Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Chemical Biology

NMR-spectroscopic analysis of mixtures: from structure to function

Ry R Forseth and Frank C Schroeder

NMR spectroscopy as a particularly information-rich method offers unique opportunities for improving the structural and functional characterization of metabolomes, which will be essential for advancing the understanding of many biological processes. Whereas traditionally NMR spectroscopy was mostly relegated to the characterization of pure compounds, the past few years have seen a surge of interest in using NMR-spectroscopic techniques for characterizing complex metabolite mixtures. Development of new methods was motivated partly by the realization that using NMR for the analysis of metabolite mixtures can help identify otherwise inaccessible small molecules, for example compounds that are prone to chemical decomposition and thus cannot be isolated. Furthermore, comparative metabolomics and statistical analyses of NMR spectra have proven highly effective at identifying novel and known metabolites that correlate with changes in genotype or phenotype. In this review, we provide an overview of the range of NMR-spectroscopic techniques recently developed for characterizing metabolite mixtures, including methods used in discovery-oriented natural product chemistry, in the study of metabolite biosynthesis and function, or for comparative analyses of entire metabolomes.

Address
Boyce Thompson Institute and Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA

Corresponding author: Schroeder, Frank C (schroeder@cornell.edu)

Current Opinion in Chemical Biology 2011, **15**:38–47

This review comes from a themed issue on
Omics
Edited by Kate Carroll and Pieter Dorrestein

Available online 9th November 2010

1367-5931/$ – see front matter

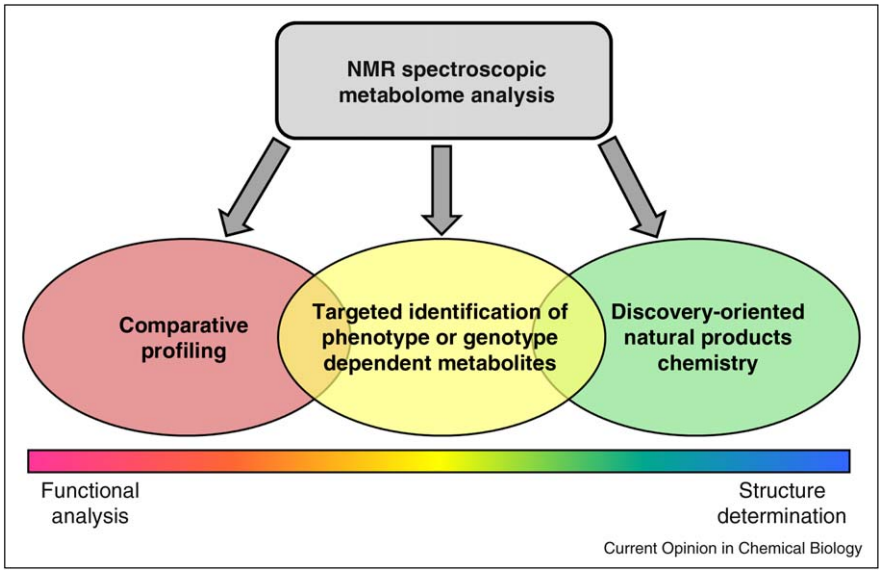
DOI 10.1016/j.cbpa.2010.10.010

Introduction

The identification and functional characterization of biogenic small molecules remains one of the most challenging tasks in chemical biology [1,2•]. Despite great advances in analytical techniques such as NMR spectroscopy or mass spectrometry, determining the structures and biological roles of non-trivial metabolites usually requires a major research effort, and even if structures of compounds can be identified, success in determining their biological function is often incomplete. Compared to recent progress in genomics and proteomics, our ability to structurally and functionally characterize the entirety of an organism's endogenous small molecules – the organism's metabolome – is still woefully limited [3]. This deficiency may seem surprising, given that knowledge of the structures of some important biogenic small molecules, for example steroid hormones, actually predates understanding of structures and functions of proteins and nucleic acids. However, the metabolomes of even the most well-studied life forms – model organisms such as yeast, Drosophila, the nematode *Caenorhabditis elegans*, or mice – have been explored only to a very limited extent, principally as a result of the daunting challenges associated with identifying the often unexpected structures, biosynthetic pathways, and functional roles of thousands of compounds. One principal reason for proteomics and genomics leaping ahead of metabolomics lies in the high degree of structural diversity (or irregularity) of biogenic small molecules. Proteins and nucleic acids are template-derived and thus enable approaches for determining structures and functions for which there are simply no equivalents in the world of small molecules.

Correspondingly, there is considerable interest in developing new, more comprehensive approaches for structural identification of biogenic small molecules and their functional characterization. Increased sensitivity and dynamic range of MS-based methods have enabled rapid profiling of metabolite samples for detection and quantification of known compounds, and routine HPLC–MS analyses can now analyze large arrays of samples for the presence of thousands of compounds [3]. Although NMR spectroscopy can also make important contributions to metabolic profiling [2•,4], one additional strength lies in its utility for the identification of unknown or unexpected compounds, in this regard complementing MS-based approaches. Analysis of multi-dimensional NMR spectra provides *structural information*, that is, a skilled spectroscopist can directly infer atom connectivity and spatial arrangements from the NMR-spectroscopic data [5]. In addition, NMR spectroscopy is distinguished from MS in that it is less biased: the results of MS-based analyses greatly depend on choice of ionization conditions and the specific instrumentation used [6,7], and in fact the most commonly used ionization technique, electrospray ionization, markedly favors detection of specific compound classes of high polarity, introducing a strong bias against many less polar classes of metabolites, for example many lipids and steroids. By contrast, the relative strength of

Figure 1



The range of applications for NMR-spectroscopic analyses of metabolite mixtures can be categorized roughly into three areas. Screening for new natural products (green), structure determination of specific metabolites that are functionally connected with genotypic or phenotypic changes (yellow), and comparative profiling for the identification of biomarkers or assessment of general metabolic changes (red). The overlap of the ovals represents how NMR-spectroscopic concepts and specific techniques find applicability in adjoining areas. The types of information sought by the three analytical vantages vary with regard to the emphasis placed on identifying new structures or associating compounds with biological activities, as indicated by placing them along a color-coded gradient. At the 'structure' end of the spectrum (blue) fall purely discovery-oriented natural product initiatives, whereas moving left across the gradient, emphasis moves away from procuring new structures towards associating structures with function.

signals in NMR spectra is usually proportional to the relative amounts of the compounds these signals represent, as long as all compounds are soluble in the NMR solvent and sufficiently stable under the conditions of analysis (two occasionally non-trivial requirements).

