Monatsh Chem (2011) 142:717–730
DOI 10.1007/s00706-011-0486-6

ORIGINAL PAPER

Structure verification of small molecules using mass spectrometry and NMR spectroscopy

Herbert Thiele · George McLeod · Matthias Niemitz · Till Kühn

Received: 2 December 2010 / Accepted: 21 March 2011 / Published online: 20 April 2011

Abstract The new complete molecular confidence (CMC) concept explores the synergies of the analytical techniques LC–MS and NMR to obtain an estimation of the purity, concentration, and identity of chemical compounds. The high mass accuracy of the MS and MS/MS data provided by the new generation of ESI-TOF and ESI-Q-TOF mass spectrometers provides an accurate determination of molecular weight, which is used specifically for the structural verification and purity determination of substances. The high separation of the isotope profile for both MS and MS/MS spectra affords further dimensions of information to achieve precise molecular formula determination. By performing a complete NMR spectral analysis, the automated consistency analysis routine provides a safe assessment of the consistency between molecular structure and ^{1}H NMR spectrum. The routine returns the fully assigned spectrum and the accurate NMR parameters extracted from the experimental data. Absolute quantification of a series of samples can be automatically performed including the whole workflow from sample setup, automatic NMR measurements, analysis, and spread-sheet reporting. This allows determining mass contents, relative amounts of substances, and purity. The strategy is explored on a set of 96 different pyrrole derivates.

Keywords Complete molecular confidence · Molecular formula determination · Accurate mass · Isotopic profile · Structure consistency analysis

H. Thiele (✉)
Bruker Daltonik GmbH, 28359 Bremen, Germany
e-mail: ht@bdal.de

G. McLeod
Bruker Bruker Daltonics Ltd., Coventry, UK

M. Niemitz
Perch Solutions Ltd., 70500 Kuopio, Finland

T. Kühn
Bruker BioSpin AG, 8117 Faellanden, Switzerland

Introduction

In the service laboratories in the pharmacy industry today, chemists synthesize huge numbers of different compounds on the basis of parallel synthesis and combinatorial chemistry, which must be verified. One main aspect is the confirmation of chemical identity and information about molecular formulae to identify possibly present impurities. Most often, a screening approach is applied in order to obtain a estimation on the purity, concentration, and identity of the synthesized products. This points out samples below certain purity or with too many impurities, and samples that do not have the intended contents. It is often impossible to use all samples at the same time; therefore, most of the samples are stored in ensembles for later use. Undertaking quality control on a regular basis is mandatory in order to check for possible degradation reactions.

The main area of application for quality control and confirmation of molecular identity is synthetic chemistry (medicinal chemistry, core facility, organic chemistry, and pharmacy NCE). Another broad field of application is the identification of small molecules in life sciences, such as metabolite identity and food and drug contaminations.

All of these analyses are handled using different spectroscopic techniques like nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC–MS), and evaluated by manual inspection of all the

spectra types generated. This activity demands an expensive reservoir of human experts with vast knowledge in spectral interpretation, a talent that is becoming very rare nowadays. It also is very cost-intensive, which is why the quality assurance of larger ensembles was often neglected in the past, and only rudimentary analysis was performed on selected samples. An automated approach to structure verification based on the integration of the different information-rich spectroscopic methods LC–MS and NMR should be the method of choice [1–8].

Results and discussions

Methodology

From mass spectrum to molecular formula

With high-resolution instruments, molecular formulae can be calculated directly from the mass spectrum [9]. Because mass spectra do not automatically convey elemental information, data analysis tools are necessary to extract the information inherent in mass spectra to provide molecular formula candidates. Software programs typically produce a list of potential candidates near the measured mass, calculate the expected exact isotope masses and isotopic intensity distribution, and compare them to the measured values. Mass accuracy is an essential parameter to limit the number of potential candidates.

Due to the high number of possible combinations, additional constraints to formula generation are used to restrict the number of solutions. Basic chemical knowledge can supply boundary constraints for formula generation [10, 11]. Some of the constraints can be derived directly from 1D NMR measurements. The integral of non-overlapping signals can give information about the number (min/max) of aliphatic hydrogen atoms and the number of olefinic and/or aromatic double bonds.

The unique ESI-TOF technology

Ideally, an instrument that can provide both high mass accuracy and stable, true isotopic pattern (TIP) information could provide greater information content. This technical concept allows for a two-dimensional analytical method: a combination of accurate mass determination with the analysis of the isotopic distribution. Combining complementary information is essential to find the correct formula for the elemental composition. Time-of-flight instruments are most often the best choice for molecular formula determination. This is especially true for electrospray-Q-TOF-MS instruments (micrOTOF/-Q^{TM}, maXis^{TM}; Bruker Daltonik, Germany), which use linear ion counting to determine the known natural isotopic ratios.

Maximum certainty in small molecule identification requires cutting-edge performance from the MS instrument: a resolution of typically 15,000–20,000 at a high acquisition speed of 20 spectra/s is mandatory to cope with ultra-fast chromatography systems. Mass resolution and mass accuracy have to be maintained in all scan modes and speeds, in MS as well as in MS/MS. The outstanding dynamic range is related to the fast repetition rate of the TOF (5,000–20,000 Hz) and the adequate analog-to-digital-conversion (ADC) technique. This allows exact mass determination over the whole dynamic range as it is not compromised by dead time effects found in the more common time-to-digital-conversion (TDC).

The ADC technique combined with high-resolution TOF-MS is also important for the accuracy of the relative intensities of the isotopic peaks. This accuracy is required for the success of the TIP matching strategy. This can be a severe problem for TDC-based TOF-MS because the intensity of isotopes following high abundant isotopes is often reduced by the dead time of the detector. In MS/MS spectra generated with the micrOTOF-Q II^{TM} (Bruker Daltonik, Germany), the isotopic pattern information and the accuracy are also retained in the fragment ions. Molecular formula proposals can be made for the fragment ions in the same way as for mass spectra, which add a third level of confidence [11].

Isotopic pattern analysis: scoring of formula candidates

Mass accuracy is not enough to reduce the number of possible hits in molecular formula generation [12]. The SmartFormula^{TM} (software delivered by Bruker Daltonik, Germany) approach considers the isotopic pattern distribution for mass spectra. After generation of a list of all possible formulae for a window around a selected mass of an LC–MS peak, the measured isotopic pattern is compared with the theoretical isotopic pattern—resulting in a similarity measure, sigma (σ). This measure is simply the root mean square deviation between the normalized measured and theoretical isotopic intensity distribution. This comparison is done for all the generated molecular formulae. Then, the sigma value is used to rank the formula candidates.

Enhanced probability-based concept

In many cases there are several molecular formulae with well-matched properties. It is not sufficient to consider only one hit with the best scoring factor; there are a few hits satisfying the overall criteria. The overall ranking of the formula candidates has to be extended into a probability-based scoring concept, modeling the distributions of mass accuracy and TIP matching for a number of possible candidates.

An isotopic pattern is described by three characteristic properties: the mass position of the peaks in the pattern, the peak intensities, and the peak distances within the pattern. It is possible to combine those values into a more informative score for the individual hits. This boils down to finding weights modeling the relative informative value of the different properties, which would depend on the accuracies, precisions, and resolutions of masses and intensities of the MS instrument. Such a score can be used to rank the list of hypothetical formulae based on a meaningful quality criterion. Considering all of the generated formula candidates using a Bayesian statistical modeling of the deviations, a score value with a range of 0–100 can be derived.

Precision in formula generation: true isotopic pattern of fragments

Several techniques have been developed, which use the information from MS/MS spectra as an additional criterion for reducing the number of possible formulae for the precursor ion. These methods sum up the potential formulae for product ion and the neutral loss to establish the identity of the precursor ion.

