Efficient Structure Elucidation of Natural Products in the Pharmaceutical Industry

Serge Moss, Günter Bovermann, Regis Denay, Julien France, Christian Guenat, Lukas Oberer, Monique Ponelle, and Harald Schröder*

Abstract: The procedure of structure elucidation of natural products at Novartis is discussed. Structure elucidation in an industrial environment has to be time efficient, which is achieved by an optimized workflow and a close integration of all analytical methods used in this process. After a preliminary dereplication step based on LC/MS, compounds are delivered to the structure analysis unit. A second dereplication step is based on IR spectroscopic data. Subsequent MS analysis leads to the molecular weight or molecular composition. NMR spectroscopy is finally used for complete structure elucidation, respectively for confirmation of structural suggestions made on the basis of IR and MS data. Challenges like a large number of structure elucidation requests, low amounts of material, variation of sample purity or complexity of structures are faced with modern analytical equipment and high expertise in all disciplines.

Keywords: IR database · Natural products · Pharmaceutical industry · Spectroscopy · Structure elucidation

Introduction

Natural products play an important role in pharmaceutical industry. Historically, the first drugs were either plant extracts or isolated compounds originating from natural sources, such as morphine or penicillin. Today about 40% of the current drugs are related to natural products.[1] In 2000–2006 about 36% of all small molecular new chemical entities were either natural products or their structures were derived from natural products.[2] Natural compounds are used as antibiotics, immunosuppressive agents, analgesics, antivirals and as drugs in all kind of disease areas.[3,4] Despite this impressive success many major pharmaceutical companies have terminated or scaled down their natural product programs in the last decade.[5,6] The emergence of high-throughput screening and combinatorial chemistry has shifted the focus of drug discovery towards better automatable technologies.[1] To counteract the downward trends new strategies to enhance the efficiency of natural product analytics had to be developed.[6]

Novartis has a long history of natural product research which has contributed significantly to the product portfolio and sales (Fig. 1). More than 20 active clinical trials with compounds originating from natural sources or having a strong structural relationship to natural products are currently in Phase I, II or III of the Novartis pipeline. Three strategies are employed to obtain natural products: i) Screening of micro-fractionated extracts and follow-up isolation of active compounds, ii) activity-independent isolation of pure natural products, and iii) licensing-in compounds from external sources. Dereplication and structure elucidation is an integrated part of the process. Structure determination in an early stage of the process is advantageous to exclude known or pharmaceutically unattractive compounds from screening and to allow determination of structure–activity relationships. Structure elucidation of natural compounds is a very special challenge in analytics. The amounts of compound are in the range of some micrograms up to a few milligrams. Samples of unknown structures and diverse purity are sent to our analytical labs and the majority of structures are expected to be elucidated within 2 to 4 weeks. More than 500 compounds with a purity >85 % isolated by Novartis scientists and with new structures not yet available in the Novartis repository are added each year to the screening collection. It currently includes more than 10'000 natural products. As a consequence, a strategy has been worked out to cope with the large number of structure elucidation requests. This strategy is based on an efficient use of natural product structure and spectroscopic databases and the combination of the different spectroscopic methods available in our analytical department.

Procedure

Scheme 1 displays the workflow for the structure identification of natural products at Novartis. Compounds are isolated by the Natural Product Unit from various natural sources such as plants, fungi and bacteria. After a preliminary dereplication step based on LC/MS, compounds are submitted to the analytical group for structure elucidation.

At a first step, the sample is analyzed with infrared spectroscopy (IR), which is generally considered to be the analytical tool with the highest identification power due to its fingerprinting ability. Spectra are collected using the FTIR microscope. This technique has an outstanding sensitivity, allowing high-quality solid state spectra to be obtained with a few nanograms and minimal sample preparation.[7] IR spectroscopy is applied to check for structure