Despite their desirable qualities and high information content, NMR spectra used to be employed primarily as a tool for identifying compounds from pure samples. Early biological examples have used NMR spectroscopy for metabolic profiling of mixtures [8], but NMR spectroscopy was not commonly accepted as an appropriate means for the definitive identification of novel or unexpected metabolites from mixtures. Only recently NMR-spectroscopic methods have begun to play a larger role in the identification of previously unknown or unexpected compounds in complex small-molecule mixtures. Development of methods for NMR-spectroscopic characterization of mixtures was driven partly by increases in sensitivity of NMR spectrometers [9] and advances in data processing, but also the realization that whole metabolome analyses via NMR can help identify otherwise inaccessible small molecules, for example compounds that are prone to chemical decomposition and thus cannot be isolated [10,11•,12]. Furthermore, it has become apparent that NMR spectroscopy-based metabolome analyses can be highly effective at identifying novel and known metabolites that correlate with changes in genotype or phenotype [2•,13,14•].

Figure 1 illustrates the current range of applications for NMR-spectroscopic analyses of small-molecule mixtures, from comparative metabolic profiling to applications in discovery-oriented natural products chemistry. In this review, we focus on recent developments that extend the use of NMR-spectroscopic mixture analysis to the search for new structural entities and the targeted identification of metabolites associated with specific phenotypes or genotypes. In addition, we discuss computational methods for deconvoluting complex 2D NMR spectra that may represent stepping stones toward future automated analyses of NMR-spectroscopic data.

NMR-based compound identification from mixtures: a paradigm shift

Until recently identification of new biogenic small molecules – usually obtained as part of a complex metabolite mixture – largely relied on the investigator's ability to isolate the compound of interest chromatographically. Chromatographic fractionation, often guided by assays for specific biological activities, followed by NMR-spectroscopic characterization of pure, isolated compounds constitutes the traditional approach of structure determination in natural products chemistry and chemical

biology. Usually, achieving a high degree of purity was deemed essential for the success of subsequent NMR-spectroscopic analyses [15]. However, this approach obviously excluded any metabolites not robust enough to survive chromatography. Moreover, recent examples show that activity-guided isolation and characterization of metabolites from complex mixtures may overlook biologically important compounds, for example because of synergistic action of several components [14•].

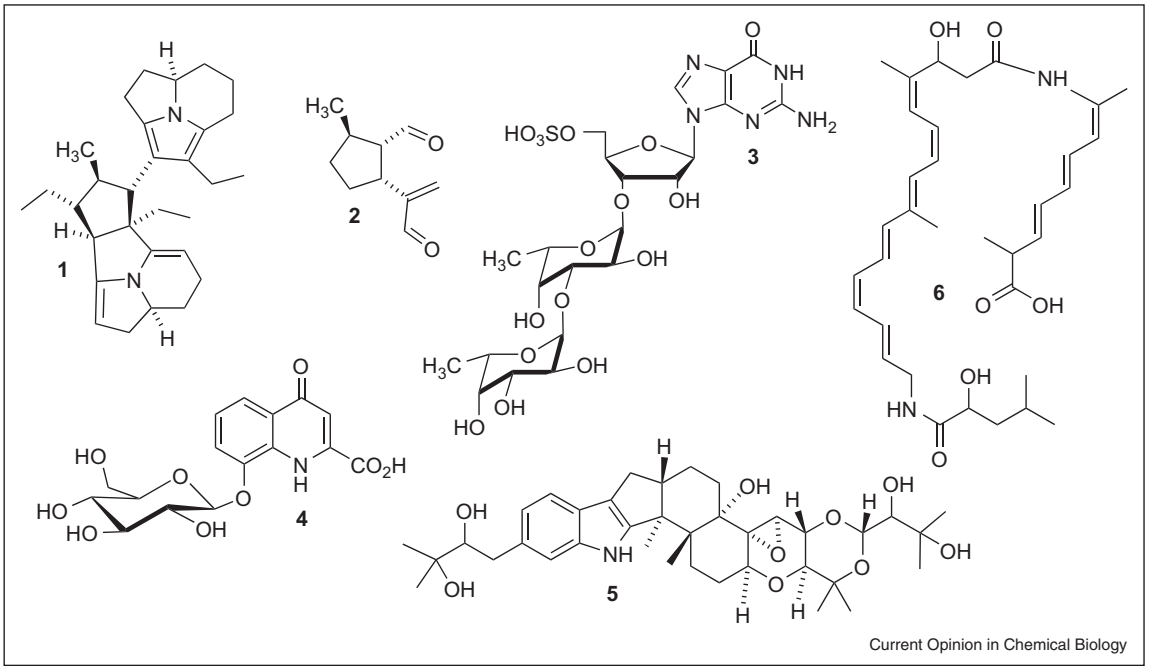
Identification of unstable alkaloids from unfractionated biofluids

The first examples for using NMR spectroscopy to identify new compounds from complex mixtures originated from arthropod natural products studies [16,17]. *Myrmicaria* ants produce copious amounts of a highly toxic poison gland secretion, however, mass spectrometric analyses of some secretion samples revealed only non-toxic monoterpene hydrocarbons. It was then observed that the secretion rapidly lost its toxic properties after collection, likely due to decomposition upon exposure to air. Consequently, fresh, unfractionated secretion was subjected to NMR-spectroscopic analyses, which lead to the identification of myrmicarin 430A (**1**, Figure 2) and related structures, representing a new family of heptacyclic alkaloids. Instrumental for the identification of myrmicarin 430A from the unfractionated ant secretion, which also contained large quantities of monoterpene hydrocarbons and other alkaloids, was the use of high-resolution dqfCOSY spectra [17]. Because dqfCOSY crosspeaks contain detailed coupling constant information and due to their high fidelity (excellent peak shape), structural assignments were possible even in cases were signals overlapped or signals of very low intensity had to be discerned.

