. They are particularly useful in situations where the pool of component candidates is limited, for example in the case of a urine sample.[16] In such cases, the observation of a single nonoverlapping peak multiplet of a mixture component permits identification of the component and determination of its relative concentration with high confidence. On the other hand, 1D ^{1}H NMR spectra of mixtures containing unknown components permit neither the determination of the 1D spectral traces that belong to individual components nor the assignment of peaks to specific atoms. In such cases, the use of 2D NMR methods becomes indispensible. *J*-resolved NMR spectra[17] are useful for simplifying overlap in 1D ^{1}H NMR but lack important atomic correlations that are the cornerstone of structure determination via 2D NMR; therefore, this approach is not included in this Account.

The main drawback of 2D Fourier transform (FT) NMR spectroscopy is the time required to collect the indirect second dimension at high digital resolution. This dimension is acquired in the time domain by repeating the same pulse

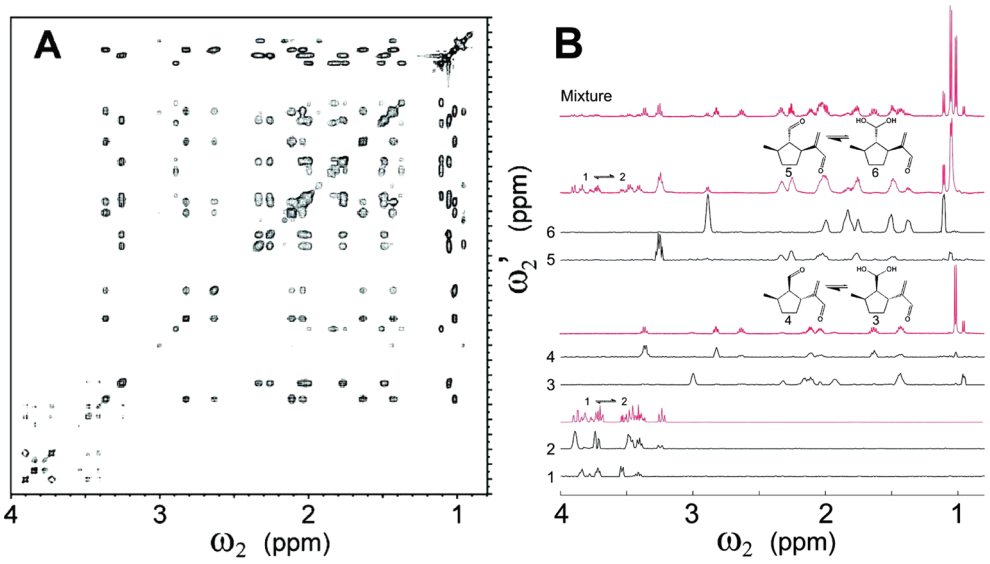


FIGURE 2. (A) Aliphatic section of covariance proton TOCSY spectrum of defensive secretion of a single walking stick insect. (B) One-dimensional ^{1}H NMR spectrum of the mixture. The six black spectra are covariance TOCSY traces extracted from covariance TOCSY of panel A using the DemixC approach. The bottom three red spectra are reference 1D spectra of purified components. Each reference spectrum contains two species, α-glucose (trace 1) and β-glucose (trace 2); dialdehyde and diol forms of the anisomorphal (traces 4 and 3, respectively); and the peruphasmal (traces 5 and 6, respectively) monoterpenes. Chemical structures of the anisomorphal and peruphasmal and their corresponding geminal diols are shown as insets. Figure is adapted from ref 25.

sequence N_1 times with the evolution time t_1 incremented from experiment to experiment. Because the digital spectral resolution is determined according to the Nyquist theorem by $\Delta\nu = \mathrm{SW}/N_1$, where SW is the spectral width, a relatively large number of increments, $N_1 > 256$, is required to obtain acceptable spectral resolution along the indirect dimension. This sampling requirement, which is independent of sensitivity, becomes more severe at higher magnetic field strengths $\mathbf{B}_0$ as $\mathrm{SW} \propto \mathbf{B}_0$.

Covariance NMR[18,19] represents an alternative approach to 2D FT that overcomes some of these limitations. Instead of applying FT along the indirect dimension, the 2D correlation information is reconstructed by statistical means through covariance yielding a symmetric spectrum **C** that has the same high resolution along the indirect and direct dimensions. In this way, the minimal number of t_1 increments required can be reduced to $N_1 < 100$.[20] It can be shown that **C** is mathematically related to the 2D FT spectrum **F** through $\mathbf{C} = sqrtm(\mathbf{F}^{\mathrm{T}} \cdot \mathbf{F})$[19] where *sqrtm* denotes the matrix square root, which can be efficiently computed via singular value decomposition (SVD).[21] The covariance method can be directly applied to TOCSY and NOESY-type spectra, whereas for COSY an additional regularization step should be applied.[22]