In contrast to these algorithms, a new approach—SmartFormula3D^{TM} (software delivered by Bruker Daltonik, Germany)—utilizes accurate measured mass and additionally accurate measured isotopic pattern. It generates a confident list of formulae simultaneously for the precursor ion and all fragment ions. In contrast to the already known algorithms, both candidate lists are already drastically pruned, thus reducing the time to evaluate all possible relationships between the potential precursor formulae and the related product ions.

Results: molecular formula determination

LC–MS results

For 1 of the 96 samples (pyrrole derivates), an example of an LC–MS chromatogram is shown in Fig. 1. For sample ID183 (a derivate of a benzyl 1*H*-pyrrole-2-carboxylate), a single component accounted for approximately 90% of the detected material. The mass spectrum in Fig. 2 shows the mass measurement of the component of interest. Table 1 shows the MS information derived from the spectrum. For each ion—the pseudomolecular ion, dimer, and adducts—all of the possible formulae within 1-mDa error and acceptable isotopic pattern fit score (σ value) were returned in table format. In this case, based on the chemical rules, mass measurement, and isotopic pattern filtering described, only one possibility was returned for each ion. The theoretical mass of the proposed ions is shown, along with the measurement errors (in ppm and mDa) and the isotopic fit value ($m\sigma$). A score value is also returned, based on combined isotopic fit and mass, normalized to a sum of 100. In this example, there is only one possibility for each ion, so the score is simply 100; if multiple formulae were proposed, then lower scoring possibilities would be considered less likely. The formula of the $[M+H]^+$ ion is thus definitively assigned as $C_{20}H_{22}NO_5$, and therefore the molecular formula is $C_{20}H_{21}NO_5$.

LC–MS/MS results

Further structural information on sample ID183 was derived from the MS/MS results. The accurate mass spectrum of the compound of interest was obtained from the auto-MS/MS experiment. The isotopic pattern information can also be obtained for the fragment ions by setting an appropriate isolation window. Figure 3 shows the MS/MS spectrum of the precursor *m/z* 356 with accurate mass measurements. Table 2 shows the assignment of the masses of the ions of interest, again using a 1-mDa window and isotopic pattern fit threshold.

All of the product ions in the spectrum were assigned to individual formulae based on the same mass measurement and isotopic pattern thresholds, again using the SmartFormula^{TM} tool. For compounds where there is ambiguity about the empirical formula, that is, where there is more than one reasonable possibility proposed for the precursor ion by SmartFormula^{TM}, the SmartFormula3D^{TM} tool was employed, in which those precursor and product ion formulae that could not logically combine to form one another would be filtered out.

For this sample, the LC–MS data gave a definitive identity of $C_{20}H_{21}NO_5$ that can inform the interpretation of NMR data and furthermore provides MS/MS fragmentation information with formulae that can be directly related to a structure proposed through NMR.

Quantitation and structural confirmation using NMR

NMR delivers the most complete information on molecular structures on an atomic scale in solution. Typically, NMR observes ^{1}H chemical shift and ^{1}H–^{1}H couplings, both allowing conclusions on the chemical environment of the observed proton. Even though, besides ^{1}H nuclei, many more isotopes of the periodic table are accessible by NMR, ^{1}H is the most sensitive one and is the most frequently used nucleus observed by NMR.

In addition to being very selective and specific about chemical structures, NMR is a fully quantitative method. The intensity of an NMR signal of any given nucleus is linearly proportional to the total amount of these nuclei in the observed volume (Eq. 1):

$$\frac{Signal}{Noise} = n\gamma_{EXE}\gamma_{DET}^{3/2}B_0^{3/2}\frac{\sqrt{NS}}{T}T_2 \quad (1)$$

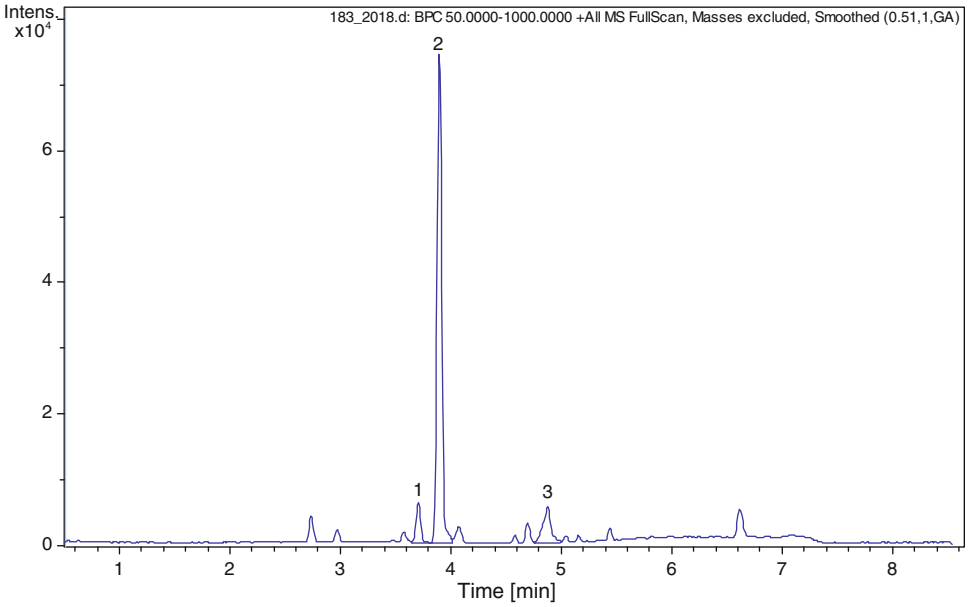

Fig. 1 Integrated base peak chromatogram for LC–MS analysis of derivate(s) of benzyl 1*H*-pyrrole-2-carboxylate (sample ID183)

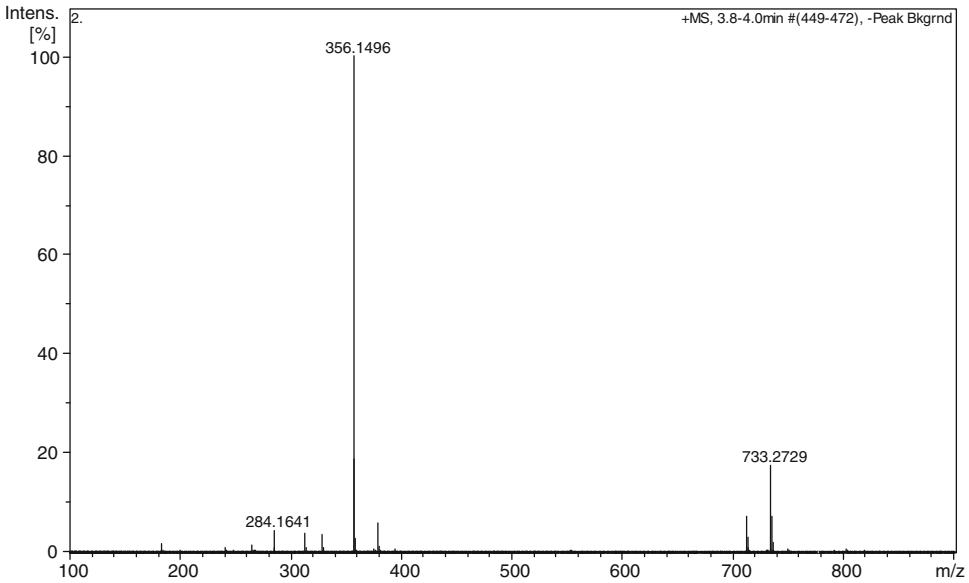

Fig. 2 Full scan TOF-MS spectrum, well H8, LC peak 2 (component >90% of signal). One major $[M+H]^+$ ion was observed; the related sodiated ion, dimer, and sodiated dimer were also seen and assigned

Table 1 Formula assignments for the UHR-TOF mass spectrum of sample ID183

Measured *m/z*	Formula	Score	Theoretical *m/z*	Err (mDa)	Err (ppm)	*mσ*
356.1496	$C_{20}H_{22}NO_5$	100	356.1492	−0.3	−0.9	19.1
378.1309	$C_{20}H_{21}NNaO_5$	100	378.1312	0.3	0.9	10.8
711.2906	$C_{40}H_{43}N_2O_{10}$	100	711.2912	0.6	0.9	23.1
733.2729	$C_{40}H_{42}N_2NaO_{10}$	100	733.2732	0.2	0.3	21.1

where n is the number of nuclei in the active volume, γ is the gyromagnetic ratio of the excited and detected nucleus, B_0 is the magnetic field of the instrument, NS is the number of scans, and T is the temperature.