Single insect NMR

Dossey *et al.* extended the scope of 2D NMR-spectroscopic analyses of unfractionated biofluids by demonstrating the utility of reduced-volume probes to increase mass sensitivity. Using a 1-mm HTS cryogenic probe that afforded 25-fold greater sensitivity than conventional probes at that time [18], unfractionated defensive secretion from individual male walking sticks (*Anisomorpha buprestoides*) was analyzed via COSY, TOCSY, ROESY, and natural abundance (^{1}H,^{13}C)-HMQC and HMBC [19]. Based on these spectra, the walking stick secretion was shown to consist of glucose and a mixture of monoterpene dialdehydes that are stereoisomers of the known dolichodial (**2**, Figure 2). Using the HTS cryogenic probe, as little as 1 μL of walking stick secretion could be characterized, enabling analyses of individual insect specimen. The study established that individual insects, even when grown under identical conditions, show highly variable venom content, demonstrating the potential relevance of chemical diversity at the level of individual animals. For the case of *A. buprestoides*, the authors further demonstrated the use of a computational

Figure 2



Structures of new natural products identified via NMR-spectroscopic analyses of complex mixtures. Myrmicarin 430A (**1**) and bacillaene (**6**) represent members of a small but growing class of metabolites that have never been isolated in pure form.

deconvolution procedure for the interpretation of TOCSY spectra of mixtures ('DemixC') which will be discussed in a following section.

In addition to known compounds, Dossey *et al.* discovered novel metabolites using 2D NMR of unfractionated extracts. Using the same low-volume HTS cryogenic probe, the defensive spray of the walking stick *Parectatosoma mocquerysi* was found to contain the novel monoterpene parectadial [20]. After partitioning the defensive spray between D_2O and benzene-d_6, COSY, (^{1}H,^{13}C)-HMQC, (^{1}H,^{13}C)-HMBC, and NOESY acquired for the otherwise unfractionated extract allowed to unambiguously establish the structure of parectadial.

Whole spider-venom analysis reveals sulfated nucleosides

Due to the modest complexity of the *Myrmicaria* and walking stick extracts, containing only two or three different compound families, these examples proved to be ideal test cases for exploring the utility of 2D NMR spectroscopy for the identification of new compounds from mixtures. Yet another arthropod study explored applicability of 2D NMR spectroscopy toward structural analysis of much more complex metabolite samples: spider venom. Spider venoms were known to constitute complex mixtures of proteins, peptides, acylated polyamines, and various small-molecule neurotransmitters, which had been identified via extensive LC–MS analyses and activity-guided fractionation. However, the surprising discovery of a sulfated nucleoside, HF-7 [21], as a component of the venom of the funnel web spider *Hololena curta* suggested that important spider venom components may have been missed by earlier analyses, possibly due to loss of chemically labile components during chromatographic fractionation of the venom. Therefore, Taggi and co-workers decided to screen a small library of fresh, unfractionated spider venom samples using 2D NMR spectroscopy for comprehensive, unbiased assessment of venom composition. These analyses led to the identification of sulfated nucleosides (e.g. **3**, Figure 2), as major venom components in several spider species, including the hobo spider, *Tegenaria agrestis*, and the infamous brown recluse spider, *Loxosceles reclusa* [10,11•]. The study demonstrated that using high-resolution dqfCOSY spectra, new compounds can be detected and identified even in complex metabolite mixtures, and moreover, that 2D NMR-spectroscopic analyses can provide a comprehensive, unbiased overview of the structural diversity within a metabolite sample.

A natriuretic hormone from 2D NMR spectra of human urine

The aforementioned examples showed that characterization of small-molecule mixtures by 2D NMR spectroscopy is particularly useful in situations where the compounds of interest appear to be unstable or otherwise unsuitable for chromatography. The identification of a human natriuretic hormone by Cain *et al.* demonstrated that this approach also offers advantages when used to abbreviate bioassay-guided fractionation schemes. Bioassay-guided fractionation of human urine – representing a highly complex mixture of metabolites – had allowed Cain *et al.* to obtain a fraction with significant natriuretic activity; however, attempts at further purification of the active components were unsuccessful as activity appeared to decrease after additional chromatography [22]. Therefore, a partially purified urine fraction with good natriuretic activity was characterized extensively using 2D NMR spectra including dqfCOSY, (^{1}H,^{13}C)-HMQC and (^{1}H,^{13}C)-HMBC. In combination with high-resolution ESI$^+$ mass spectra, these data allowed identification of three potential candidates for the sought-after active component, 2-(6-amino-3-hydroxyphenyl)-2-hydroxypropanoic acid, xanthurenic acid 8-*O*-β-D-glucoside (**4**, Figure 2), and xanthurenic acid 8-*O*-sulfonate, all three of which, intriguingly, had not previously been described as human metabolites. From the observation that natriuretic activity of the active urine fractions did not appear to correlate with the amounts of 2-(6-amino-3-hydroxyphenyl)-2-hydroxypropanoic acid present, it was hypothesized that the xanthurenic acid derivatives represented the active components. Independent synthesis of these compounds confirmed the structures assigned based on the NMR-spectroscopic mixture analysis and confirmed their strong activity in natriuretic hormone assays. This example shows that when conducting bioassay-guided fractionation for compound identification, it is useful to weigh the benefits of (1) possibly achieving further enrichment of an active component with additional fractionation, against (2) carrying out NMR-based structural analysis on partially purified fractions and using this information to choose the most likely molecular candidates responsible for the observed activity.

New compounds from 2D NMR-based comparison of different phenotypes or genotypes

The studies described in the previous section employ 2D NMR-spectroscopic analysis of crude metabolite mixtures to identify components of interest without extensive fractionation. However, this approach does not address today's perhaps biggest challenge in chemical biology: to develop better methods for connecting identified small molecules with their biological functions and biosynthesis. Various approaches are being pursued to conquer this challenge via comparative metabolomics, based on metabolic profiling data obtained from a variety of mass spectrometric and NMR-spectroscopic methods [4,23–25]. By comparing the metabolomes of organisms exhibiting different phenotypes or representing different genotypes, it is often possible to infer the biological role of individual metabolites or determine their biosynthetic heritage.

Several recent studies extend the scope of NMR-based comparative metabolomics beyond comparative profiling of known metabolites. In the following examples, 2D NMR-based comparison of metabolite samples derived from different genotypes or phenotypes is used to identify previously unknown or unexpected compounds and place them into a specific biological context.

New fungal metabolites from NMR-spectroscopic whole-metabolome analyses

Differential analyses of 2D NMR spectra (DANS), a method for graphic comparison of 2D NMR spectra representing different biological states, was first applied to a small library of metabolite extracts derived from cultures of the filamentous fungus, *Tolypocladium cylindrosporum*. This study aimed to detect and identify products of PKS/NRPS pathways induced by specific environmental conditions, for example nutritional stress, and used DANS to compare dqfCOSY spectra obtained for metabolite extracts derived from seven different culturing protocols [13]. The dqfCOSY spectra representing different conditions were overlayed and processed using a simple algorithm that highlighted spectroscopic signals that are strongly differential between spectra, that is signals that represent compounds whose biosynthesis is strongly induced under specific culturing conditions. This approach led to detection and identification of two new indole alkaloids, TC-705A (**5**, Figure 2) and TC-705B. These structures were proposed based on NMR-spectroscopic analyses of the unfractionated extracts and subsequently were confirmed through additional spectroscopic analyses of chromatographically enriched samples of TC-705A and TC-705B. In addition to highlighting those extracts that contained structurally interesting new metabolites, the DANS analyses of the unfractionated fungal metabolite extracts continued to be of use, as (1) the fractionation could be guided by NMR-spectroscopic signals recognized as representing differentially expressed compounds, and (2) the original DANS spectroscopic data provided positive evidence that the enriched metabolites were not artifacts of the isolation process.