Correspondence: Dr. H. Schröder
Novartis Institutes for Biomedical Research
Lichtstrasse 35
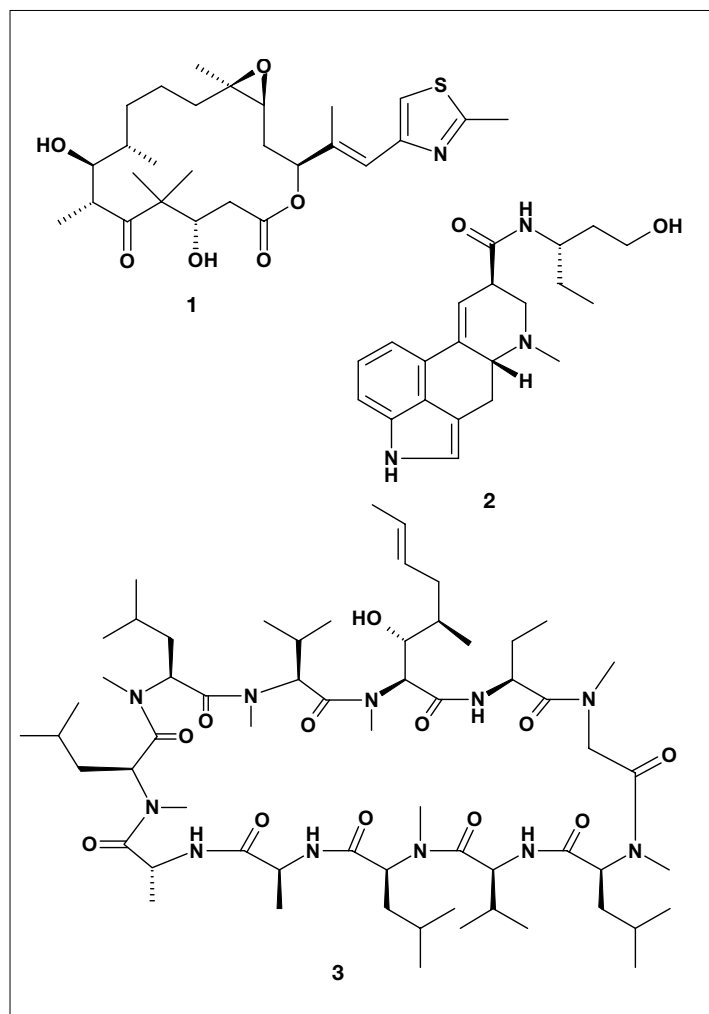CH-4056 Basel
E-Mail: harald.schroeder@novartis.com

Fig. 1. Novartis drugs: Epothilon B (**1**): Phase III, application in Oncology against solid tumors; Methylsergide (**2**): used against migraine; Cyclosporin A (**3**): Sales in 2006 were 900 Mio US$, used as immunosuppressor after transplantations

Scheme 1. Compounds are isolated by the Natural Product Unit and submitted to the analytical group. A dereplication step based on IR spectroscopic data is followed by MS to determine the molecular weight or molecular composition. NMR spectroscopy is finally used for complete structure elucidation. Structure searches and spectroscopic queries support the elucidation process.

identity and to determine the class of compound and functional groups exploiting our in-house library of FTIR spectra of natural compounds by using the search program of SPECTACLE (LabControl). This library contains currently more than 12 000 condensed phase reference spectra of different compounds of high purity. Most spectra were collected from structures which have been identified or elucidated in our group during the last ten years.

The IR query yields three different types of results:

i) If a reference spectrum of the compound to be identified is present in the library, this spectrum is normally found at the top of a match list. All spectra are measured with the IR-microscope in the transmission mode to ensure a homogeneous spectroscopic database for recent data. Older solid-state spectra have been measured with the potassium bromide pellet technique and can show slight differences, but in general these are not significant whereas a deviation in purity or crystallinity would have a greater impact of the matching. Crystallinity effects could in principle be avoided by conducting measurements in solution, but the drawback of this alternative would be the greater amount of sample needed and the higher time consumption. Measuring the sample in the aggregate state has also the advantage that only tiny amounts of sample is needed, leaving the main part of the sample untouched and thus preventing the sample from a possible decomposition in solution. Nevertheless, pure known compounds rarely pass through the IR filter even when the IR microscope is used. Thus identified compounds are confirmed by MS in order to achieve an unequivocal identification. About 20% of the samples can be identified by IR and MS and therefore be ruled out at this level.