While for short mixing times the information content of TOCSY and COSY spectra are similar, for longer mixing times (~100 ms) typically all protons that belong to the same molecule or spin system show 2D cross-peaks to each other. This property conveniently permits the spectral deconvolution of the mixture into 1D NMR traces that correspond to individual components. Such a decomposition can be achieved by non-negative matrix factorization (NMF)[23] or bottom-up clustering of all 1D cross sections of a covariance TOCSY spectrum using the DemixC method.[24] Application of the method to the analysis of a single defensive spray milking from an adult female walking-stick insect *Anisomorpha buprestoides* using a 1-mm high-temperature superconducting NMR probe revealed that it contains, in addition to glucose, two stereoisomeric terpenes, each with a dialdehyde and a diol in slow chemical exchange (Figure 2).[25]

DemixC traces can be treated like 1D NMR spectra, and hence, they are well suited for compound identification by database searching. For this purpose, a peak list is generated for a given DemixC trace, which is then queried against the peak lists of the components of an NMR database, such as the metabolomics BMRB[26] or HMDB,[27] as implemented in the COLMAR suite of web servers[28,29] (http://spinportal.magnet.fsu.edu). NMR databases[26,27,30] are improving and expanding rapidly, and correspondingly, their importance for metabolomics and natural products research is growing. Detailed descriptions of these databases are available at the listed Web sites.