Identification of the product of an orphan gene cluster via DANS

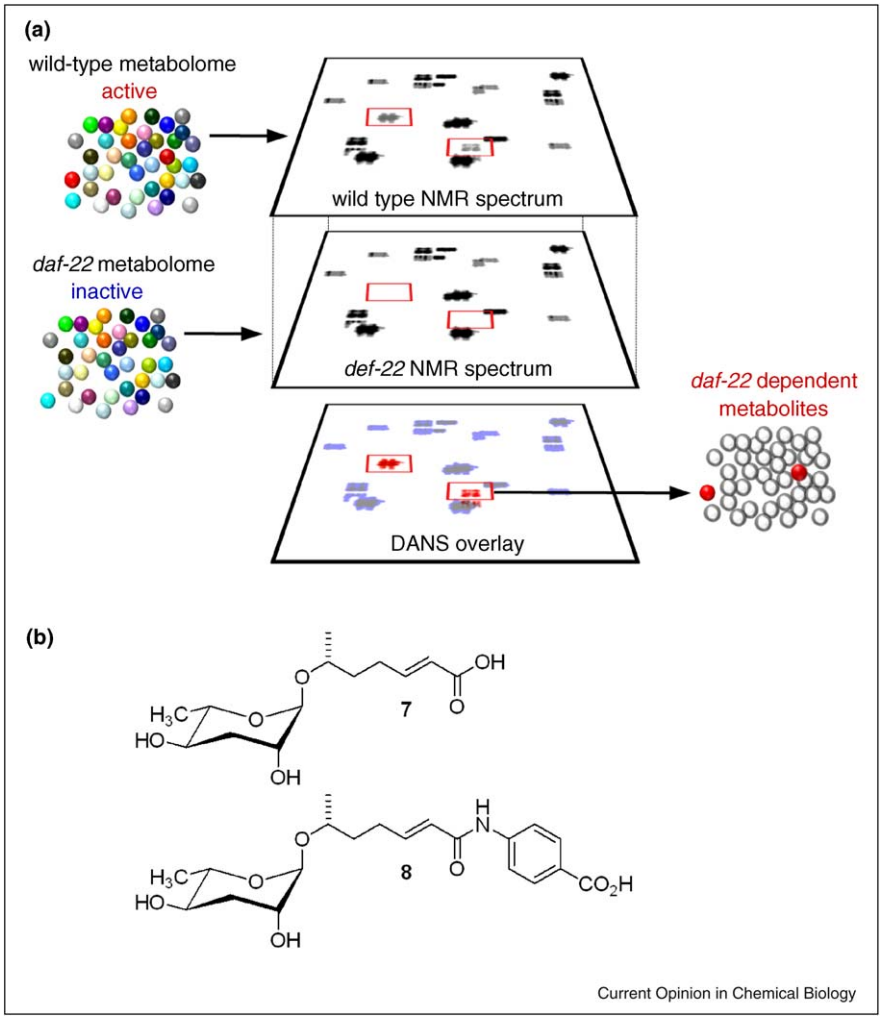
In a second example, DANS was used to detect and identify the product of the orphan gene cluster pksX in *Bacillus subtilis*. Previous work has shown that this very large (~80 kb) pksX gene cluster encodes an unusual hybrid polyketide/nonribosomal peptide synthase that produces an antibiotic metabolite, named bacillaene. However, the structure of bacillaene remained elusive, because attempts at chromatographic isolation of bacillaene were unsuccessful and the complex nature of the pksX gene cluster precluded a bioinformatic prediction of the structure [12]. Therefore, Clardy and co-workers pursued identification of bacillaene based on NMR spectra of largely unfractionated bacterial extracts. dqfCOSY spectra obtained for pksX-expressing strains and corresponding knock-out strains were compared via DANS, which revealed several structural fragments that were present in $pksX^+$ extracts but absent in $pksX^-$ extracts. Additional (^1H,^{13}C)-HMQC, (^1H,^{13}C)-HMBC and (^1H,^{15}N)-HMBC, ROESY, and high-resolution MS data were acquired for the $pksX^+$ extracts and provided the information needed to complete structural assignments, revealing the polyene bacillaene (**6**, Figure 2) and several derivatives. Given the presence of a highly unstable conjugated polyene, elimination-prone β-hydroxy amide, and conjugated polyene-amide functionalities, the bacillaenes must be considered as highly reactive, explaining the failure of attempts to isolate these compounds chromatographically. In this example, DANS proved highly effective at linking genotype (pksX expression), phenotype (antibiotic activity) and small-molecule structure.

Using DANS for the identification of signaling molecules

The success of DANS methodology for the characterization of the bacillaenes inspired application of this approach to investigate small-molecule signaling in the model organism *Caenorhabditis elegans* [14•]. Earlier work had shown that small-molecule extracts of *C. elegans* cultures have potent biological activity in two different assays. These extracts were shown (1) to induce arrest in developing *C. elegans* larvae at the so-called 'dauer' stage, a highly enduring, non-feeding larval stage, and (2) to act as a pheromone by attracting male *C. elegans*. Activity-guided fractionation of *C. elegans* metabolite extracts led to the identification of a series of glycosides of the dideoxy sugar ascarylose, the ascarosides ascr#2–ascr#4 [26,27]. These compounds showed significant activity as dauer inducing signals or male attractants, but failed to reproduce the activity of native *C. elegans* extracts in full, suggesting that important components of the mating and dauer pheromones had been missed.