ii) If an exact match is not found, the similarity of the best hits with the unknown spectrum can bring conclusive information on the structural class of the unknown compound. Peptides, diketopiperazines, polyenes, terpenoids and even more specific structure families like anthracyclines, beauverolides, chaetoglobosins, lipiarmycins, niphimycins and many more can be determined at a glance at this very early stage.[8] By using our library, which contains spectra of compounds of high chemical diversity from different biological sources, the majority of the unknown compounds can be related to a registered structure or assigned to a known class of compound at this step. It is worth noting that an early identification of the structure class speeds up the overall process, especially
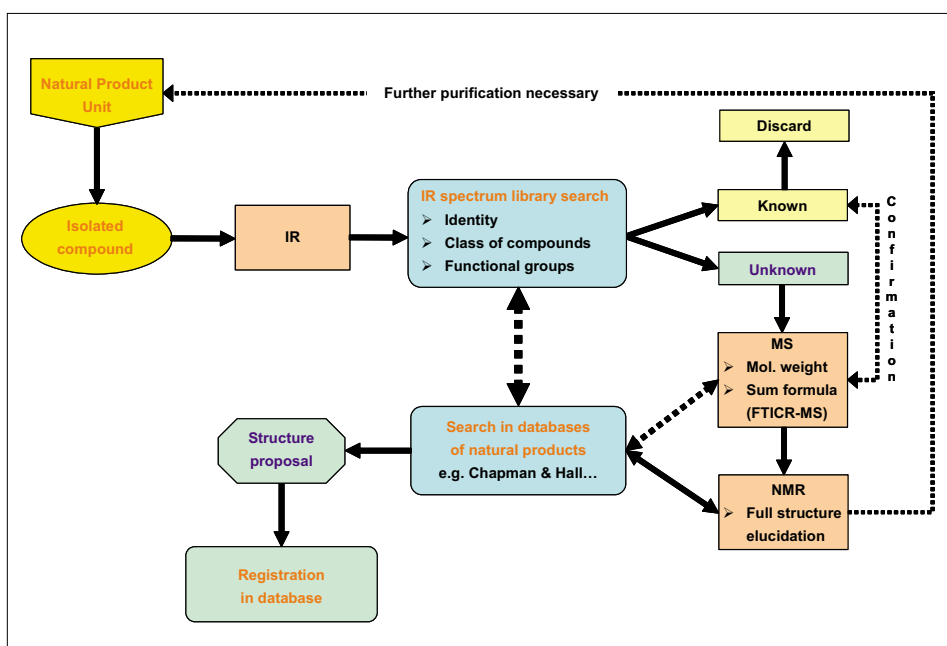
when the compound is not described in the literature.

iii) When a new compound is not documented in the library and cannot be assigned to a known type of compound, the IR spectrum provides at least information on functional groups and molecular fragments, which is also very useful for subsequent structure elucidation steps.

In the second step of the process, all samples are analyzed by mass spectrometry (MS) or liquid chromatography/mass spectrometry (LC/MS) to determine the mass and if needed the molecular composition by applying Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS).[9] MS measurements are generally performed in two steps. First, molecular weight information is determined by acquiring low resolution mass spectra in positive and negative ion modes. Flow injection is the preferred sample introduction method because the compounds are purified and it allows several measurements to be carried out under different experimental conditions. This setup has proven to be successful with labile natural products, which may readily fragment (*e.g.* water molecule loss) under harder ionization conditions, such as atmospheric pressure chemical ionization. If no satisfactory results are obtained, other mass spectrometric tools are applied, such as LC/MS (if matrix compounds generate ion suppression) or electron impact ionization (if compounds do not ionize in electrospray).

For compounds that could not be sufficiently characterized by mass and IR spectroscopy, accurate mass measurement in the sub ppm range is performed using FTICR-MS. To minimize the number of possible formulae, especially at higher masses (*e.g.* above 800) the input from other spectroscopic is often helpful.

In the third step of the process, NMR spectra are acquired. Using the NMR shift information combined with results from IR and MS a search in internal and commercial databases like the Dictionary of Natural Products[10] is performed and further structural suggestions can often be made. For a summary of natural product databases see ref. [11].

The structure confirmation of such a suggestion, respectively the final structure elucidation is based on NMR spectroscopy.[12] One- and two-dimensional datasets, such as COSY, ROESY, TOCSY, HSQC and HMBC allow a complete *de novo* structure elucidation for most of the compounds. ^{13}C NMR spectra are rarely measured due to higher time and sample requirements. Although NMR is the most material demanding method compared to IR and MS, new techniques like cryoprobes or microprobes have reduced the necessary amount to 20–50 µg of pure sample to perform 2D NMR

experiments. In the case of small amounts of available samples we use the 1-mm microliter NMR probe, which gives a higher spectrum quality compared to the cryoprobe due to smaller solvent signals and fewer impurities from the NMR solvent.