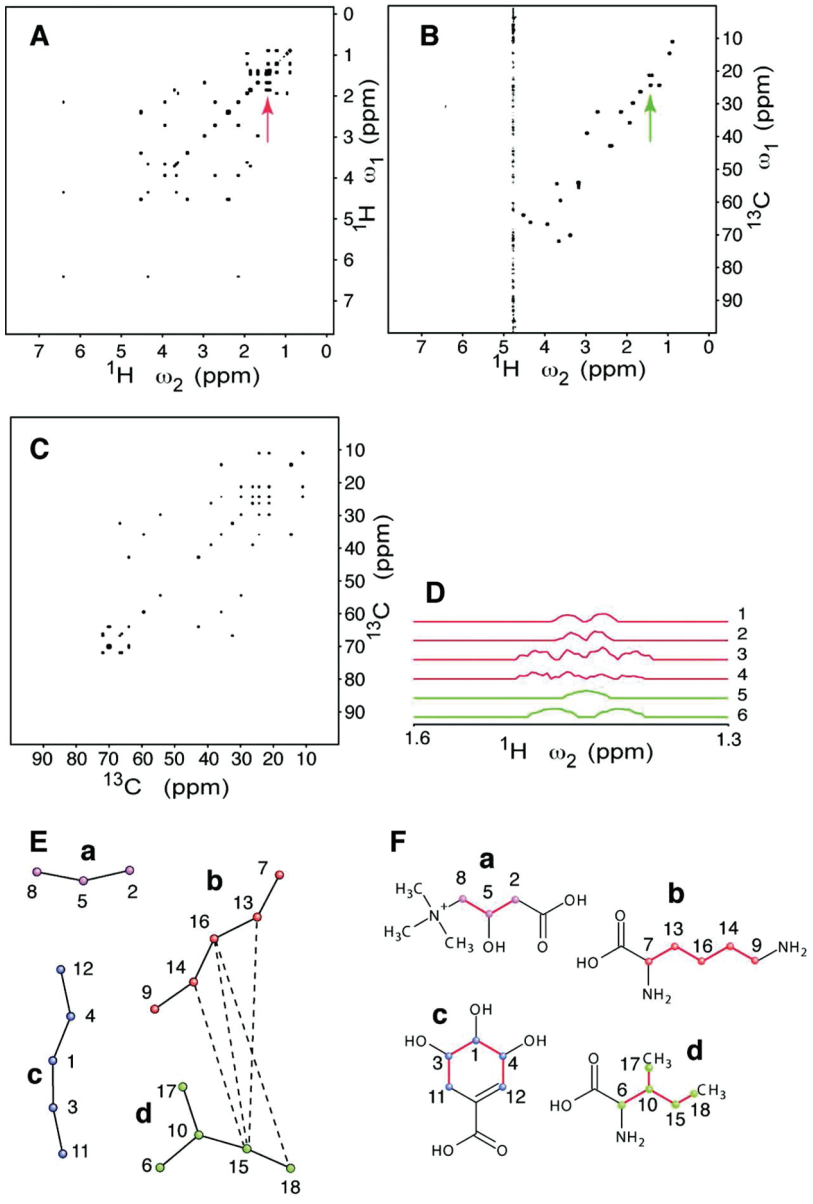


FIGURE 3. Doubly indirect covariance spectroscopy of model mixture containing carnitine, lysine, isoleucine, and shikimate: (A) 2QF-COSYspectrum; (B) ^{13}C−^{1}H HSQC spectrum; (C) doubly indirect covariance spectrum; (D) absolute value cross sections along ^{1}H dimension of COSY and HSQC for the overlap at $\omega_2 = 1.45$ ppm (traces 1−4 belong to COSY and traces 5 and 6 to HSQC); (E) carbon−carbon connectivity graphs of mixture components. Dashed lines indicate extraneous connectivities to nodes 15 and 16 due to ^{1}H overlap at 1.45 ppm as indicated by arrows in panels A and B. They are identified by filtering according to differential peak positions and shapes (1st and 2nd moments) (panel E) or, alternatively, by diffusion order spectroscopy (DOSY). The labels a, b, c, and d in panels E and F belong to carnitine, lysine, shikimate, and isoleucine, respectively. Figure is adapted from ref 33.

Database matching has also been demonstrated for heteronuclear 2D ^{13}C−^{1}H HSQC[31] and HSQC-TOCSY NMR[32] spectra of mixtures. Although at natural ^{13}C abundance the intrinsic sensitivity of HSQC is lower than for TOCSY, along the ^{13}C dimension the fully decoupled HSQC spectrum displays very sharp peaks with low probability of cross-peak overlap. Moreover, HSQC spectra only yield correlations between directly bonded ^{1}H−^{13}C nuclei and are therefore unsuitable to trace the carbon backbones of individual mixture components. This objective can be achieved by the doubly indirect covariance NMR method combining a 2D COSY spectrum **Y** with a 2D HSQC spectrum **H** via basic matrix operations: $\mathbf{D} = \mathbf{H}\cdot\mathbf{Y}\cdot\mathbf{H}^T$.[33] **D** is a symmetric ultrahigh resolution ^{13}C−^{13}C correlation spectrum that can be analyzed using graph theory to identify the carbon skeletons of individual mixture components, as shown in Figure 3. In case of proton overlap, additional connectivities do occur. They can be suppressed by an unsupervised "moment"

filtering method[34] that removes false correlations based on differential peak positions and peak widths (Figure 3). Besides the carbon backbone topology, the method also provides ^{13}C and, via HSQC, ^{1}H chemical shifts, which are useful reporters on the nature of additional chemical groups, for example, a phosphate or amino group, attached to these carbons that are not directly visible in the COSY and HSQC spectra. The doubly indirect covariance approach is designed for studying samples from uncharted territories where not only the concentrations of mixture components but also their chemical structures are not *a priori* known, thereby bridging the fields of natural product research and metabolomics.

2.3. Comparative Analysis of Biological Mixtures. Statistical techniques used in comparative metabolomics will be particularly useful for identifying new BSMs associated with specific phenotypes or genotypes, perhaps the largest looming challenge in chemical biology and natural products chemistry. For most known BSMs in nonmammalian systems, their biological functions and biosynthetic heritage are not known, and correspondingly, many genes believed to play a role in small molecule biosynthesis have remained orphans. The recent identification of the antibiotic bacillaene from *Bacillus subtilis* using differential analyses of 2D NMR spectra (DANS) illustrates the potential for adapting comparative approaches to natural products research.[35,36] Previous studies had shown that the very large pksX gene cluster (~2% of the *B. subtilis* genome) encodes an unusual hybrid PKS/NRPS (polyketide synthase/nonribosomal peptide synthetase) that produced a small molecule with antibiotic properties. Despite copious production of this metabolite, its structure had remained undetermined because all isolation attempts had failed as a result of chemical instability. Ultimately, bacillaene was identified via 2D NMR spectroscopic comparison of the unfractionated metabolomes of a pksX knockout and a pksX-expressing strain. DANS analysis of 2QF-COSY spectra obtained for these two metabolite samples enabled straightforward identification of spectral features that were present in the pksX-expressor but absent in the knockout, which formed the basis for subsequent identification of bacillaene's complete structure (Figure 1).[35]

A similar approach was used for the identification of a component of the male-attracting pheromone of the nematode *Caenorhabditis elegans*. Earlier studies based on activity-guided fractionation of worm metabolite extracts had shown that *C. elegans* hermaphrodites produce a male-attracting blend of derivatives of the dideoxysugar ascarylose.[3] However, due to strong synergism, not all components of this pheromone blend could be identified via fractionation. Based on the observation that *daf-22* mutant worms do not produce the male attractant, Pungaliya et al. employed DANS to compare the *C. elegans* wild-type and *daf-22* mutant metabolomes, which revealed the unanticipated *p*-aminobenzoic acid derivative ascr#8 as an important component of the male-attracting pheromone (Figure 4).[37] The identification of ascr#8 from *C. elegans* and bacillaene from *B. subtilis* illustrate that using comparative approaches for the analysis of whole-metabolome 2D NMR spectra offers significant opportunities for natural products chemistry and chemical biology.

