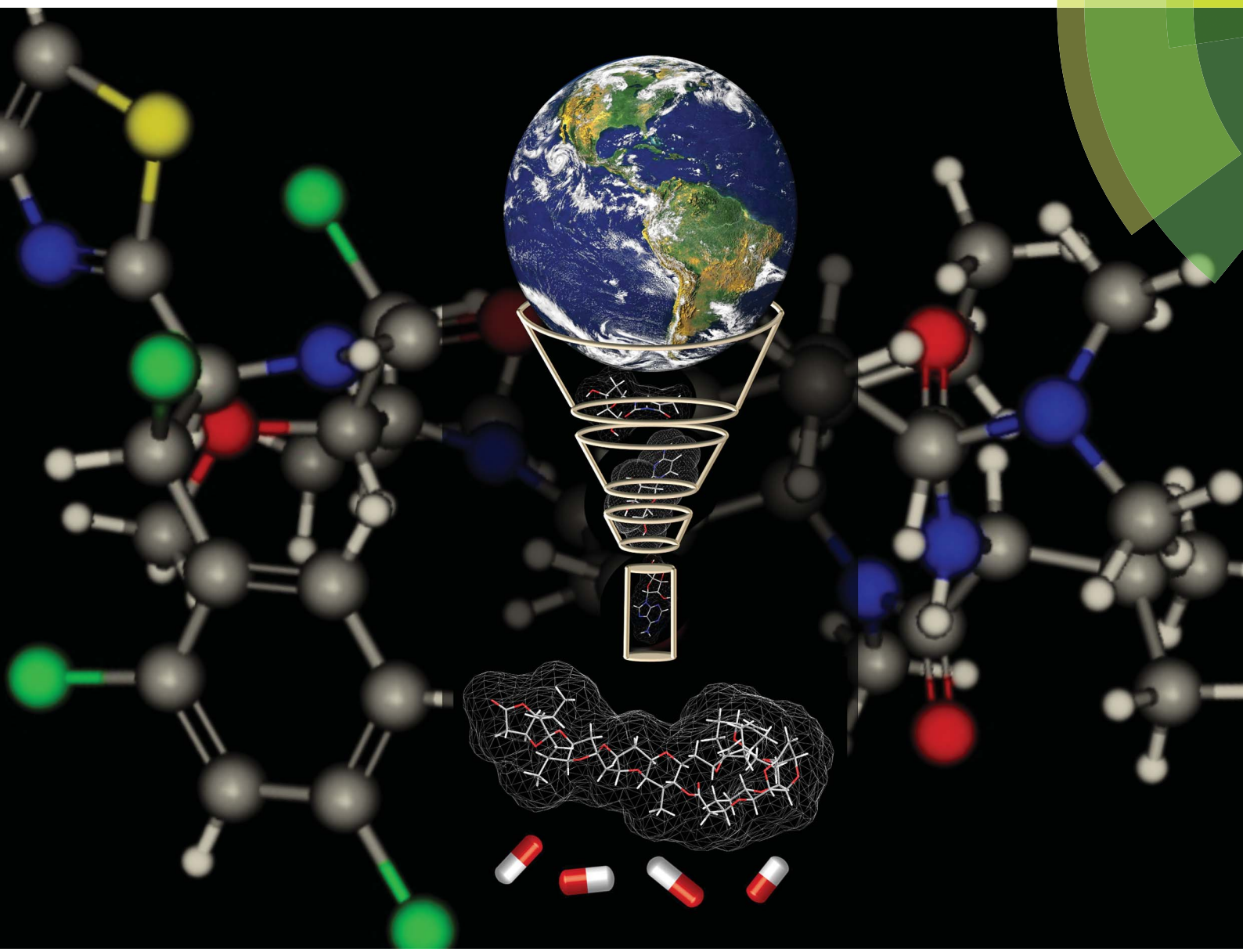
Volume 32 | Number 6 | June 2015 | Pages 755–878

# NPR

Natural Product Reports

www.rsc.org/npr



ISSN 0265-0568

ROYAL SOCIETY OF CHEMISTRY

REVIEW ARTICLE

Susana P. Gaudêncio and Florbela Pereira

Dereplication: racing to speed up the natural products discovery process

NPR

REVIEW

View Article Online
View Journal | View Issue



Cite this: *Nat. Prod. Rep.*, 2015, **32**, 779

# Dereplication: racing to speed up the natural products discovery process

Susana P. Gaudêncio*[ab] and Florbela Pereira[a]

Covering: 1993–2014 (July)

To alleviate the dereplication holdup, which is a major bottleneck in natural products discovery, scientists have been conducting their research efforts to add tools to their "bag of tricks" aiming to achieve faster, more accurate and efficient ways to accelerate the pace of the drug discovery process. Consequently dereplication has become a hot topic presenting a huge publication boom since 2012, blending multidisciplinary fields in new ways that provide important conceptual and/or methodological advances, opening up pioneering research prospects in this field.