A practical, simple example of a rapid structure elucidation is given (Fig. 2–4). Compound **5** was submitted for structure elucidation by the Natural Product Unit. The IR spectrum search of the unknown (**5**) yielded the spectrum of Geodin (**4**) as the best match of the query (Fig. 2). A mass of 364 amu was determined (Fig. 3), confirming by the isotopic pattern that a monochloro-derivative of Geodin could be present. Five compounds having the same mass and containing one chlorine atom were found in the Dictionary of Natural Products.[10] Three of them could be excluded based on the IR and ^{1}H NMR (Fig. 4) spectra. To distinguish between the remaining two and to exclude other structures 2D HSQC and HMBC NMR experiments were performed. By combining the information from IR, MS and NMR, the structure of compound **5** could be elucidated.

Summary

Each year, more than 1000 natural product samples are sent to the analytical group at Novartis for structure elucidation. The IR spectroscopic database query helps to identify known structures (20%) and to propose the compound class or structure family for the majority of the remaining samples. Combining the expertise in IR, MS and

NMR allows structures to be proposed for about 80% of the requests. The remaining 20% of the samples are terminated without a final structure proposal because of insufficient amount of sample, purity, spectroscopic quality or instability of the sample. These ratios have not remarkably changed in the past years. The combined application of three analytical disciplines and the tight collaboration of the experts in the Analytical and Natural Product Units ensure an efficient process of structure elucidation, which supports the Drug Discovery by adding each year more than 500 pure structure-elucidated natural products to the Novartis compound collection. About 65% of these isolated compounds are even unknown in literature.

Experimental

IR Spectroscopy

IR spectra are usually collected on a Bruker Vertex 70 FT-IR spectrometer (Bruker Optics, Fällanden, Switzerland) linked to a Hyperion 2000 FTIR microscope. The spectrometer is fitted out with a DTGS detector and allows measurements from 4000 to 400 cm^{-1}. The measurement range for the microscope equipped with a liquid nitrogen cooled MCT detector is 4000 to 600 cm^{-1}.

For measurements with the FTIR microscope, a tiny amount of substance is transferred to an IR-transparent plate and then flattened to an appropriate thickness by rolling with a metallic probe. Areas of 50 × 50 µm are usually selected with an adjust-
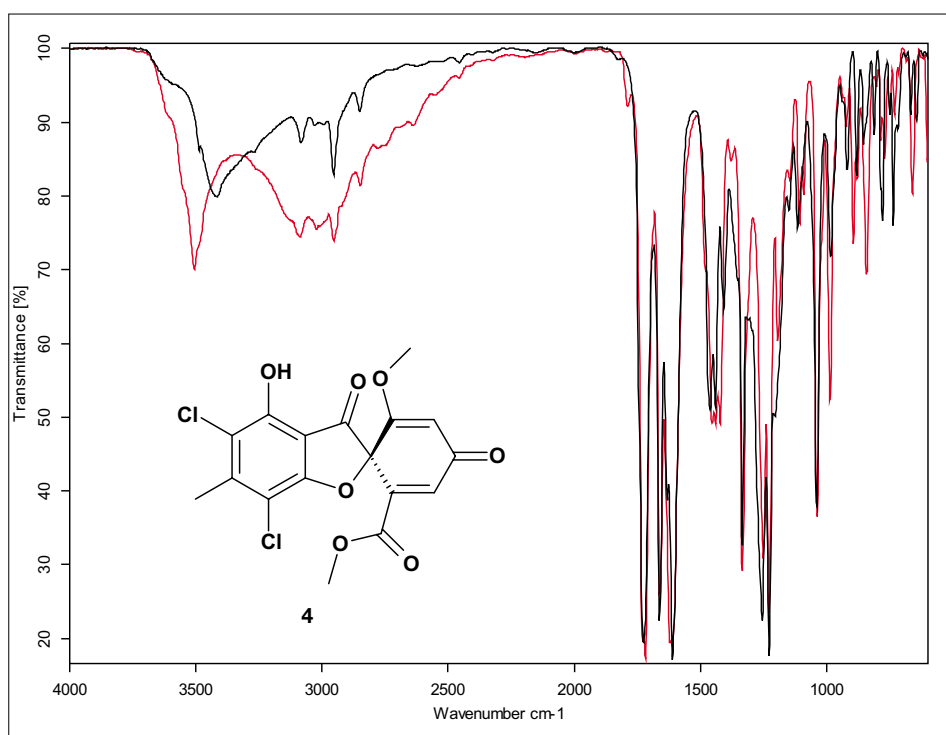