This means that the NMR signal is not dependent on any molecular properties other than the number of nuclei (e.g., ^{1}H atoms) in the molecule. This is very different from optical methods, for example, where the extinction coefficient is a unique molecular property that can be very different from molecule to molecule.

NMR therefore is often used as a "standard-free" method for quantification. It is used in both applications,

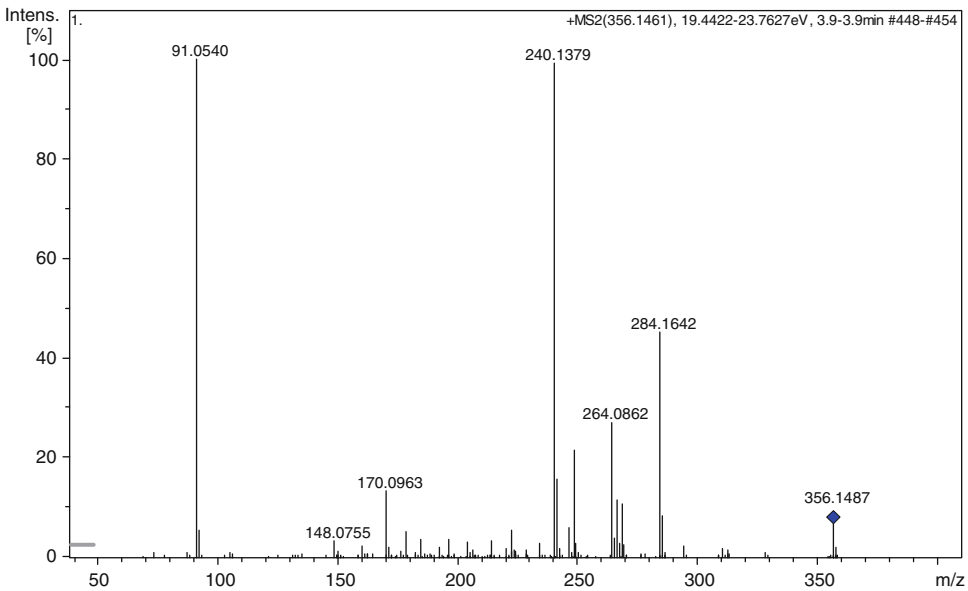

Fig. 3 UHR-TOF-MS/MS spectrum of *m/z* 356

Table 2 Formula assignments for the UHR-TOF-MS/MS spectrum of sample ID183

Measured *m/z*	Formula	Score	Theoretical *m/z*	Err (mDa)	Err (ppm)	*mσ*
91.0540	C_7H_7	100	91.0542	0.3	2.9	12.9
170.0963	$C_{12}H_{12}N$	100	170.0964	0.1	0.5	8.3
240.1379	$C_{16}H_{18}NO$	100	240.1383	0.3	1.4	10.8
248.0919	$C_{13}H_{14}NO_4$	100	248.0917	−0.1	−0.6	16.3
264.0862	$C_{13}H_{14}NO_5$	100	264.0866	0.5	1.8	11.9
284.1642	$C_{18}H_{22}NO_2$	100	284.1645	0.3	1	10.8

either for absolute quantification to determine concentrations [without the need for certified reference compounds and with higher accuracy than optical methods like UV, ELSD (evaporative light scattering), or on a larger concentration range like CLND (chemoluminescence nitrogen detection)]. NMR may also be used for relative quantification, e.g., in order to determine the amount of an impurity with respect to a main compound. NMR therefore is ideally suited for structural confirmation and for quantification of compounds in all areas of chemistry or pharmaceutical research, and is routinely used here.

An example for pharmacy research illustrates the importance of a combined MS and NMR approach for quality assurance aspects: the accurate characterization of the contents of lead discovery substance libraries used in pharmaceutical research and development is of vital importance for drug discovery. These are typically collections of solutions of small molecules used for bioactivity screening. Three aspects define the quality and reliability of a company's liquid screening repository:

Purity: Is the substance in the depository really pure or does the sample contain impurities or a mixture?

Verification: Is the molecular structure of the submitted compound known and correct?

Concentration: Knowledge of substance concentration is of utmost importance for affinity assays.

To address all of these aspects of quality control, all substances that enter the depository must undergo reliable quality control (QC) or quality assurance (QA) procedures. Ideally, this includes liquid chromatography coupled with mass spectrometry (LC–MS). The LC part gives a rough estimation of the sample purity, where the MS detection allows for the conformation of the molecular formula.

In addition to this, NMR gives an answer about the structural integrity as well as the absolute concentration of the sample. However, up until recently, NMR has been regarded as too insensitive for routine QC. NMR was considered to consume too much of the precious materials that are often available only in very small amounts, e.g., from parallel syntheses. Further drawbacks were the need for significant amounts of expensive deuterated solvents and the slow manual analysis of the information-rich spectra.

With a new generation of high-sensitivity probes such as the CryoProbe family or the 1 or 1.7 mm MicroProbes (Bruker BioSpin, Germany), the amount of sample required for NMR measurement can be reduced by an order of magnitude. It is now possible to measure and quantitate the ^{1}H NMR spectra of ca. 100-µg samples of small synthetic molecules in about 3 min.

The amount of deuterated solvent can also be reduced dramatically or even omitted completely through the use of small-volume microprobes with sample volumes as low as 5 mm^3. With such small volumes the quality of solvent suppression is excellent, so that it is now possible to even work without or with only very small percentages of deuterated solvents. This capability simplifies the typical workflow in a chemistry environment; relatively small aliquots of the synthesized products, already dissolved in DMSO, for example, can be diverted for direct NMR analysis.

New automation hardware enables fully automated sample handling, including sample preparation, tube filling, and sample changing. By using an adapted Gilson Liquid Handler 215 (Bruker BioSpin, Germany), it is now possible to automatically fill NMR tubes that are evenly arranged in 96-well plate racks. These racks can then be transferred to the SampleJet autosampler (Bruker BioSpin, Germany) directly. The SampleJet allows the storage of 5 of these 96 tube racks, which enables users to run NMR in a hands-off fully automated fashion with these types of samples for over the weekend. In addition to the rack storage, the autosampler is ideally suited for open access NMR with single samples. Therefore, today traditional routine NMR laboratories equipped with this hardware can handle high-throughput repository QA tasks in addition to their normal walk-up routine work load.

The bottleneck for automated quality assurance by NMR is now shifted away from hardware automation, which is solved by the products described above. Today the bottleneck is rather the analysis of the data, and solutions are provided to answer to these demands.