Received 21st October 2014

DOI: 10.1039/c4np00134f

www.rsc.org/npr





## 1 Introduction

Nowadays, the presumably high productivity in terms of the number of novel bioactive compounds isolated has not yet led to a corresponding increase in the number of new drug candidates. Instead, after more than two decades of combinatorial chemistry research, a declining number of new chemical entities (NCEs) in the drug development pipeline has been observed,[1] Fig. 1.

This phenomenon combined with the higher success rate of drug discovery obtained from the marine world (1 in 3140 marine natural products) compared with the industry average (1 in 5000–10 000 compounds) has led to the rekindled interest in natural products-like scaffolds.[2] Natural products (NPs) sources are well known to produce chemical metabolites with unique features, highly complex structures and properties for human health care and well-being, exhibiting a wide range of applications that have inspired a number of industrial arenas. There is no doubt that NPs are the most consistently successful source of drug leads, as can be seen from Fig. 1. The urge to fill the industrial pipeline and to discover novel lead-like compounds for drug discovery, which can meet the societal challenge of the lack of suitable therapeutic agents for a broad range of diseases, has never been greater. Antibiotic resistance, for instance is a "ticking time bomb", we are currently on "red alert", having a poor drug repertoire, in which commonly treated infections are becoming lethal. One dominant tailback in NPs discovery is dereplication, *i.e.* the discard of known compounds. With the ultimate objective of speeding up and improving drug discovery program efficiency, researchers have been using multifaceted approaches, either merging different areas of knowledge or creating totally innovative ways to advance this field. Consequently, dereplication, which is the object of our review, has become a matter of great interest in recent years. The

[a]*LAQV, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. E-mail: s.gaudencio@fct.unl.pt; Fax: +351 212948550; Tel: +351 212948300*

[b]*UCIBIO, REQUIMTE, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*

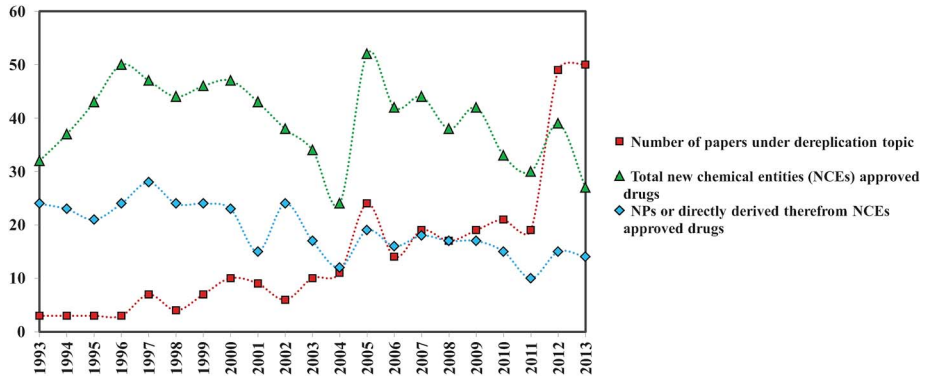


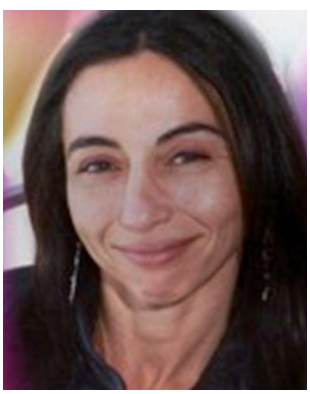

Fig. 1 Comparison between dereplication outputs, NPs approved drugs and total NCEs approved drugs. (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/ucm20025676.htm).[1]

importance of dereplication can be proven by the significant increase of publications covering this topic since 2012 (Fig. 1). Therefore, we believe that the development of multidisciplinary dereplication processes, which will be highlighted in this review, will certainly result in an impressive enhancement in the number of NPs (or directly derived therefrom) and approved drugs in the imminent future. Our insight into this theme will cover the period from 1993 to 2014 (July). In this review we do not intend to give an exhaustive description of the focused techniques and methodologies. Instead we will refer to their use from the NPs dereplication point of view, giving priority to significant reported progress in this area, emphasizing key developments that have shaped the field, trends and describing future directions.