A more complete characterization of the dauer and mating signals was achieved using DANS-based comparison of the metabolomes of wild-type *C. elegans* and a signaling-defective mutant strain, *daf-22*. *daf-22* mutant worms had been shown to be deficient in production of both the dauer and mating signals, suggesting that comparison of wild-type and *daf-22*-mutant metabolomes could reveal the sought-after signaling molecules. For DANS, dqfCOSY spectra obtained for wild-type and *daf-22* worms were overlayed and processed to highlight signals present in wild-type but entirely absent in *daf-22* extracts (Figure 3). This analysis detected a small number of metabolites in the wild-type extracts that were not produced by *daf-22* worms, including the already known ascarosides ascr#2–ascr#4 and several previously undescribed compounds, which represented likely candidates for the sought-after pheromone components. Each of

Figure 3



Schematic representation of differential 2D NMR spectroscopy (DANS) applied to comparing *C. elegans* wild-type metabolome with that of *daf-22* mutant worms [14•]. Reprinted from [14] with permission from the National Academy of Sciences, USA.

these *daf-22*-dependent metabolites accounted for much less than 0.1% of the entire wild-type metabolite sample, demonstrating that proton-detected 2D NMR spectroscopy offers sufficiently high dynamic range to detect even trace amounts of signaling molecules. As a result of this study, four additional ascarosides were identified, including the unusual *p*-aminobenzoic acid derivative ascr#8, which was shown to represent an important component of both the dauer and mating signals [14•].

Computational approaches for NMR-based compound identification from mixtures

2D NMR spectra of small-molecule mixtures are extremely information-rich, promising to advance understanding of metabolic processes and small-molecule signaling significantly. However, as noted above, the spectra's high complexity poses great challenges for data interpretation aiming to identify specific spectral features relevant within a given biological context. Therefore, continued development of computational tools that aid or partially automate the interpretation process will be required in order to take full advantage of NMR-spectroscopic mixture analysis. Differential comparison of 2D spectra, for example via DANS, constitutes a simple method for selecting signals in complex spectra that warrant more detailed analysis. Computationally more ambitious approaches include methods for the deconvolution of 2D NMR spectra of mixtures, for example DemixC [28,29,30•,31•,32,33], as well as statistical methods for the analysis mixtures using arrays of 1D NMR spectra, for example STOCSY [34–41,42•,43,44,45•].

Automated deconvolution of 2D NMR spectra using DemixC

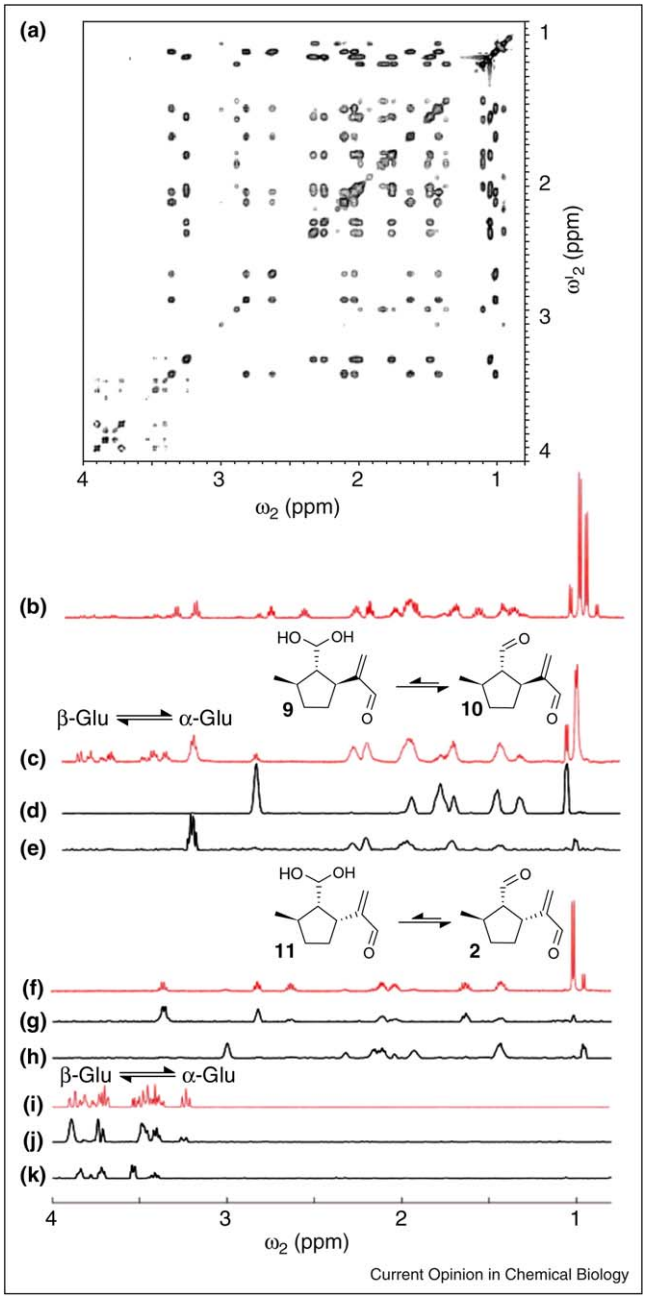
NMR spectra of mixtures, regardless of whether they are one-dimensional, two-dimensional or three-dimensional,

represent superpositions of spectra of the individual components. Due to the mostly linear response of NMR spectroscopy, spectra of mixtures can be thought of as linear combinations of the component spectra, whereby each component contributes proportionally to its concentration in the sample. Therefore, in principle, it should be possible to deconvolute 1D and 2D spectra of mixtures into individual spectra of their components. It should be noted, however, that in order to accomplish effective deconvolution of NMR spectra, some form of additional information or constraints are required in order to determine which groups of signals belong to the same molecule or form a specific structural fragment. Such additional constraints can be derived, for example, from statistical correlation of signals within a spectrum or within an array of spectra.

Brüschweiler *et al.* developed a method named DemixC for deconvoluting (^{1}H,^{1}H)-TOCSY spectra of mixtures into one-dimensional traces (or subspectra) that represent the individual proton-spin systems of the mixture's components [28,30•,32,33]. The one-dimensional subspectra produced by DemixC closely resemble normal ^{1}H NMR spectra, and thus enable identification of the components by screening the subspectra against rapidly growing NMR-spectral databases such as the Biological Magnetic Resonance Data Bank (BMRB, http://www.bmrb.wisc.edu, data for >1000 compounds [46]), the Human Metabolome Database (HMDB, http://www.hmdb.ca, data for ~8000 human metabolites [47]), or the Swedish NMR metabolomics database of Lingkoping (http://www.liu.se/hu/mdl/main).