A software package that allows the full setup, data acquisition, data processing, spectra interpretation, and result reporting for batch wise compound quality assurance is available (Bruker BioSpin, Germany). The software package is named CMC-q for *c*omplete *m*olecular *c*onfidence for *q*uality assurance and *q*uantification. The workflow is based on a file-in-file-out basis, and the user typically simply supplies a well plate of samples along with an industry standard SD file that describes the contents of the well plate. The SD file holds the information on the chemical structures in the wells along with further information on the samples such as compound names, laboratory journal numbers, etc. The CMC-q package interprets this SD file, generates the NMR orders, and runs the NMR experiments. The data then is automatically analyzed using a human logic emulation algorithm, which checks for inconsistencies between the spectra and the provided structures. In this way it is possible to reliably pick out for example pipetting errors in the well plate or to identify compounds that have been decomposed. The CMC-q analysis algorithm automatically assigns concentrations to the samples using the PULCON method [13, 14]. It also determines the water content on samples acquired in non-deuterated DMSO. The CMC-q software package includes a data viewer, which allows the user to very easily check and possibly re-evaluate the automatically generated results (see also Fig. 5).

The post-processing module of CMC-q automatically analyzes 1D ^{1}H NMR spectra using an algorithm that emulates the human logic for spectra interpretation. It uses information on chemical shift ranges, signal intensities, coupling constants, and line shape, and applies chemical and spectroscopic rules and "common sense" in order to interpret the 1D NMR spectra. It can handle a certain level of spectra imperfections and impurities, and therefore is able to analyze a wide range of data. The following information is extracted from the spectra:

1. CMC-q analysis yields integration regions and automatically calibrates the corresponding integral values to the number of protons for these regions.
2. CMC-q analysis suggests the concentration for the major component in the sample.
3. It provides a coarse multiplet analysis for first order multiplets and identifies possible higher order multiplets.
4. CMC-q analysis can identify known impurities in the spectrum and assigns the solvent signal.
5. CMC-q analysis identifies signals that possibly may originate from other impurities in the sample.
6. If a molecular structure in the form of a mol file is present in the data set, CMC-q analysis performs a plausibility check of whether or not the spectrum could correspond to the given structure.
7. It also suggests an assignment if the NMR spectrum can be explained with the given molecular structure.
8. If the spectrum has been acquired in non-deuterated DMSO (HDMSO), CMC-q analysis estimates the water content in the sample.

For all these automatic analysis tasks, NMR spectra of a certain data quality are required, and this is achieved by using the CMC-q NMR parameter sets for acquisition. In this way, the highest confidence in the automatic analysis can be reached. CMC-q analysis may also be used on individual data sets where its spectra interpretation functionality can be used to prepare spectra for publication or for further analysis.

Results: concentration and structural plausibility

The NMR measurements were set up from the SD file that accompanied the samples using the CMC-q setup tool. The data were acquired in full automation with the CMC-q standard parameter sets with 32 scans. The total experiment time was <5 min per sample, including the sample exchange time.

Figure 4 shows the sample ID186 data set acquired on just a few micrograms of sample with these parameters and the automatic analysis result by the CMC-q routine, indicating the concentration and structural plausibility. The multiplet at 7.3 ppm has been automatically assigned to the five aromatic protons by the algorithm, the integrals have been scaled accordingly, and the concentration has been automatically determined. The CMC-q analysis algorithm has internally assigned the other signals as well and has found no inconsistencies between supplied spectrum and structure. The solvent peak at around 2.5 ppm and the water peak in DMSO, which in this case appears at around 3.5 ppm, have of course been recognized by the CMC-q analysis algorithm automatically. The results for the whole batch of 96 samples as displayed by the CMC-q batch analysis tool are displayed in Fig. 5.

Using the batch analysis tool, the user can easily control and modify the automatically generated results, and finally the results are exported into an EXCEL spreadsheet.

From NMR to 3D structure

NMR spectroscopy also gives information about stereochemistry and even dynamic effects. However, extracting this vast amount of information from the NMR data usually requires profound knowledge and long years of experience. New automated software methods can deliver insight also to the less experienced user and increase the efficiency of the experienced user by extracting the essential information from the overwhelming amount of data in full automation.

Computational NMR spectral analysis

The consistency between a given structure and its ^{1}H NMR spectrum can be assessed by comparing predicted or tabulated NMR chemical shifts and couplings with the ones actually found in the experimental data. However, predicted and tabulated NMR parameters usually do not directly allow unambiguous and complete assignments especially for 1D proton spectra because they do not resemble the experimental spectrum well enough. In addition, the extraction of the actual NMR parameters from the experimental data can be tedious, difficult, or even impossible if signals overlap or show higher order effects. In the latter case only complete quantum mechanical spectral analysis can accurately retrieve the actual NMR parameters [15].

Optimizing predicted or tabulated chemical shifts and *J* couplings to match the experimental data can be achieved using iterative quantum mechanical spectral analysis [16]. However, starting values for the NMR parameters need to be reasonable, solvent signals removed, and the chemical shift assignments in correct order before the iterative process can succeed.

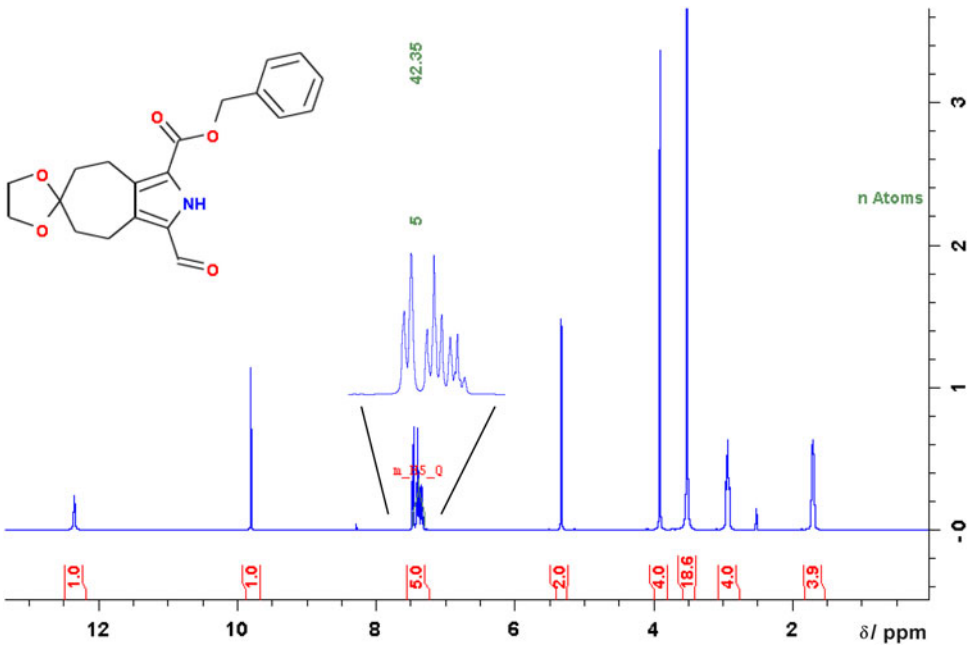

Fig. 4 A derivate of benzyl 1*H*-pyrrole-2-carboxylate (sample ID183): fully automatic NMR data acquisition, processing, and analysis using CMC-q