## 2 Reviews

Over forty reviews on the "dereplication" topic have been reported, according to the information obtained from ISI Web of Science™ and NPR archive. Hence, a selection of several reviews focusing on several and diverse NPs dereplication aspects will be highlighted. A vast number of reviews describing the enhanced technological progress of hyphenated tools and combined dereplication strategies that have been pivotal to accelerate the speed of novel NPs isolation process have been reported.[3–7] Others particularly debate the emerging advances in chromatography, mass spectrometry (MS) and NMR technologies to overcome the challenges encountered in screening NPs libraries, increasing the role of NPs in high-throughput screening (HTS) based drug discovery.[8–11] The approaches used in biodiversity- and taxonomy-guided microbial NPs library construction, combined with HTS and with Liquid Chromatography-Mass Spectrometry (LC-MS) have been promoted as efficient dereplication processes.[12] Furthermore, significant advances in the field of computational mass spectrometry for NPs research, highlighting recently developed methods for the detection and investigation of small molecules, namely MetFrag/MetFusion, ISIS, FingerID, and FT-BLAS, were reviewed.[13] The latest applications of Imaging Mass Spectrometry (IMS)

*Susana P. Gaudêncio is currently a Researcher Chemist at LAQV and UCIBIO REQUIMTE-Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Portugal. She holds degrees in technological organic chemistry (M.S. 2001) and in organic chemistry (Ph.D. 2006), was a postdoc at Scripps Institution of Oceanography (SIO), UC-San Diego (2007–2010). Susana's early interests have crossed several disciplines including the chemistry and ecology of marine sponges, and the total synthesis of staurosporine precursors. Presently she is the PI of several projects related to marine microbiology, marine natural products drug discovery and chemical ecology, particularly targeting microorganisms from the Macaronesia Atlantic ecoregion.*



*Florbela Pereira is currently a Researcher Chemoinformatics at LAQV REQUIMTE-Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Portugal. She holds degrees in chemistry – analytical chemistry (1991) and in natural products (Ph.D. 1997) from the University of Aveiro. She undertook postdoctoral research on the development of pharmaceutical products at the Instituto de Biologia Experimental e Tecnológica (IBET). During 2000–2005, Florbela worked in industry as a synthetic chemist in organic synthesis of agrochemical active substances, before returning to academia to join the chemoinformatics and marine natural products groups. Her main scientific interests include machine learning and data mining techniques.*

technologies, challenges, pitfalls and prospects for improvement and application to microbial NPs were reviewed.[14] Existing and emerging technological advances referring to MS related to NPs dereplication have been extensively reviewed and described.[15] Recent studies emphasized metabolomics-driven analysis, using modern mass spectrometry techniques, as a key approach for the discovery of NPs from microbial sources.[16] Chemotyping/metabolomics as part of screening, using direct metabolite profiling techniques such as direct injection MS or NMR have been reviewed, focusing on how it can be used for the discovery of novel compounds in combination with modern methods for dereplication, avoiding redundancy in the selection of microorganism species.[17]

As far as NMR is concerned, in a recent review,[18] technical improvements were headlined, such as miniaturized and cryogenic NMR probes along with hyphenation capabilities and computational support and showing that nowadays it is not always necessary to separate the components of a mixture in order to obtain spectroscopic information from its constituents, minimizing isolation and dereplication efforts. However, capital and operating costs of commercial instrumentation for this purpose are still highly-priced.[18] Another published review article described and discussed the analytical techniques of dereplication and related technologies for quantification and structure identification of small-amount compounds from limited amounts of natural source extracts.[19] Technical improvements such as miniaturized and cryogenic NMR probes, which allow good spectral data to be obtained with 1 mg of sample or less, are in vogue and are highlighted in more than one review.[20] Reviews of the state of the science in what regards computer assisted structure elucidation (CASE) developments were produced by Elyashberg (2010),[21] Jaspars (1999),[22] and Steinbeck (2004).[23] The CASE systems aim to minimize structure elucidation difficulties, attribution errors and speed up structure elucidation. Ultimately, becoming a tool for structure elucidation and enabling the determination of stereochemistry. Approaches involving chemical biology applied to marine bacteria and *in silico* methods have been overviewed as successfully useful to the discovery of novel antibiotics.[24] Technological and philosophical strategies were reviewed addressing antifungal NPs discovery bottlenecks, associated with NPs screening and dereplication.[25]