Fig. 2. Result of the IR query of **5 (black line)**, Geodin (**4, red line**) matched best
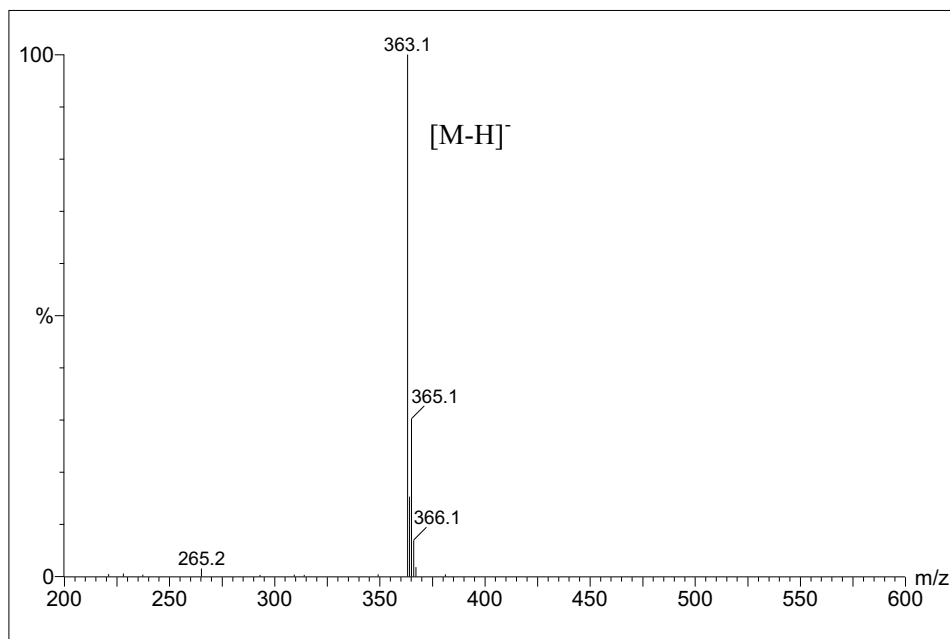
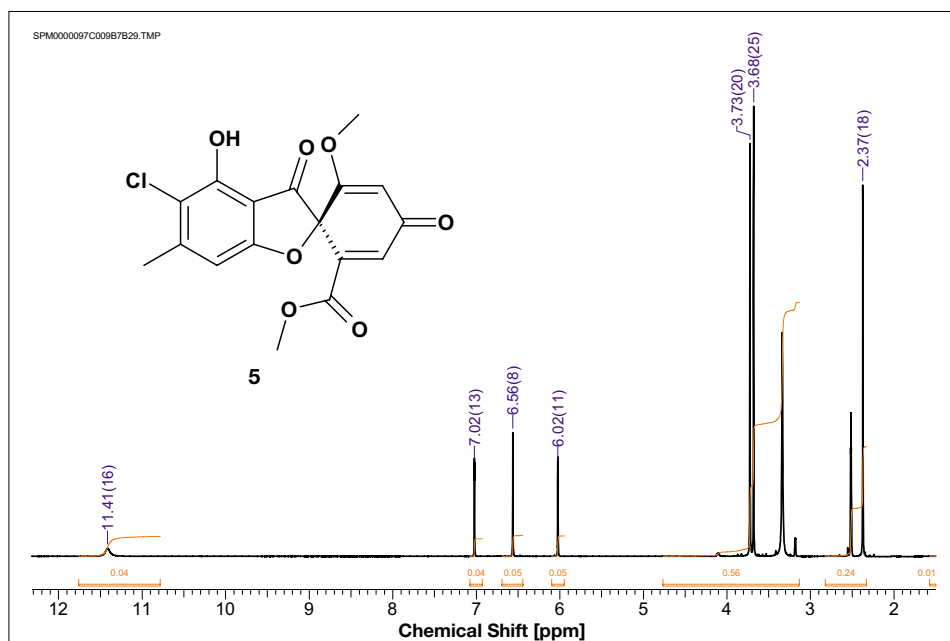Fig. 3. Negative ESI-MS spectrum of **5** showing the [M-H]⁻ ion with a characteristic chlorine atom isotopic pattern.