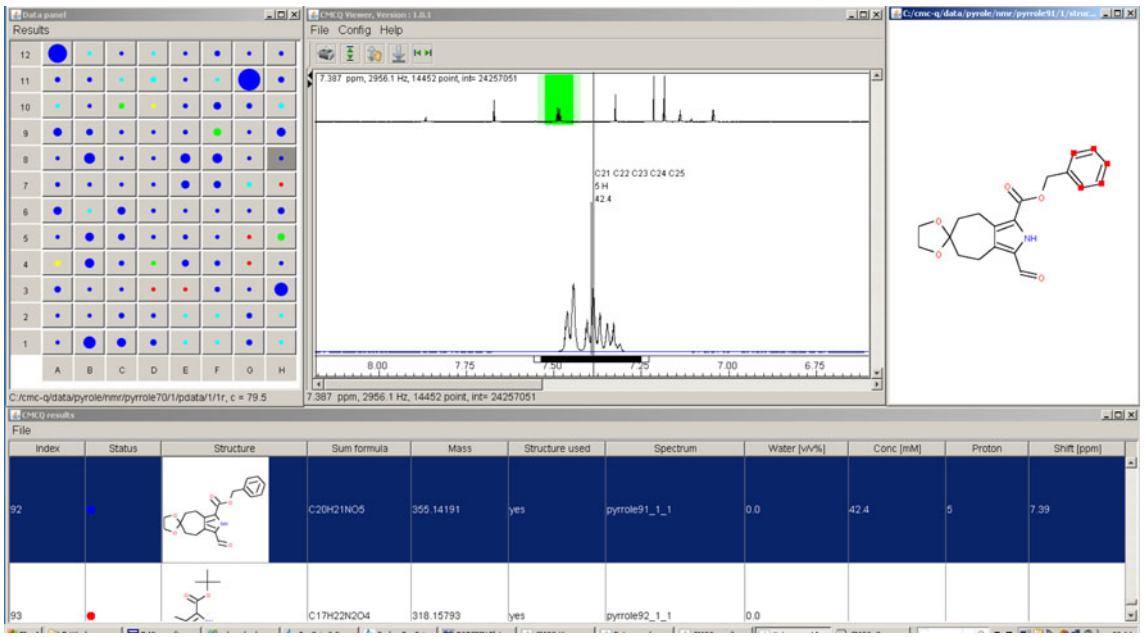

Fig. 5 Batch display of the whole well plate as result control. *Dark blue dots* indicate that the CMC-q algorithm has not found inconsistencies between spectra and structure based on the chemical shift prediction. *Light blue dots* indicate that the automation suggests checking the results for consistency, even though a concentration is determined. *Red*, *yellow*, and *green* indicate that a human analyst has checked the result and has made an expert interpretation on the respective spectrum. The dot size scales to the concentration. The viewer automatically zooms the spectral view to the selected multiplet for quantification. The assigned areas in the molecule light up

Automated consistency analysis (ACA)

A procedure is presented that performs such an iterative quantum mechanical spectral analysis in full automation even from poor starting values and including solvent signals, finally overcoming tedious manual procedures required before. The procedure includes the following steps:

Step 1: spectral post-processing
Accurate peak and intensity information is extracted from the spectrum using automated peak-picking, integration including background removal (wide-moving-minima filter), and calculation of the spectral linewidth, lineshape, and *S/N* (signal to noise ratio). A total-line-shape analysis (TLS) [17] is used for complete de-convolution also detecting and fitting broad lines (for example, water signals).

The peak, intensity, and lineshape information is used later to correlate the spectral information with the NMR prediction when generating reasonable solutions to be tested by the iterative quantum mechanical calculation (Fig. 6).

Step 2: NMR spectrum prediction
Starting from the given structure, molecular mechanics including geometry optimization (GO), Monte-Carlo (MC), and molecular dynamics (MD), is used to establish an ensemble of conformers representing the conformational space of the solute molecules. This also allows the prediction of stereochemical and even dynamic effects, such as the flexibility of the structure, which are essential especially for proton NMR spectroscopy (Fig. 7).

The NMR prediction is based on a complex semi-empirical model using up to 400 parameters to describe proton chemical environments as generically as possible, also including parameters such as solvent, pH, and ion concentration. The NMR spin system is generated by grouping NMR particles taking magnetic and two-fold chemical equivalence symmetry into account. The spectrum is calculated using quantum mechanical principles including *X*-approximation. Reliable error estimates for the chemical shifts and coupling constants are given as well (Fig. 8).

Step 3: spectral assignment
The information retrieved by the first two steps is used to generate all reasonable assignments called solutions correlating the predicted with the experimental NMR spectrum. Unfortunately, the available information is usually neither complete nor accurate, and often even self-contradictory. Therefore, the principle of cost analysis is used to generate all reasonable assignments and rank them based on the distance between predicted and actual chemical shifts weighted by the estimated prediction error and the compatibility with corresponding peak and

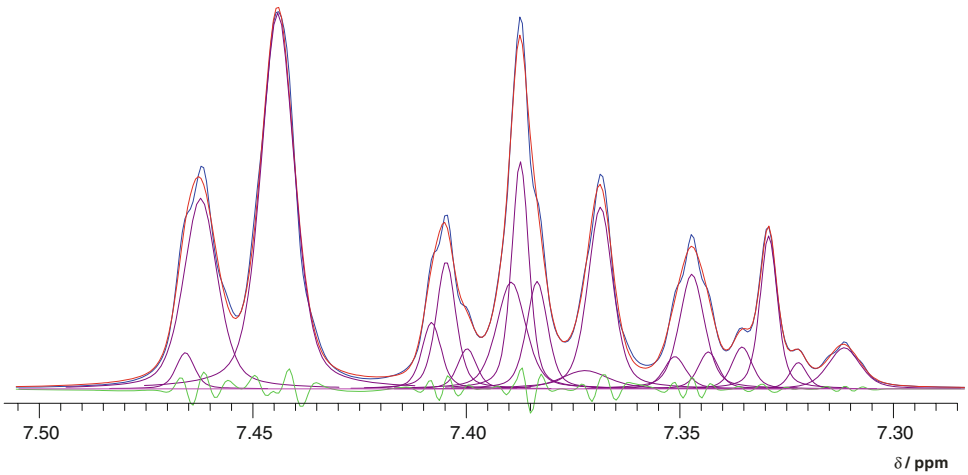

Fig. 6 The result of the total-line-shape analysis for the aromatic region of sample ID183 (a derivate of benzyl 1*H*-pyrrole-2-carboxylate) with the experimental spectrum in *blue*, the individually fitted lines in *violet*, the calculated spectrum in *red*, the difference between experimental and calculated in *green*, and the baseline in *magenta*

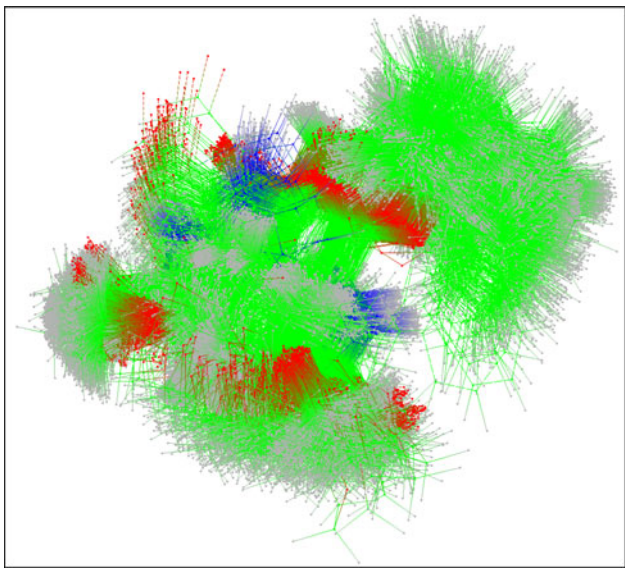

Fig. 7 A set of conformers generated by the MC/MD calculation to achieve an average of chemical environments to be used by the NMR prediction algorithm

intensity information. Each violation of prior information imposes a penalty (cost) upon the tested solution; the graver the violation and the higher the reliability of the prior information are, the higher the cost.

The reasonable solutions are then ranked by their overall cost. The smaller the overall cost of a solution, the more likely is this solution. To reduce the vast set of possible solutions to a more manageable number, the cost analysis employs sophisticated dead end elimination algorithms to weed out solution candidates that are deemed to be not reasonable as early as possible. This is done in two steps, first assigning NMR groups into integration areas and then placing them within each integration area.