*In vitro* bioassays incorporated during various stages of research and development are regarded as playing a vital role in evaluating botanicals.[26] Several chromatographic separation techniques of plant material, and other related issues such as dereplication procedures during NPs isolation have been discussed in a couple of reviews.[27,28]

# 3 Facts and statistics

As illustrated in the following graphics, Fig. 2 and 3, dereplication has become a markedly pursued subject in recent years.

The present statistical analysis comprises dereplication literature indexed in Web of Science™ Core Collection, Current Contents Connect®, Derwent Innovations Index℠, MEDLINE® and SciELO Citation, prior to July 8th of 2014, in a total of 340 publications (277 articles, 21 proceedings and 42 reviews). All publications incorporating the dereplication subject were analyzed. Prior to 1993 until 2011 there was a steady increase in the number of publications/year rising from 3 in 1993 to 19 during 2011, there was a notable increase in the number of reports since 2011 (approximately 89%), corresponding to 49 and 50 publications in 2012 and 2013, respectively. Worldwide more than three-quarters of the publications (*i.e.* 205 out of 243, corresponding to 84%), during the last ten years, are linked to ten countries, namely USA, Denmark, Switzerland, Brazil, Germany, China, Australia, Netherlands, Canada and England. However, 170 publications (out of 243, corresponding to 70%) are related with several other countries, Fig. 4 demonstrates these data. It was taken into account that the papers may include researchers from several countries.

The most productive country in this field has been USA, contributing 69 outputs across this period (28%). However, Denmark is the country with more citations, 829 citations out of 5297 citations during the mentioned period (16%), compared with 778 citations (15%) obtained by USA. The publications associated with the top ten mentioned countries (Fig. 4) have accumulated the highest number of citations, corresponding to approximately 79% of the overall citations.

Over 120 journals have been selected as containing publications relating to dereplication, during 1993–2014, the top ten journals reporting the above theme are listed in Fig. 5.

The papers and citations related to these top ten journals correspond to *ca.* 43% and 43% of the complete publications and the accumulated citation number during the mentioned period, respectively. Interestingly, *J. Nat. Prod.* is the top selected journal, by a considerable margin (35 publications, corresponding to 14%), and the second most cited (603 accumulated citations, corresponding to 11%). Only *J. Antibiot.* exhibits more accumulated citations (682 citations, corresponding to 13%). Nevertheless, *J. Antibiot.* includes the record cited paper in the field,[29] with 638 citations to date, having an amazing average of 63.8 citations per year.

Although we perceive that older publications accumulate more citations, the exponential increment of citations per year, for dereplication interrelated publications obtained during 1993–2014 is quite remarkable, Fig. 3. In fact, the upsurge is more impressive taking into consideration that approximately 80% of the total number of reports dealing with this matter have been published in the last ten years (*i.e.* since 2004, inclusive) and that nearly half of these 36% were published from 2011 to 2014 (July).

# 4 Dereplication trends and approaches

Chronologically the use of the word "dereplication" and the first issues related to this theme were reported in 1978.[30] After a long period of "silence" in 1990 a second manuscript was published.[31] Later, in 1993 research efforts started to be progressively implemented in this field.[32,33] At the end of the last

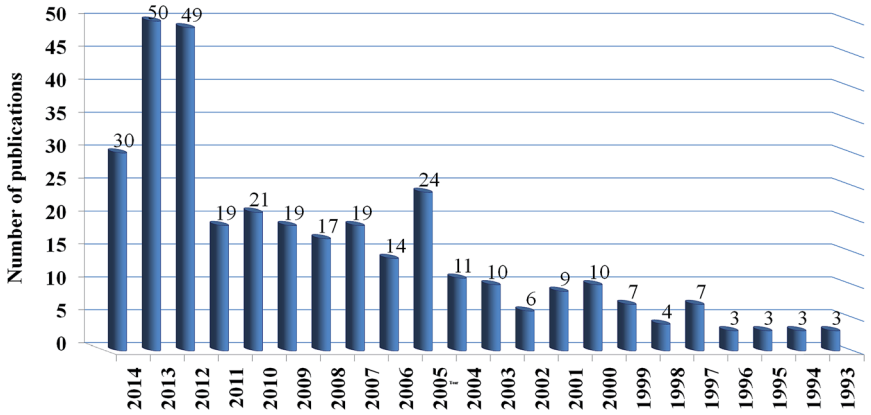



Fig. 2 Number of publications per year covering dereplication topic, period 1993–2014. Data source from Web of Science™ Core Collection.