2.4. Statistical Methods for Identifying BSMs from Multiple Biological Mixtures. The combination of untargeted spectral analysis of complex small molecule mixtures with statistical pattern recognition techniques for direct spectral comparison has transformed the study of metabolite associations with disease, gene function, and drug metabolism.[38,39] Quantitative comparisons of spectra from two or more distinct biological states has been a defining feature of an approach that has been referred to as "metabolomics", "metabonomics", and "metabolic profiling".[38,39] As such, the statistical data analysis methodology for comparing spectra is as significant a part of metabolomics as the underlying analytical chemistry technique that is used to generate data sets. Two major classes of statistical tools are used in metabolomics: those that identify relationships between biological states and spectral signals and those that look for correlations between spectral signals themselves.

Most metabolomics studies make use of statistical techniques to look for correlations between spectral signals and the biological states of the analyzed samples. Examples of these tools include principal component analysis (PCA), partial least squares (PLS), orthogonal projection onto latent structures (O-PLS), and discriminant analysis (PCA-DA, PLS-DA, O-PLS-DA).[6] In the general implementation of a metabolomics study, two biological states are chosen such that the majority of BSMs are unchanged, but certain BSMs that are sensitive to differences in the two biological states are quantitatively or qualitatively different. Pattern recognition tools then act as a filter to remove the spectroscopic signals from the shared metabolites from consideration while highlighting the signals arising from the metabolites present in one state or the other (Figure 5).[6]

Although the use of 2D NMR spectra in metabolomics has generally lagged behind the use of 1D NMR or mass

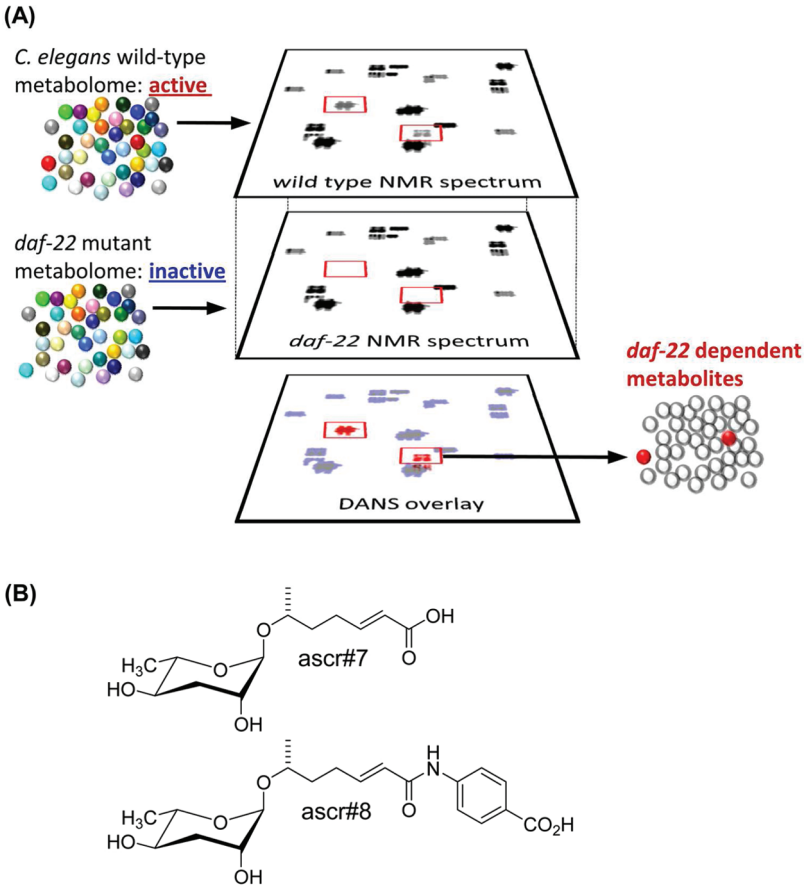


FIGURE 4. Identification of male-attracting pheromones in *C. elegans* via differential analysis of 2D NMR spectra (DANS): (A) overlay of COSY spectra from *C. elegans* wild-type and *daf-22* mutant metabolomes reveals *daf-22*-dependent signals; (B) structures of two new metabolites, ascr#7 and ascr#8, that were identified from additional analysis of the *daf-22*-dependent signals.[37]

spectra,[40] a recent example demonstrates that full-resolution 2D NMR spectra can be used for statistical analysis (e.g., via PCA) following proper alignment. Using a peak alignment algorithm termed HATS-PR (hierarchical alignment of two-dimensional spectra - pattern recognition), Robinette et al. compared sets of TOCSY spectra derived from metabolome samples of two nematode species, *Pristionchus pacificus* and *Panagrellus redivivus.*[41] The method combined the strengths of traditional 1D statistical methods and 2D NMR data, illustrating how a combination of natural products and metabolomics approaches can be used together and providing a framework for applying comparative statistical analysis to BSM discovery.