The functionality of DemixC was first demonstrated for several model mixtures. In a first study the covariance TOCSY spectrum of mixtures of three or four amino acids were deconvoluted via DemixC, which extracted one-dimensional subspectra closely resembling the ^{1}H NMR spectra of the amino acid constituents of the mixture, in a sense providing *in silico* separation of the amino acid mixture [31•]. In a second example, a TOCSY of a model metabolic mixture consisting of five components, sorbitol, histidine, lysine, serotonin and glucose was analyzed via DemixC, which extracted seven one-dimensional subspectra from the TOCSY [30•]. Three of the subspectra, representing sorbitol, histidine, and lysine directly corresponded to data base ^{1}H NMR spectra available for these compounds, whereas serotonin and glucose were each represented by *two* subspectra. In case of serotonin, the two subspectra represent the two disjoint spin systems (aromatic and aliphatic) of this molecule, whereas in case of glucose, two separate subspectra representing the interconverting α-anomers and β-anomers were obtained. This example illustrated the considerable utility of DemixC for deconvoluting complex mixtures of metabolites. Brüschweiler and co-workers further demonstrated the method's applicability to mixtures of several interconverting chemical species using an insect example (Figure 4). As described in the preceding section,

Figure 4



Application of DemixC in the analysis of an unfractionated insect secretion [30•]. **(a)** TOCSY spectrum of walking stick (*A. buprestoides*) defensive secretion. **(b)** Top red trace, representing the ^{1}H NMR spectrum of the secretion. **(d, e, g, h, j, k)** Black traces, representing one-dimensional subspectra obtained from DemixC analysis of the TOCSY, corresponding to the six different components of the secretion, including α-D-glucose and β-D-glucose, dolichodial (**2**), peruphasmal (**10**), as well as the diol-derivatives **11** and **9** of **2** and **10**. **(c, f, i)** Bottom three red traces, representing ^{1}H NMR reference spectra of purified components. Each reference spectrum contains two interconverting species, dialdehyde and diol forms of peruphasmal (trace c), dialdehyde and diol forms of dolichodial (trace f), as well as α-D-glucose and β-D-glucose (trace i). Adapted with permission from [30]. Copyright 2007 American Chemical Society.

walking sticks (*A. buprestoides*) produce a defensive secretion representing a mixture of glucose and two dialdehydes, anisomorphal and peruphasmal. These two dialdehydes are in equilibrium with corresponding geminal diol forms, which contribute to a very crowded appearance of the ^{1}H NMR and TOCSY spectra obtained for the secretion. As shown in Figure 4, DemixC deconvolution of the TOCSY extracted six subspectra representing the dialdehydes, the corresponding diol forms, as well as the α-anomers and β-anomers of glucose.

DemixC has also been applied to (^{1}H,^{13}C)-HSQC-TOCSY spectra of human prostate cancer cell line extracts, which demonstrated that the method can be used to reliably identify known compounds from biological extracts [29]. In using (^{1}H,^{13}C)-HSQC-TOCSY spectra, Zhang *et al.* took advantage of the high spectral dispersion along the ^{13}C-dimension in HSQC data for the identification of individual spin systems and compounds. From the (^{1}H,^{13}C)-HSQC-TOCSY spectra, DemixC derived both ^{1}H and ^{13}C subspectra, which could be queried against NMR databases using a dedicated web interface named COLMAR (for complex mixture analysis by NMR, http://spinportal.magnet.fsu.edu/) [28,33].

DemixC/COLMAR offers an impressive suite of tools for analyzing mixtures. The method appears to be particularly suitable for characterization of samples of limited complexity and identification of major components from more complex mixtures. However, it seems likely that refinement of the DemixC method and integration of other computational approaches (for a study using a matrix factorization approach, see [48]) will further extend applicability.

Using statistical correlation in NMR spectra to correlate metabolites and phenotype

For assigning molecular structures, NMR spectroscopists traditionally depend on physical interactions (such as *J*-coupling or nuclear Overhauser effect) between the individual atoms in the compounds of interest. 2D NMR spectra such as (^{1}H,^{1}H)-TOCSY, (^{1}H,^{13}C)-HMBC, (^{1}H,^{13}C)-HSQC, NOESY, and many more are designed to detect such interatomic interactions, and interpretation of these spectra requires understanding of the exact nature of the interaction, for example *J*-coupling or nuclear Overhauser effects. By contrast, statistical total correlation spectroscopy (STOCSY), developed by Nicholson and co-workers, takes advantage of the multi-collinearity of the intensity variables in a set of spectra, for example ^{1}H NMR spectra [45]. The idea behind this method is that changes in the intensity of NMR signals belonging to a specific metabolite correlate with changes in the amount of this metabolite across a series of different samples. In two-dimensional representations of STOCSY-analysis, crosspeaks indicate signals that strongly correlate with each other, for example because they represent protons belonging to the same molecule. It should be noted that in contrast to TOCSY or NOESY-type spectra, two different protons belonging to STOCSY crosspeaks do not have to be *J*-coupled or in spatial proximity. STOCSY crosspeaks may occur because two protons belong to the same molecule, alternatively a crosspeak may indicate that proton signals from two different compounds are correlated (or anti correlated) because they are involved in same metabolic pathway. STOCSY has been used extensively for biomarker discovery, usually based on comparing sets of ^{1}H NMR spectra acquired for crude biological samples obtained from two or more different cohorts of organisms [2•,34–41,42•,43,44,45•].

For example, STOCSY was used to characterize the metabolic changes in rats treated with the renal cortical disrupting agent mercury(II) chloride [44]. ^{1}H NMR spectra were acquired for urine samples from a cohort of rats treated with mercury(II) chloride and a control cohort. Comparison of the spectra from the two rat cohorts clearly revealed increases in the amounts of glucose, lactate and several amino acids in the ^{1}H NMR spectra of mercury(II) chloride treated rats relative to controls. However, STOCSY also detected significant changes in the concentrations of low-abundance metabolites as indicators of mercury toxicity in the treated rat, including *N*-methylnicotinic acid, *N*-methylnicotinamide and medium chain dicarboxylic acids. Small but reproducible changes in minor metabolites were found to be more indicative of toxicity than variation in major metabolites, which tended to vary significantly even in the control set. STOCSY reaches beyond older statistical treatments of NMR spectra such as principal component analysis (PCA) in that it can identify correlated sets of NMR-spectroscopic signals belonging to a specific compound or groups of compounds, even if these represent minor components in highly complex metabolite mixtures [2•,45•].

Summary and outlook

NMR spectra are extremely information rich and thus offer great promise for the characterization of complex metabolite mixtures. The past few years have seen a surge in the number of applications and the scope of NMR-spectrometric mixture analyses, ranging from primarily novelty-oriented natural products screening to studies aimed at elucidating the structures of specific signaling molecules or the identification of biomarkers in metabolomic studies. Looking ahead, one principal challenge consists in finding ways to access the enormous amount of information contained in NMR spectra more effectively.