The whole spectral assignment is based on the assumption that the spectrum consists of one primary solute and a number of solvents that can have arbitrary molar quotients relative to the solute.

Step 4: iterative spectral analysis

Classical quantum mechanical spectral analysis is used to test the most likely assignments generated by step 3. By iteratively optimizing the pre-assigned NMR parameters to match the experimental data, the difference between the calculated and experimental spectrum is reduced until the two spectra resemble each other as well as possible. This is done in several successive steps, first optimizing chemical shifts and then couplings from large to small. Finally, a complete total-line-shape fitting is used to also optimize the individual linewidths of each spin particle and the overall lineshape (Lorentzian/Gaussian contribution). This extremely accurate fitting method delivers a very high level of specificity because all NMR parameters need to be in full consistency, yielding fully assigned spectra and very accurate spectral parameters extracted from the data even for overlapping signals and strongly coupled spin systems. Solvent peaks are recognized and treated accordingly. Thus, no exclusions ("dark regions") are needed, and peaks overlapping with solvent peaks can be analyzed as well (Fig. 9).

Step 5: final evaluation

A match index combining individual likelihoods of all assignments with the quality of the final fit (RMS) between the calculated and experimental spectrum is given for the best solution and provides an extremely safe assessment of the consistency between the given structure and its NMR data. In case of doubt, sub-scores are available providing information regarding the discrepancy between found and expected spectral parameters as well as the spectral regions where the fitting is poor. This main score ranks from 0 to 100 and is given with a descriptive evaluation also for the

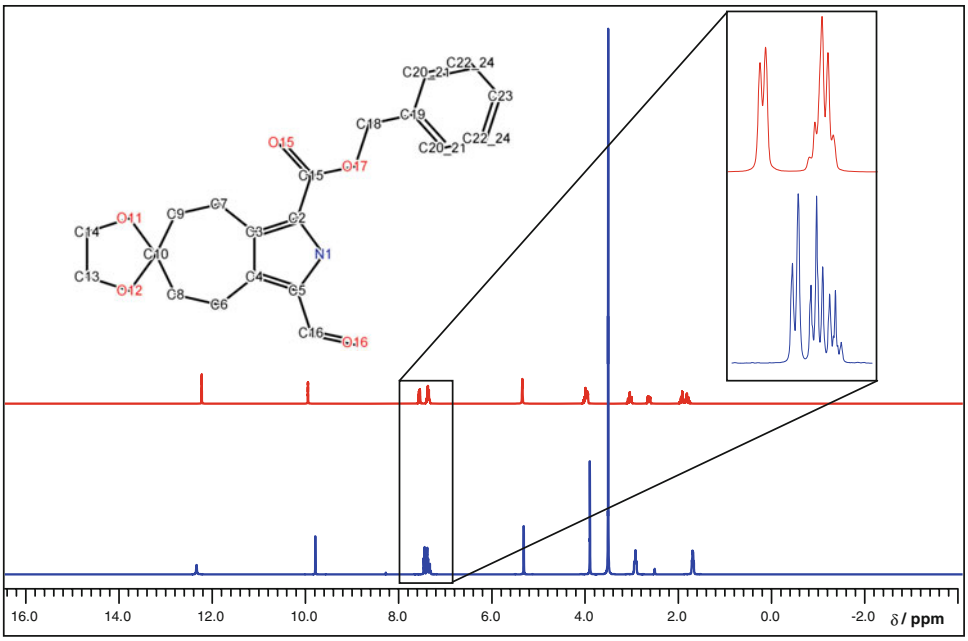

Fig. 8 The predicted (*red*) versus the experimental spectrum (*blue*) for sample ID183. The expansion shows the small but distinct difference between predicted and experimental spectra

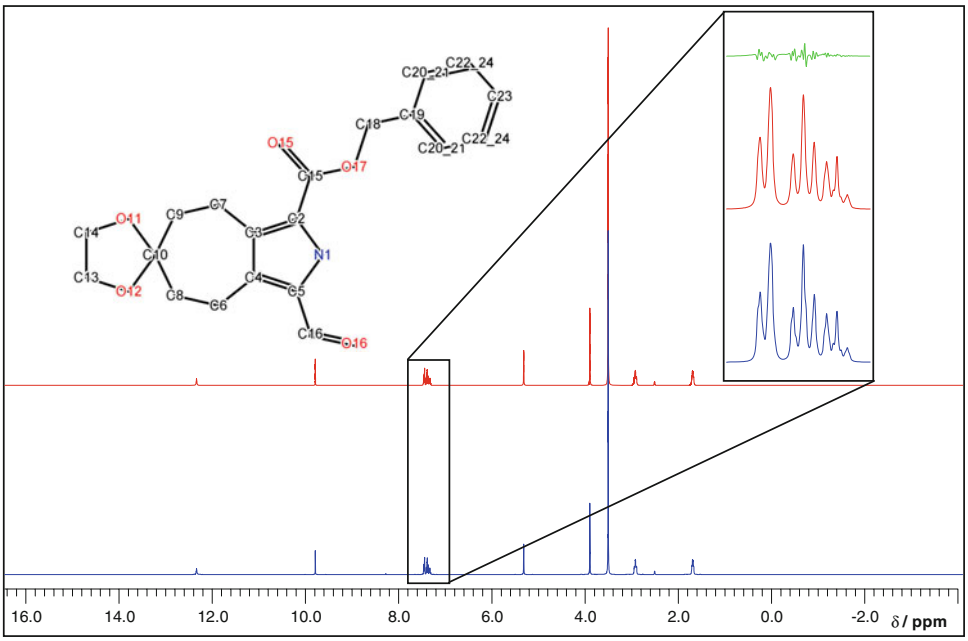

Fig. 9 The fitted (*red*) versus the experimental spectrum (*blue*) for sample ID183. The expansion shows the perfect agreement between fitted and experimental spectra including the difference (*green*)

Table 3 Full list of descriptors

Name	Range (worst–best)	Description and interpretation guidelines
Match index	0.0–100.0	Overall score for the consistency between given structure and spectral data
Total RMS	100.0–0.0%	Overall root mean square between theoretical and experimental spectrum
Shift similarity	0.0–100.0%	Similarity index between predicted and found chemical shifts
Coupling similarity	0.0–100.0%	Similarity index between predicted and found *J* couplings
Highest local RMS	100.0–0.0%	Local root mean square between theoretical and experimental spectrum for the NMR particle with the poorest fit
R-factor	0.0–100.0%	Percentage of total experimental spectrum intensity explained by the theoretical spectrum intensity

sub-scores. The full list of descriptors is presented in Table 3.

Results: automated consistency analysis

The results of all 96 samples were also checked manually. Samples where structure and spectrum were consistent were classified as "true-positive." Samples with no consistency between structure and NMR spectrum were classified as "true-negative." Samples with impurities more than 40% were classified as "highly contaminated." Samples with <40% but more than 20% impurities were classified as "contaminated."

The ACA results were classified as follows: samples with a match index >90 were classified as "positives." Samples that for various reasons did not give a result were classified as "no-solution" (NS). No separate test for "false-positives" was conducted, but at least all of the "true-negatives" and "highly contaminated" samples were correctly identified as "not positive." Table 4 summarizes the results for the 96 samples: all 13 "true-negatives" and 4 "highly contaminated" scored a match index of 0. The "contaminated" samples gave match indices between 0 and 39.1.

In 28 cases the method did not deliver a result ("no solution"); 20 of them were either "true-negative" or "highly contaminated." From the remaining eight cases, one shift prediction was poor. In another case the spin system was not recognized correctly. In the rest of the cases "no solution" was caused by still too complex spectral regions with too many possible assignments for which no reasonable solutions could be found or that could not be fitted properly. Therefore a result "no solution" does not mean a safe "true-negative."