While the majority of the statistical tools in metabolomics seek to identify correlations between signals and biological states, in the past 5 years there have been a number of methods developed to identify correlations of spectroscopic signals to other signals. The best known of these methods is statistical total correlation spectroscopy (STOCSY).[42,43] While covariance TOCSY[18,19,21] (see section 2.2) identifies correlations between signals by assessing their covariances across multiple t_1-increments collected from a single sample, STOCSY uses 1D spectra of multiple biological replicates in the place of t_1-increments. Because all the signals from a given compound increase and decrease together with the concentration of the compound, peaks arising from the same compound exhibit high positive statistical correlations. Because STOCSY does not depend on physical phenomena such as magnetization transfer, it can correlate signals even across multiple spin systems that could not be correlated by TOCSY.

Though STOCSY was applied to identify correlations between protons within the same spectrum, it was quickly realized that statistical correlations could be identified across multiple types of spectra. Heteronuclear STOCSY, or Het-STOCSY, has been applied to identify $^{1}H-^{19}F$ and $^{1}H-^{31}P$ correlations from multiple biological replicates of 1D ^{1}H and 1D ^{19}F or ^{31}P spectra.[44,45] Additionally, peaks from NMR and mass spectra have been correlated using statistical heterospectroscopy (SHY).[46] This approach is particularly promising given that while ascertaining both molecular formula and the

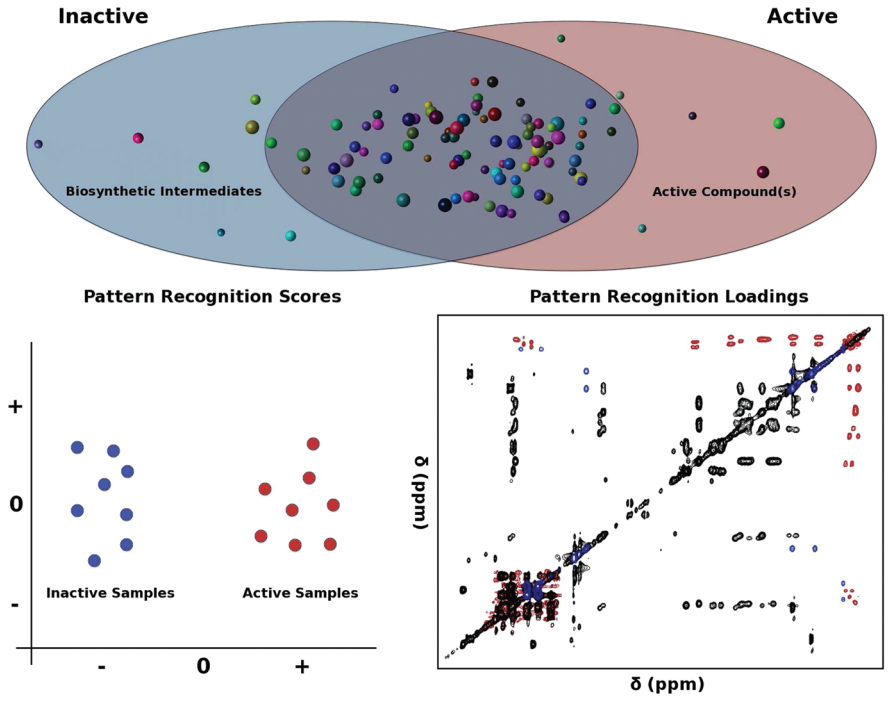


FIGURE 5. When two biological mixtures (blue and red ovals) share most compounds (colored spheres) but differ in biological activity, pattern recognition techniques applied to NMR spectra of the mixtures can identify the spectral signatures of the differential compounds. Multivariate statistical tools such as PCA and PLS decompose a set of mixture spectra into a score for each sample spectrum and loadings for each chemical shift in the NMR data set. When pattern recognition produces scores that separate the active (red circles) and inactive mixtures (blue circles), the loadings indicate which signals arise from compounds specific to the active (red cross-peaks) and inactive (blue cross-peaks) mixtures.

structural restraints encoded in NMR spectra are critical for structural elucidation, there are few methods for correlating nuclear resonances with molecular weights in mixtures.

One drawback of these techniques is that any small molecules that covary in concentration due to biological factors such as coregulation will exhibit statistical correlations as well, which can lead to ambiguity about whether a given correlation between two peaks represents a structural or biological relationship.[47] A recent extension of STOCSY termed cluster analysis statistical spectroscopy (CLASSY) has focused on using the network structure of correlations to resolve local clusters of tightly correlated peaks.[48] In the future, clustering both signals and states using two-way or biclustering techniques should provide a means of both identifying signals specific to an active biological state and grouping correlated signals as a first step in the identification of BSMs. While "signal−state" correlation tools are useful for identifying signals arising from a biologically active small molecule, they also could be extremely useful in structure elucidation.