century dereplication techniques mainly involved the use of biological screening processes,[33,34] LC-MS techniques[35–37] with implementation of MS libraries and databases,[32,38] allowing hit searching to some extent (*e.g.* chemical structure, molecular formula, molecular weight, bioactivity and taxonomy). Meanwhile, enhancements in the dereplication process such as substructure investigation started to be promoted,[32] the use of similarity searches over databases of estimated $^{13}C$ NMR-spectra for NPs structure identification,[38] the first computer assisted structure 2D NMR elucidation methodologies (CASE)[22,39–42] as well as HTS.[9] During the considered period we specially highlight the contribution of Cordell *et al.*[43,44] At the turn of the century and during the early 2000's taxonomic identification started to be used as a dereplication tool (16S rDNA sequencing)[45–50] in addition to the improvement and advent of novel hyphenated technologies/techniques (*e.g.* LC-MS, LC-DAD, LC-NMR, LC-NMR-MS).[37,51–54] Methodologies using hyphenated techniques combined with bioactivity guided assays (*e.g.* HPLC-ESMS-bioassays, TLS/ESMS/bioassays)[55–58] have also been reported. MS methods were improved and used as dereplication tactics (*e.g.* Q-TOF-MS-MS multistage, IT-MSn, LC-MS-MS, ESI-TOF-MS),[59,60] as well as cryoprobe, gradient and chiral NMR technologies[61–63] and further advances in the field of CASE.[23,64,65] Commercial databases became available in the

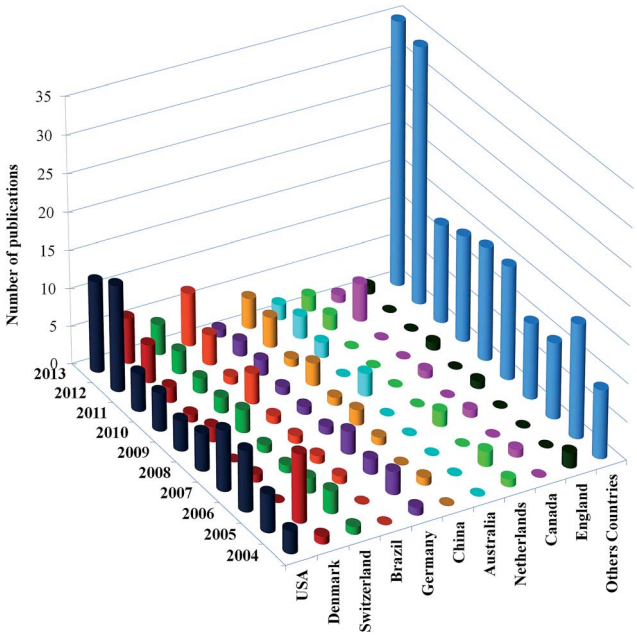



Fig. 4 Number of publications per year and per country covering dereplication topic, period 2004–2013. Data source from Web of Science™ Core Collection.

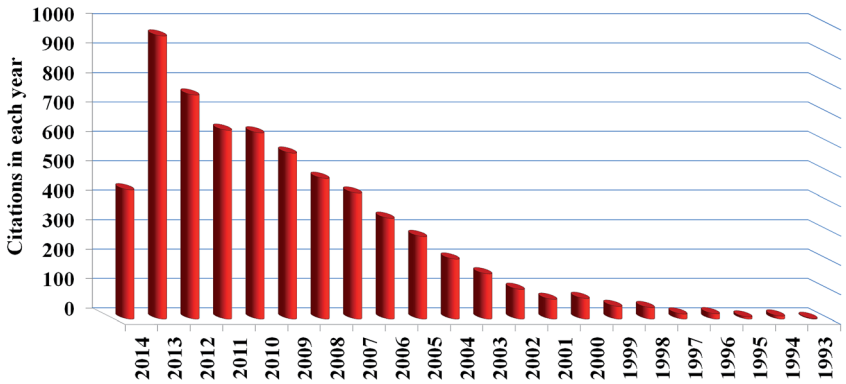



Fig. 3 Number of citations per year covering dereplication topic, period 1993–2014. Data source from Web of Science™ Core Collection.