Approaches exploring statistical correlation such as STOCSY and simple comparative methods such as DANS are driving progress towards a more comprehensive understanding of how metabolomes are influenced

by external stimuli or genetic changes, and how individual metabolites can be connected to pathways and phenotypes. However, better computational methods will be needed to take full advantage of the capabilities of NMR spectroscopy for characterizing biological systems. As a tool to accelerate both molecular discovery-oriented methods and metabolomic studies, DemixC/COLMAR provides a first example for robust deconvolution of 2D spectra. As manual interpretation of NMR-spectroscopic data of complex mixtures is almost inevitably time intensive and often cumbersome, computational platforms aiding spectral analysis will become increasingly important. For example, comparative analyses such as DANS could be greatly accelerated if an integrated software platform could overlay spectra, reliably pick out signals differential between spectra (and thus characteristic for a specific genotype or phenotype), assess the statistical significance of spectral differences, determine whether the detected differential signals represent known metabolites, and highlight spectral features that cannot be associated with known metabolites. DemixC, though specifically constructed to evaluated TOCSY-type spectra, and STOCSY as well as other statistical approaches not discussed here (e.g. PCA and related methods, see [2•]) represent important steps in this direction.

Acknowledgements

This work was supported by the National Institutes of Health (Grant No. GM079571, GM088290, and GM008500) and DuPont Crop Protection. We thank Arthur S. Edison and Stephen T. Deyrup for helpful comments.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Edison AS, Schroeder FC, Lew M, Hung-Wen L: **NMR – small molecules and analysis of complex mixtures**. *In Comprehensive Natural Products II*. Edited by Elsevier; 2010: 169–196.

2. • Nicholson JK, Lindon JC: **Systems biology: metabonomics**. *Nature* 2008, **455**:1054-1056.
This question and answer-style article broadly frames the field of metabonomics/metabolomics and offers an appropriate starting point for readers interested in the field. The brief article walks the reader through the history, methodology, utility, and future potential of metabonomics as it applies to systems biology.

3. MacBeath G, Saghatelian A: **The promise and challenge of ['] omic' approaches**. *Curr Opin Chem Biol* 2009, **13**:501-502.

4. Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK: **Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts**. *Nat Protoc* 2007, **2**:2692-2703.

5. Bross-Walch N, Kuhn T, Moskau D, Zerbe O: **Strategies and tools for structure determination of natural products using modern methods of NMR spectroscopy**. *Chem Biodivers* 2005, **2**:147-177.

6. Cai SS, Short LC, Syage JA, Potvin M, Curtis JM: **Liquid chromatography-atmospheric pressure photoionization-mass spectrometry analysis of triacylglycerol lipids—effects of mobile phases on sensitivity**. *J Chromatogr A* 2007, **1173**:88-97.

7. Cai SS, Syage JA: **Atmospheric pressure photoionization mass spectrometry for analysis of fatty acid and acylglycerol lipids**. *J Chromatogr A* 2006, **1110**:15-26.

8. Nicholson JK, Connelly J, Lindon JC, Holmes E: **Metabonomics: a platform for studying drug toxicity and gene function**. *Nat Rev Drug Discov* 2002, **1**:153-161.

9. Molinski TF: **NMR of natural products at the 'nanomole-scale'**. *Nat Prod Rep* 2010, **27**:321-329.

10. Schroeder FC, Taggi AE, Gronquist M, Malik RU, Grant JB, 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.

11. • Taggi AE, Meinwald J, Schroeder FC: **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.
This article highlights the discovery of a family of chemically labile sulfated nucleosides found to be present in the crude venoms of the spider *Tegenaria agrestis*. Addressed by this article is the concern that many biologically-derived materials are not compatible with routine purification techniques including chromatography. As a new look into what, at the time, was an already well-studied system, the authors of this article apply 2D NMR-spectroscopic techniques to gain chemical information about crude spider venom samples, subsequently developing chromatographic enrichment strategies compatible with the detected, easily hydrolysable target compounds.

12. Butcher RA, Schroeder FC, Fischbach MA, Straight PD, Kolter R, Walsh CT, 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.

13. Schroeder FC, Gibson DM, Churchill AC, Sojikul P, Wursthorn EJ, Krasnoff SB, Clardy J: **Differential analysis of 2D NMR spectra: new natural products from a pilot-scale fungal extract library**. *Angew Chem Int Ed Engl* 2007, **46**:901-904.

14. • Pungaliya C, Srinivasan J, Fox BW, Malik RU, Ludewig AH, Sternberg PW, Schroeder FC: **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.
DANS was used in this work as the key analytical tool for identifying a synergistic blend of small molecules responsible for hermaphrodite male attraction and the induction of an alternative long lived life-stage ''dauer'' in the model organism *C. elegans*. This work emphasizes the utility of 2D NMR spectroscopy for detecting changes in an organism's metabolome associated with genetic perturbation.

15. Koehn FE, Carter GT: **The evolving role of natural products in drug discovery**. *Nat Rev Drug Discov* 2005, **4**:206-220.

16. Schroder F, Franke S, Francke W, Baumann H, Kaib M, Pasteels JM, Daloze D: **A new family of tricyclic alkaloids from Myrmicaria ants**. *Tetrahedron* 1996, **52**:13539-13546.

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

18. Brey WW, Edison AS, Nast RE, Rocca JR, Saha S, Withers RS: **Design, construction, and validation of a 1-mm triple-resonance high-temperature-superconducting probe for NMR**. *J Magn Reson* 2006, **179**:290-293.

19. Dossey AT, Walse SS, Rocca JR, Edison AS: **Single insect NMR: a new tool to probe chemical biodiversity**. *ACS Chem Biol* 2006, **1**:511-514.

20. Dossey AT, Walse SS, Conle OV, Edison AS: **Parectadial, a monoterpenoid from the defensive spray of Parectatosoma mocquerysi**. *J Nat Prod* 2007, **70**:1335-1338.

21. Lin MF, Shapiro MJ, Wareing JR: **Screening mixtures by affinity NMR**. *J Org Chem* 1997, **62**:8930-8931.

22. Cain CD, Schroeder FC, Shankel SW, Mitchnick M, Schmertzler M, Bricker NS: **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.

23. Zerikly M, Challis GL: **Strategies for the discovery of new natural products by genome mining**. *Chembiochem* 2009, **10**:625-633.