3. Conclusion/Future Perspectives

Long-established NMR techniques in natural product research and the more recent statistical techniques of metabolomics both strive to identify BSMs and correlate them to activity. As described in this Account, several approaches using NMR take advantage of the strengths of both approaches in solving important biological problems. Advances in analytical instrumentation in NMR and mass spectrometry as well as improved databases and computational resources are fueling these new capabilities. However, a great amount of work remains. Some important challenges include the following:

- *Improved integration of NMR and MS data.* The two techniques are complementary and both are often required for BSM identification. However, they have very different sensitivities and dynamic ranges. MS is very sensitive, but not all compounds are easily detected. NMR is almost a universal detector, but it suffers from low sensitivity. Despite advances, such as SHY,[46] this is still a difficult problem that limits many studies.
- *Improved Databases.* This has improved significantly in the past few years, but the size of NMR databases is still relatively small, and many NMR chemical shifts are dependent on solvent, sample concentration, and other acquisition parameters, which makes exact matching a challenge.

- *Rapid Identification of BSMs.* Ultimately, the goal in virtually every natural products or metabolomics study is the identification of active BSMs. Acceleration and automation of the identification process remains the major bottleneck, both because of limitations described above, and perhaps more importantly, because of the great chemical diversity in BSMs. Although comparison to experimental reference spectra may enable identification of common abundant metabolites, a large number of minor and trace components, especially organism-specific metabolites, usually remain unidentified in automated analyses. There is no universal identification algorithm or protocol that can be robustly applied to all classes of molecules, and thus exists a major need for the further development of methods that can build on the recent progress made, including the work described in this Account.

This work was supported by the National Institutes of Health (Grant R01 GM066041 to R.B., Grant R01 GM088290 to F.C.S., and Grant R01 GM085285 to A.S.E.). S.L.R acknowledges support from the NSF Graduate Research Fellowship Program and from the Marshall Aid Commemoration Commission.

BIOGRAPHICAL INFORMATION

Steven L. Robinette received his B.S. in Biochemistry and Molecular Biology from the University of Florida in 2010 and is currently a Ph.D. student in Jeremy Nicholson's group at Imperial College London. He is the recipient of a Marshall Scholarship and an NSF predoctoral fellowship and is a NIH Oxford−Cambridge Scholar.

Rafael Brüschweiler studied Physics at ETH Zürich and obtained his Ph.D. in 1991 at the Laboratory for Physical Chemistry at ETH under the direction of Prof. Richard R. Ernst. He then joined the Scripps Research Institute, La Jolla, California, as a postdoctoral fellow before returning to ETH as an Oberassistent. In 1998, he accepted the Carlson Chair position of Chemistry and Biochemistry at Clark University, Worcester, Massachusetts, before joining Florida State University as George M. Edgar Professor of Chemistry and Biochemistry and the National High Magnetic Field Laboratory as Associate Director for Biophysics in 2004. His interests in Biophysical Chemistry focus on the development and application of NMR and computational methods for the analysis of the structure, dynamics, interactions, and function of biological molecules.

Frank C. Schroeder received undergraduate degrees in chemistry and physics and conducted his Ph.D. research in natural product chemistry with Wittko Francke at the University of Hamburg, Germany. He continued with postdoctoral studies in Jerrold Meinwald's laboratory at Cornell University and subsequently joined the laboratory of Jon Clardy at Harvard Medical School, where he developed NMR spectroscopic approaches for the identification and functional characterization of small molecules in model organisms. In 2007, he joined the faculty of the Boyce Thompson Institute and Department of Chemistry and Chemical Biology at Cornell University.

Arthur S. Edison obtained a B.S. in chemistry from the University of Utah where he conducted research with William Epstein and David Grant. He completed his Ph.D. in biophysics from the University of Wisconsin—Madison under the supervision of John L. Markley and Frank Weinhold. In 1993, Dr. Edison joined the laboratory of Anthony O. W. Stretton at the University of Wisconsin—Madison as a Jane Coffin Childs postdoctoral fellow, and he joined the faculty at the University of Florida and the National High Magnetic Field Laboratory in 1996 where he is currently Professor of Biochemistry & Molecular Biology and Director of Chemistry & Biology at the NHMFL.

FOOTNOTES

*Corresponding author. E-mail address: aedison@ufl.edu.

REFERENCES

1 Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461–477.

2 He, W.; Miao, F. J.; Lin, D. C.; Schwandner, R. T.; Wang, Z.; Gao, J.; Chen, J. L.; Tian, H.; Ling, L. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature* **2004**, *429*, 188–193.