However, the method still distinguishes between spectra that are at least consistent with respect to the intensity and chemical shift information and spectra that fail this test. The latter can be considered as "true-negatives," and from these only two samples were actually "false-negatives." Table 5 shows the NMR parameters (chemical shifts and couplings) extracted from the experimental data for sample 183 compared to the predicted values.

Table 4 The results for the 96 samples

	Number	Percentage
Assigned manually		
True-positives (TP)	76	79.2
True-negative (TN)	13	13.5
Highly contaminated (HC)	4	4.2
Contaminated (C)	3	3.1
Assigned by ACA		
Positive (completely assigned by ACA)	68	70.8 (overall)
		86.1 (excluding TN + HC)
No solution	28	29.2 (overall)
		10.1 (excluding TN + HC)
False-positive	0	0 (see text)
False-negative	2	2.6 (see text)

To test the ability of CMC-i to distinguish very similar but wrong structures from the correct one, also wrong structures were analyzed regarding their consistency with the spectral data (Fig. 10). All false structures failed in the automated consistency analysis and returned match indices of 0.

Complete NMR spectral analysis can be successfully automated yielding fully assigned spectra for more than 86% of the tested pyrrole samples with impurities <20%. Samples with higher content of impurities cannot be analyzed yet because additional signals of considerable intensity cause a combinatory explosion, which requires additional developments to be handled.

The overall success rate was above 70% (above 86% when excluding the samples classified as "true-negatives" and "highly contaminated"). This rather satisfactory result cannot be generalized though. The data set is very small and the chemistry is very similar. Also, most of the structures are fairly small. The extremely high selectivity could be demonstrated by testing several very similar but false structures with the data set as in sample ID183. No "false-positive" was found.

With larger structures and more complex spectra, the overall success rate can be expected to gradually decrease, again because the number of possible assignments is increasing by factorial. Additional information like 2D NMR information (for example, HSQC) is then needed to counteract the combinatory explosion.

Experimental

LC–MS analysis

Sample preparation

The samples were prepared for LC–MS analysis by adding 100 mm^3 acetonitrile to each dried sample well. The resulting sample solutions were diluted 1:200 by adding a 5 mm^3 aliquot to 995 mm^3 of 1:1 (v/v) water:acetonitrile. These samples were used for the LC–MS analysis.

LC–MS method

Mass spectrometry was performed using a maXis^{TM} Ultra-High Resolution Time-of-Flight mass spectrometer (UHR-TOF) with Qq-TOF geometry (Bruker Daltonik, Germany). Data were acquired over mass in the range 50–1,000 Da with an acquisition rate of 2 spectra/s. All data were

Table 5 The NMR-parameter for sample ID183

Shift	Obs	Pred	Range	Delta	Rel. D	Coupling	Obs	Pred	Range	Delta	Rel. D
ACA analysis 1D 1H: shift and CC											
Hl	12.336	12.225	1.946	0.111	0.057	J(H6A, H6B)	−10.034	−14.970	2.560	−4.936	−1.928
H16	9.789	9.945	0.412	−0.156	0.380	J(H6A, H8A)	5.816	1.790	4.400	4.026	0.915
H20_21	7.452	7.553	0.204	−0.101	0.494	J(H6A, H8B)	6.761	7.320	6.800	−0.559	−0.082
H22_24	7.387	7.355	0.208	0.032	0.153	J(H6B. H8A)	7.515	13.220	4.800	−5.705	−1.189
H23	7.333	7.389	0.207	−0.056	0.273	J(H6B, H8B)	1.214	1.670	4.400	−0.456	−0.104
H18	5.316	5.346	0.178	−0.030	0.167	J(H7A.H7B)	11.128	−14.970	2.560	−3.842	−1.501
H14B	3.900	4.015	0.185	−0.115	0.620	J(H7A, H9A)	9.846	1.740	4.400	8.106	1.842
H14A	3.897	3.985	0.162	−0.088	0.544	J(H7A. H9B)	13.573	13.800	4.000	−0.227	−0.057
H13A	3.896	3.943	0.150	−0.047	0.311	J(H7B, H9A)	5.628	6.660	5.600	−1.032	−0.184
H13B	3.894	3.959	0.205	−0.065	0.317	J(H7B, H9B)	3.275	1.870	4.400	1.405	0.319
H6B	2.942	3.039	0.283	−0.097	0.343	J(H8A, H8B)	−10.243	−12.650	1.500	−2.407	−1.604
H7A	2.924	3.039	0.297	−0.115	0.386	J(H8B, H9A)	3.650	2.380	1.200	1.270	1.059
H6A	2.912	2.629	0.185	0.283	1.527	J(H9A, H9B)	−13.404	−12.640	1.500	0.764	0.509
H7B	2.906	2.629	0.200	0.277	1.386	J(H13A, H13B)	−9.067	−7.890	2.800	1.177	0.420
H8A	1.705	1.933	0.332	−0.228	0.687	J(H13A, H14A)	0.195	0.880	3.800	−0.685	−0.180
H9B	1.690	1.787	0.231	−0.097	0.418	J(H13A, H14B)	8.390	5.070	6.000	3.320	0.553
H9A	1.687	1.834	0.535	−0.147	0.275	J(H13B, H14A)	7.475	5.070	6.000	2.405	0.401
HBB	1.680	1.908	0.521	−0.228	0.437	J(H13B, H14B)	10.281	9.740	6.000	0.541	0.090
						(J1H14A, H14B)	−8.930	−7.890	2.800	1.040	0.372
						J(H20_21, H20_21)	0.003	1.980	1.200	−1.977	−1.648
H2O	3.505	3.504	0.150	0.001	0.005	J(H22_24, H20_21)	0.373	0.550	0.740	−0.177	−0.239
DMSO	2.507	2.507	0.150	0.000	0.001	J(H20_21, H22_24)	7.623	7.660	0.340	0.037	0.110
						J(H20_21, H23)	1.142	1.260	1.200	−0.118	−0.098
						JIH22_24, H22_24)	0.019	1.480	1.200	−1.461	−1.218
						J(H22_24, H23)	7.471	7.400	0.240	0.071	0.294
						J(DMSO, DMSOD1)	1.307	1.700	0.150	−0.393	−2.620
						J(IDM5O, DM5OD2)	1.307	1.700	0.150	−0.393	−2.620

acquired using an automatic external calibration routine. MS/MS spectra were acquired in auto-MS/MS mode (data-dependent acquisition). Collision energy was estimated dynamically based on appropriate values for the mass and stepped across a ±10% magnitude range to ensure good quality fragmentation spectra.

Samples were introduced as 5 µl injections on a Ultimate3000RS UHPLC system (Dionex) with a Phenomemex Luna-HST C18(2) 2.5 µm 5 cm × 2.0 mm HPLC column. A gradient elution using solvent A: water + 0.2% formic (v/v) acid; solvent B: acetonitrile + 0.2% formic acid (v/v) was applied as follows: start 20% B for 0.2 min, ramp to 100% B for 3.5 min, hold at 100% B for 2 min, return to starting conditions in 0.1 min, and re-equilibrate for 2.5 min. The flow rate was 0.3 cm^3 min^{-1}, and the column was maintained at 40 °C. All solvents were LC–MS Chromasolv grade (Sigma-Aldrich Ltd., Poole, UK).