Fig. 4. ¹H-NMR spectrum of **5** in d6-DMSO

ferred into a capillary NMR tube and measured 24–48 h to obtain a usable set of 2D NMR spectra. Depending on the molecular weight of the compound, the required amount of material would be in the range of 20–50 μg. Simple 1D ¹H spectra can be measured with less than 1 μg.

The following 2D NMR experiments are used: COSY (phase sensitive with double quantum filter), ROESY[13] (phase sensitive with 180x/180-x spin-lock and a mixing time of 200–500 ms), TOCSY (mixing time 80 ms), HSQC, DEPT-HSQC[14] and HMBC.

[1] O. Potterat, M. Hamburger, *Curr. Org. Chem.* **2006**, *10*, 899.

[2] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2007**, *70*, 461.

[3] D. G. Kingston, D. J. Newman, *IDrugs* **2005**, *8*, 990.

[4] V. P. Gullo, J. McAlpine, K. S. Lam, D. Baker, F. Petersen, *J. Ind. Microbol. Biotechn.* **2006**, *33*, 523.

[5] M. S. Butler, *Nat. Prod. Rep.* **2005**, *22*, 162.

[6] K. U. Bindseil, J. Jakupovic, D. Wolf, J. Lavayre, J. Leboul, D. van der Pyl, *Drug Discov. Today* **2001**, *6*, 840.

[7] 'Practical Guide to Infrared Microspectroscopy', Ed. H. J. Humecki, Marcel Dekker, Inc., Basel, **1995**.

[8] a) B. Roggo, P. Hug, S. Moss, A. Stämpfli, H.-P. Kriemler, H. H. Peter, *J. Antibiot.* **1996**, *49*, 374; b) B. Roggo, P. Hug, S. Moss, F. Raschdorf, H.H. Peter, *J. Antibiot.* **1994**, *47*, 143.

[9] J. Zhang, G. McCombie, C. Guenat, R. Knochenmuss, *Drug Discov. Today* **2005**, *10*, 635.

[10] 'Dictionary of Natural Products', CD-ROM, **2005**, version 15.2, Chapman & Hall, CRC.

[11] M. Füllbeck, E. Michalsky, M. Dunkel, R. Preissner, *Nat. Prod. Rep.* **2006**, *23*, 347.

[12] a) L. Révész, P. Hiestand, L. LaVecchia, R. Naef, H.U. Naegeli, L. Oberer, H.J. Roth, *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1521; b) T. Fehr, J. Kallen, L. Oberer, J. Sanglier, W. Schilling, *J. Antibiot.* **1999**, *52*, 474.

[13] T. L. Hwang, A. J. Shaka, *J. Am. Chem. Soc.* **1992**, *114*, 3157.

[14] W. Willker, D. Leibfritz, R. Kerssebaum, W. Bermel, *Magn. Reson. Chem.* **1993**, *31*, 287.

able aperture to perform measurements in the transmission mode. The measurement time is less than 1 min. by accumulating 32 scans with a spectral resolution of 4 cm⁻¹. Usually two different measurements are performed with the same sample to check its homogeneity.

MS Spectrometry

Flow injection measurements are performed on a Waters Quattro II triple quadrupole mass spectrometer equipped with an electrospray ionization (ESI) source. A Bruker APEX *III* Fourier transform ion cyclotron resonance mass spectrometer is used with internal calibration to derive molecular formula.

NMR Spectroscopy

NMR spectra are measured on different spectrometers, depending on the available amount of compound. The following NMR spectrometers and equipment are used: DPX400 with 5-mm ¹H {¹³C}{¹⁵N} probe, DMX500 spectrometer with 5-mm ¹H {¹³C}{¹⁵N} probe, DRX500 with 5-mm ¹H {¹³C}{¹⁵N}cryoprobe, AV600 with 1-mm ¹H {¹³C}{¹⁵N} microliterprobe and AV600 with 5mm ¹³C {¹H}dual probe. All spectrometers are from BRUKER (Bruker Biospin AG, Fällanden, Switzerland) and all probes are equipped with z-gradients.

The 1-mm microliter probe is used for mass-limited samples. For instance 5–7 μl of a 10 mM compound solution are trans-