3 Srinivasan, J.; Kaplan, F.; Ajredini, R.; Zachariah, C.; Alborn, H. T.; Teal, P. E.; Malik, R. U.; Edison, A. S.; Sternberg, P. W.; Schroeder, F. C. A blend of small molecules regulates both mating and development in Caenorhabditis elegans. *Nature* **2008**, *454*, 1115–1118.

4 Edison, A. S.; Schroeder, F. C.; Lew, M.; Hung-Wen, L.: NMR - Small Molecules and Analysis of Complex Mixtures. In *Comprehensive Natural Products II*; Elsevier: Oxford, 2010; pp 169−196.

5 Wolfender, J. L.; Ndjoko, K.; Hostettmann, K. Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. *J. Chromatogr. A* **2003**, *1000*, 437–455.

6 Madsen, R.; Lundstedt, T.; Trygg, J. Chemometrics in metabolomics--a review in human disease diagnosis. *Anal. Chim. Acta* **2010**, *659*, 23–33.

7 Steuer, R.; Morgenthal, K.; Weckwerth, W.; Selbig, J. A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.* **2007**, *358*, 105–126.

8 Kim, H. K.; Choi, Y. H.; Verpoorte, R. NMR-based metabolomic analysis of plants. *Nat. Protoc.* **2010**, *5*, 536–549.

9 Forseth, R. R.; Schroeder, F. C. NMR-spectroscopic analysis of mixtures: from structure to function. *Curr. Opin. Chem. Biol.* **2011**, *15*, 38–47.

10 Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2007**, *2*, 2692–2703.

11 Schroder, F.; Sinnwell, V.; Baumann, H.; Kaib, M.; Francke, W. Myrmicarin-663: A new decacyclic alkaloid from ants. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 77–80.

12 Schroder, F.; Sinnwell, V.; Baumann, H.; Kaib, M. Myrmicarin 430A: A new heptacyclic alkaloid from Myrmicaria ants. *Chem. Commun.* **1996**, 2139–2140.

13 Snyder, S. A.; Elsohly, A. M.; Kontes, F. Synthetic and theoretical investigations of myrmicarin biosynthesis. *Angew. Chem., Int. Ed. Engl.* **2010**, *49*, 9693–9698.

14 Taggi, A. E.; Meinwald, J.; Schroeder, F. C. A new approach to natural products discovery exemplified by the identification of sulfated nucleosides in spider venom. *J. Am. Chem. Soc.* **2004**, *126*, 10364–10369.

15 Schroeder, F. C.; Taggi, A. E.; Gronquist, M.; Malik, R. U.; Grant, J. B.; Eisner, T.; Meinwald, J. NMR-spectroscopic screening of spider venom reveals sulfated nucleosides as major components for the brown recluse and related species. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14283–14287.

16 Liao, W. I.; Lin, Y. Y.; Chu, S. J.; Hsu, C. W.; Tsai, S. H. Bradyarrhythmia caused by ginseng in a patient with chronic kidney disease. *Am. J. Emerg. Med.* **2010**, *28*, 538.e5–538.e6.

17 Viant, M. R. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Commun.* **2003**, *310*, 943–948.

18 Brüschweiler, R.; Zhang, F. Covariance nuclear magnetic resonance spectroscopy. *J. Chem. Phys.* **2004**, *120*, 5253–5260.

19 Brüschweiler, R. Theory of covariance nuclear magnetic resonance spectroscopy. *J. Chem. Phys.* **2004**, *121*, 409–414.

20 Chen, Y.; Zhang, F.; Bermel, W.; Brüschweiler, R. Enhanced covariance spectroscopy from minimal datasets. *J. Am. Chem. Soc.* **2006**, *128*, 15564–15565.

21 Trbovic, N.; Smirnov, S.; Zhang, F.; Brüschweiler, R. Covariance NMR spectroscopy by singular value decomposition. *J. Magn. Reson.* **2004**, *171*, 277–283.

22 Chen, Y.; Zhang, F.; Snyder, D.; Gan, Z.; Bruschweiler-Li, L.; Brüschweiler, R. Quantitative covariance NMR by regularization. *J. Biomol. NMR* **2007**, *38*, 73–77.

23 Snyder, D. A.; Zhang, F.; Robinette, S. L.; Bruschweiler-Li, L.; Brüschweiler, R. Non-negative matrix factorization of two-dimensional NMR spectra: application to complex mixture analysis. *J. Chem. Phys.* **2008**, *128*, No. 052313.

24 Zhang, F.; Brüschweiler, R. Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angew. Chem., Int. Ed.* **2007**, *46*, 2639–2642.

25 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*, 7748–7752.