The resulting LC–MS data were processed in an automated software routine within the Compass DataAnalysis software (Bruker Daltonik, Germany) as follows: a base peak chromatogram was produced and integrated. Then the spectrum of each detected peak (relative threshold 5% peak area) was extracted. The formulae of all significant detected ions were automatically assigned using the SmartFormula™ tool. Based on the expected chemistry, elements carbon, hydrogen, oxygen, nitrogen, bromine, and iodine were permitted. Sodium was also included for calculation of adduct masses. The number of nitrogen atoms was limited to an upper threshold of ten. The number of rings plus double bonds was checked to be chemically meaningful (between 0 and 50). The nitrogen rule was enforced. Even-electron species only were permitted. A window of 1-mDa mass accuracy was allowed. A threshold was applied to the goodness of fit of the isotope pattern,

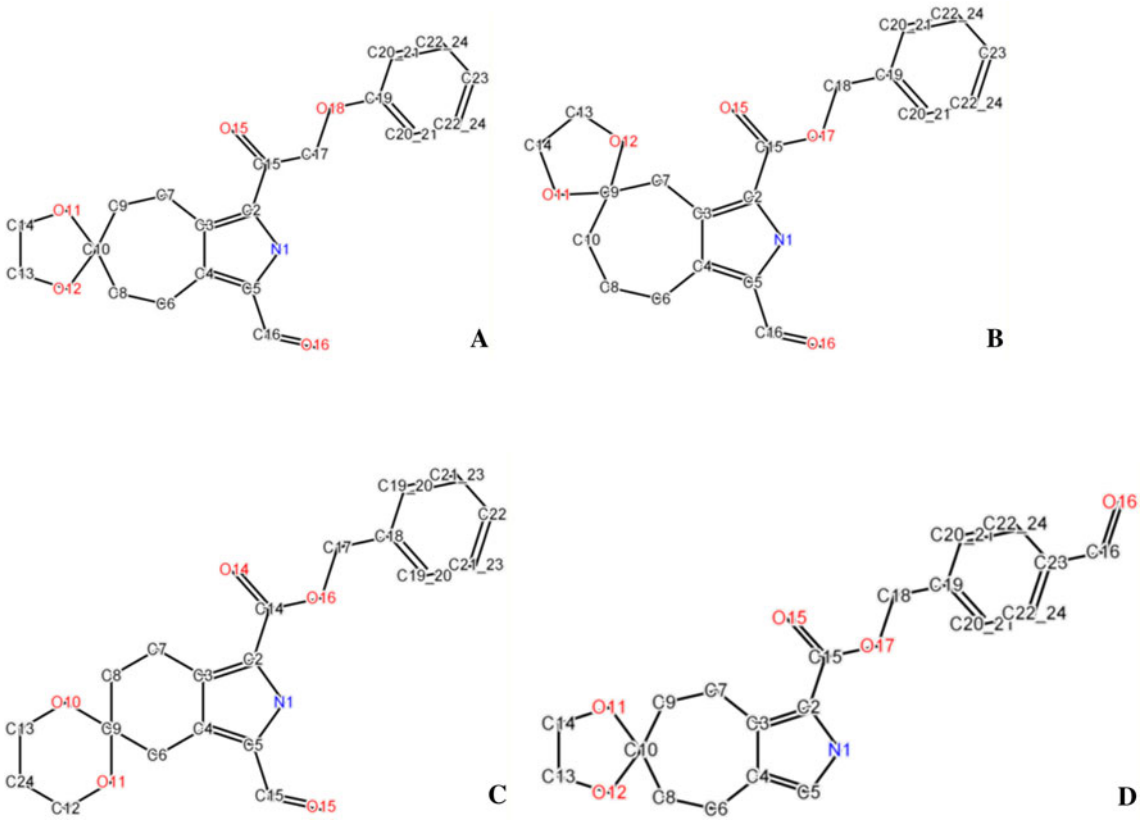

Fig. 10 Four regioisomers. These isomers are selected to demonstrate the ACA capability to discriminate the correct structure even within an ensemble of very similar structures

where lower values of the calculated term σ indicate smaller differences between the patterns of the measured and proposed empirical formulae (set value $m\sigma$ <50).

NMR experiments

Samples for a small NMR batch job consisting of 96 samples of different pyrroles, supplied by the University of Bremen, were dissolved in 10 mm^3 DMSO-d_6 each. Each of the solutions (8 mm^3) was then transferred from the source 96-well plate into 1 mm NMR tubes, which were arranged in a 96-well plate tube rack. All liquid handling steps were fully automated using a Gilson 215 Liquid handler (supplied by Bruker BioSpin, Germany). The NMR tube rack containing the 96 tubes filled with 1 mm NMR was then transferred to a SampleJet autosampler on a 400 MHz Avance III NMR instrument, equipped with a 1-mm MicroProbe (Bruker BioSpin, Germany). The quantitative 1D proton NMR experiments were set up from an industry-standard SD file using the CMC-q package (Bruker BioSpin, Germany), and the NMR experiments ran in full automation with 32 scans per sample with the standard parameter set supplied in this package. The overall experiment time was <5 min per sample, including the time for the automatic sample exchange.

NMR-based quantitation (CMC-q)

The NMR data were processed and analyzed automatically directly after the data acquisition using the CMC-q analysis tools provided within the CMC-q package (Bruker BioSpin, Germany).

NMR-based structure verification (CMC-i)

The 96 pyrrole samples were automatically analyzed using the latest version of the automated consistency analysis from PERCH Solutions Ltd., Kuopio, Finland (http://www.perchsolutions.com), which is also available under the product name CMC-i from Bruker BioSpin (Rheinstetten, Germany). The overall calculation time for all 96 samples was 257 min on a standard PC (Intel Q9400, 2.66 GHz, 3.25 GB RAM). This corresponds to a calculation time of <3 min per sample.

Acknowledgments We are grateful to Prof. Dr. F.-P. Montforts and Dr. M. Osmers (Institute of Organic Chemistry, University Bremen) for supplying the samples and fruitful discussions. We are grateful to Dr. Don Richards (Pfizer/UK) for his tremendous contribution to the concept of SmartFormula3D^{TM}.

References

1. Lacorte S, Fernandez-Alba AR (2006) Mass Spectrom Rev 25:866
2. Böcker S, Letyel MC, Lipták Z, Pervukhin A (2009) Bioinformatics 25:218
3. Roussis SG, Proulx R (2003) Anal Chem 75:1470
4. Ojanperä S, Pelander A, Pelzing M, Krebs I, Vuori E, Ojanperä I (2006) Rapid Commun Mass Spectrom 20:1161
5. Moriya F, Hashimoto Y (2003) Forensic Sci Int 131:108
6. Liotta E, Gottardo R, Bertaso A, Polettini A (2009) J Mass Spectrom 45:261
7. Rockwood AL, Van Orden SL, Smith RD (1996) Anal Chem 68:2027
8. Rockwood AL, Haimi P (2006) J Am Soc Mass Spectrom 17:415
9. Kaufmann A (2010) Rapid Commun Mass Spectrom 24:2035
10. Kind T, Fiehn O (2006) BMC Bioinforma 7:234
11. Kind T, Fiehn O (2007) BMC Bioinforma 8:105
12. Rockwood AL, Kushnir MM, Nelson GJ (2003) J Am Soc Mass Spectrom 14:311
13. Lane S, Boughtflower B, Mutton I, Paterson C, Farrant D, Taylor N, Blaxill Z, Carmody C, Borman P (2005) Anal Chem 77:4354
14. Wider G, Dreier L (2006) J Am Chem Soc 128:2572
15. Castellano S, Bothner-By AA (1964) J Chem Phys 41:3863
16. Laatikainen R, Niemitz M, Weber U, Sundelin J, Hassinen T, Vepsäläinen J (1996) J Magn Reson A 120:1
17. Laatikainen R, Niemitz M, Malaisse WJ, Biesemans M, Willem R (1996) Magn Reson Med 36:359