24. Atherton HJ, Jones OA, Malik S, Miska EA, Griffin JL: **A comparative metabolomic study of NHR-49 in Caenorhabditis elegans and PPAR-alpha in the mouse**. *FEBS Lett* 2008, **582**:1661-1666.

25. Shyur LF, Yang NS: **Metabolomics for phytomedicine research and drug development**. *Curr Opin Chem Biol* 2008, **12**:66-71.

26. Srinivasan J, Kaplan F, Ajredini R, Zachariah C, Alborn HT, Teal PE, Malik RU, Edison AS, Sternberg PW, Schroeder FC: **A blend of small molecules regulates both mating and development in Caenorhabditis elegans**. *Nature* 2008, **454**:1115-1118.

27. Butcher RA, Fujita M, Schroeder FC, Clardy J: **Small-molecule pheromones that control dauer development in Caenorhabditis elegans**. *Nat Chem Biol* 2007, **3**:420-422.

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

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

30. • Zhang F, Dossey AT, Zachariah C, Edison AS, Brüschweiler R: **Strategy for automated analysis of dynamic metabolic mixtures by NMR. Application to an insect venom**. *Anal Chem* 2007, **79**:7748-7752.

The authors of this paper showcase DemixC, a computational method for deconvoluting TOCSY data into individual 1D-like spectra corresponding to individual components of a mixture. The article focuses on DemixC analysis applied to both model mixtures and insect-derived natural product mixtures.

31. • Zhang F, Brüschweiler R: **Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy**. *Angew Chem Int Ed Engl* 2007, **46**:2639-2642.

This reference details the analytical approach applied in [30]. The authors explain the concept of transforming TOCSY data, acquired for mixtures of small molecules, into individual 1D-like representations of individual molecular components of a mixture.

32. Zhang F, Brüschweiler R: **Spectral deconvolution of chemical mixtures by covariance NMR**. *Chemphyschem* 2004, **5**:794-796.

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

34. Maher AD, Cloarec O, Patki P, Craggs M, Holmes E, Lindon JC, Nicholson JK: **Dynamic biochemical information recovery in spontaneous human seminal fluid reactions via 1H NMR kinetic statistical total correlation spectroscopy**. *Anal Chem* 2009, **81**:288-295.

35. Maher AD, Crockford D, Toft H, Malmodin D, Faber JH, McCarthy MI, Barrett A, Allen M, Walker M, Holmes E *et al.*: **Optimization of human plasma 1H NMR spectroscopic data processing for high-throughput metabolic phenotyping studies and detection of insulin resistance related to type 2 diabetes**. *Anal Chem* 2008, **80**:7354-7362.

36. Keun HC, Athersuch TJ, Beckonert O, Wang Y, Saric J, Shockcor JP, Lindon JC, Wilson ID, Holmes E, Nicholson JK: **Heteronuclear 19F-1H statistical total correlation spectroscopy as a tool in drug metabolism: study of flucloxacillin biotransformation**. *Anal Chem* 2008, **80**:1073-1079.

37. Wang Y, Cloarec O, Tang H, Lindon JC, Holmes E, Kochhar S, Nicholson JK: **Magic angle spinning NMR and 1H-31P heteronuclear statistical total correlation spectroscopy of intact human gut biopsies**. *Anal Chem* 2008, **80**:1058-1066.

38. Smith LM, Maher AD, Cloarec O, Rantalainen M, Tang H, Elliott P, Stamler J, Lindon JC, Holmes E, Nicholson JK: **Statistical correlation and projection methods for improved information recovery from diffusion-edited NMR spectra of biological samples**. *Anal Chem* 2007, **79**:5682-5689.

39. Holmes E, Loo RL, Cloarec O, Coen M, Tang H, Maibaum E, Bruce S, Chan Q, Elliott P, Stamler J *et al.*: **Detection of urinary drug metabolite (xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy**. *Anal Chem* 2007, **79**:2629-2640.

40. Bohus E, Racz A, Noszal B, Coen M, Beckonert O, Keun HC, Ebbels TM, Cantor GH, Wijsman JA, Holmes E *et al.*: **Metabonomic investigations into the global biochemical sequelae of exposure to the pancreatic toxin 1-cyano-2-hydroxy-3-butene in the rat**. *Magn Reson Chem* 2009, **47(Suppl. 1)**:S26-S35.

41. Sands CJ, Coen M, Maher AD, Ebbels TM, Holmes E, Lindon JC, Nicholson JK: **Statistical total correlation spectroscopy editing of 1H NMR spectra of biofluids: application to drug metabolite profile identification and enhanced information recovery**. *Anal Chem* 2009, **81**:6458-6466.

42. • Alves AC, Rantalainen M, Holmes E, Nicholson JK, Ebbels TM: **Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets**. *Anal Chem* 2009, **81**:2075-2084.

The focus of this work is to establish a set of relevant statistical parameters that researchers can use when implementing STOCSY in chemical structure assignment. For studies of metabolic profiling, STOCSY is a statistically rigorous method that extracts signals from sets of 1D-NMR spectra representing individual unique biomarkers.

43. Johnson CH, Athersuch TJ, Wilson ID, Iddon L, Meng X, Stachulski AV, Lindon JC, Nicholson JK: **Kinetic and J-resolved statistical total correlation NMR spectroscopy approaches to structural information recovery in complex reacting mixtures: application to acyl glucuronide intramolecular transacylation reactions**. *Anal Chem* 2008, **80**:4886-4895.

44. Holmes E, Cloarec O, Nicholson JK: **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.

45. • Cloarec O, Dumas ME, Craig A, Barton RH, Trygg J, Hudson J, Blancher C, Gauguier D, Lindon JC, 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**:1282-1289.

In this article, STOCSY was applied in tandem with orthogonal projection on latent structure-discriminant analysis (OPLS-DA) to identify biomarkers of insulin resistance in mice models. This article provides a very accessible explanation of STOCSY. The authors demonstrate that STOCSY provides not only TOCSY-like connectivities between atoms of individual molecules but also correlations between signals of different molecules, for example sets of chemical species involved in the same pathway.

46. Seavey BR, Farr EA, Westler WM, Markley JL: **A relational database for sequence-specific protein NMR data**. *J Biomol NMR* 1991, **1**:217-236.

47. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S *et al.*: **HMDB: the human metabolome database**. *Nucleic Acids Res* 2007, **35**:D521-526.

48. Snyder DA, Zhang F, Robinette SL, Bruschweiler-Li L, Bruschweiler R: **Non-negative matrix factorization of two-dimensional NMR spectra: application to complex mixture analysis**. *Journal of Chemical Physics* 2008, **128**:.