26 Markley, J. L.; Anderson, M. E.; Cui, Q.; Eghbalnia, H. R.; Lewis, I. A.; Hegeman, A. D.; Li, J.; Schulte, C. F.; Sussman, M. R.; Westler, W. M.; Ulrich, E. L.; Zolnai, Z. New bioinformatics resources for metabolomics. *Pac. Symp. Biocomput. 2007* **2007**, 157–168.

27 Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610.

28 Robinette, S. L.; Zhang, F.; Bruschweiler-Li, L.; Brüschweiler, R. Web server based complex mixture analysis by NMR. *Anal. Chem.* **2008**, *80*, 3606–3611.

29 Zhang, F.; Robinette, S. L.; Bruschweiler-Li, L.; Brüschweiler, R. Web server suite for complex mixture analysis by covariance NMR. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S118–S122.

30 Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008**, *26*, 162–164.

31 Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. Method for determining molar concentrations of metabolites in complex solutions from two-dimensional $^{1}H-^{13}C$ NMR spectra. *Anal. Chem.* **2007**, *79*, 9385–9390.

32 Zhang, F.; Bruschweiler-Li, L.; Robinette, S. L.; Brüschweiler, R. Self-consistent metabolic mixture analysis by heteronuclear NMR. Application to a human cancer cell line. *Anal. Chem.* **2008**, *80*, 7549–7553.

33 Zhang, F.; Bruschweiler-Li, L.; Brüschweiler, R. Simultaneous de novo identification of molecules in chemical mixtures by doubly indirect covariance NMR spectroscopy. *J. Am. Chem. Soc.* **2010**, *132*, 16922–16927.

34 Bingol, K.; Salinas, R. K.; Brüschweiler, R. Higher-rank correlation NMR spectra with spectral moment filtering. *J. Phys. Chem. Lett.* **2010**, *1*, 1086–1089.

35 Butcher, R. A.; Schroeder, F. C.; Fischbach, M. A.; Straight, P. D.; Kolter, R.; Walsh, C. T.; Clardy, J. The identification of bacillaene, the product of the PksX megacomplex in Bacillus subtilis. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1506–1509.

36 Moldenhauer, J.; Chen, X. H.; Borriss, R.; Piel, J. Biosynthesis of the antibiotic bacillaene, the product of a giant polyketide synthase complex of the trans-AT family. *Angew. Chem., Int. Ed.* **2007**, *46*, 8195–8197.

37 Pungaliya, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. A shortcut to identifying small molecule signals that regulate behavior and development in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708–7713.

38 Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171.

39 Nicholson, J. K. L. J. C.; Holmes, E. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis in biological NMR spectroscopic data. *Xenobiotica* **2000**, *9*, 1181–1189.

40 Zheng, M.; Lu, P.; Liu, Y.; Pease, J.; Usuka, J.; Liao, G.; Peltz, G. 2D NMR metabonomic analysis: a novel method for automated peak alignment. *Bioinformatics* **2007**, *23*, 2926–2933.

41 Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S. Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Anal. Chem.* **2011**, *83*, 1649–1657.

42 Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Anal. Chem.* **2005**, *77*, 1282–1289.

43 Holmes, E.; Cloarec, O.; Nicholson, J. K. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: Application to $HgCl_2$ toxicity. *J. Proteome Res.* **2006**, *5*, 1313–1320.

44 Coen, M.; Hong, Y. S.; Cloarec, O.; Rhode, C. M.; Reily, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Heteronuclear $^{1}H-^{31}P$ statistical total correlation NMR spectroscopy of intact liver for metabolic biomarker assignment: Application to galactosamine-induced hepatotoxicity. *Anal. Chem.* **2007**, *79*, 8956–8966.

45 Keun, H.; Athersuch, T.; Beckonert, O.; Wang, Y.; Saric, J.; Shockcor, J.; Lindon, J.; Wilson, I.; Holmes, E.; Nicholson, J. Heteronuclear $^{19}F-^{1}H$ Statistical Total Correlation Spectroscopy as a Tool in Drug Metabolism: Study of Flucloxacillin Biotransformation. *Anal. Chem.* **2008**, *80*, 1073–1079.

46 Crockford, D.; Holmes, E.; Lindon, J.; Plumb, R.; Zirah, S.; Bruce, S.; Rainville, P.; Stumpf, C.; Nicholson, J. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Anal. Chem.* **2006**, *78*, 363–371.

47 Alves, A. C.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. Analytic properties of statistical total correlation spectroscopy based information recovery in ^{1}H NMR metabolic data sets. *Anal. Chem.* **2009**, *81*, 2075–2084.

48 Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M. D.; 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*, 6581–6589.

49 Cain, C. D.; Schroeder, F. C.; Shankel, S. W.; Mitchnick, M.; Schmertzler, M.; Bricker, N. S. Identification of xanthurenic acid 8-O-beta-D-glucoside and xanthurenic acid 8-O-sulfate as human natriuretic hormones. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17873–17878.