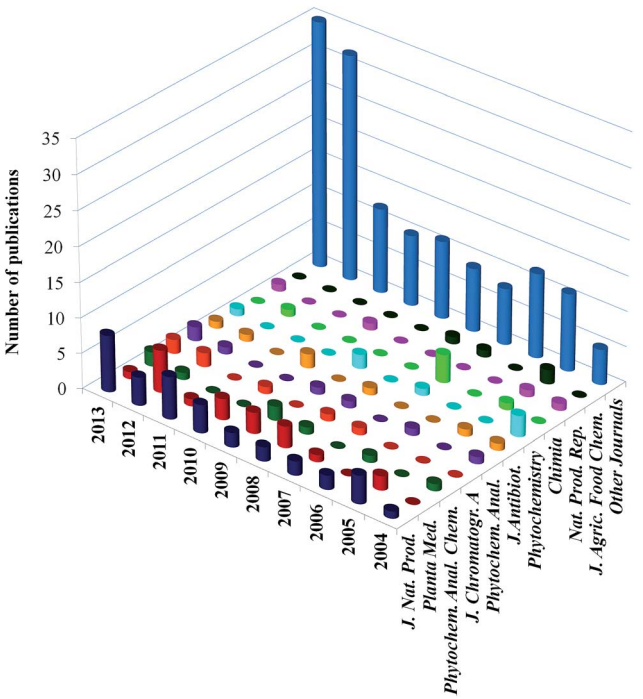



Fig. 5 Analysis of the selected journals reporting dereplication topics, period 2004–2013. Data source from Web of Science™ Core Collection.

market (*e.g.* NPs libraries such as MarinLit, LC-MS-MS libraries).[32,66,67] From 2005 to 2011 analytical spectroscopic and spectrometric methods levelled up exceptionally, dereplication methodologies met these new techniques, promoting and developing novel multifaceted approaches. For hyphenated techniques a wide array of possible combined sets was used (*e.g.* HPLC-DAD-SPE-NMR, LC-MS-ELSD),[68] and exhaustive detailed descriptions have been reported.[4,6,69–83] The introduction of X-hitting algorithm,[84] capillary scale NMR probes,[85] MALDI-TOF imaging,[86–88] advanced public and commercial database availability (*e.g.* Dictionary NPs, Antibase, MarinLit, AntiMarin, Pubchem, ZINC, NAPROC-13, NMRShiftDB, GNPS),[12,89–95] *de novo* sequencing techniques,[96] *in silico* dereplication,[24] computer assisted numerical analysis,[97,98] bioinformatics,[24,29] genomics,[99] proteomics,[80] metabolomics have all been reported.[17,100–104] From 2012 to mid-2014 we have seen an incredible increasing amount of publication data with a dereplication focus, using up dated and enhanced trends comprising analytic techniques, databases, combined procedures and blending scientific fields with the main goal of accomplishing faster, accurate and efficient dereplication,[105] using smaller amounts of samples.[106,107] In particular, hyphenated/combined techniques,[108–111] MS,[112–115] MS-MS networking,[15,95,116] IMS,[117] and NMR[118] as well as genomics,[119] bioinformatics[120,121] and metabolomics[122–124] approaches[19,28,95,125–127] have played a major role in dereplication. Many of the above mentioned methods report important advances in this field and contributed to fundamental developments that have made dereplication as it is performed nowadays. These will be emphasized in leading topics presented in the following sections.

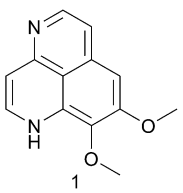
## 5 High throughput screening (HTS)

In general, high throughput research relies on the automation of large scale repetition experiments making such trials achievable. At the turn of the century the potential of new technologies to enhance HTS and use it as a parallel dereplication method started to be discussed.[9] HTS allows millions of chemical, genetic or pharmacological tests to be quickly conducted. Through this process it is possible to rapidly identify active compounds, antibodies, or genes that modulate a specific biomolecular pathway, providing hits for drug design and for the understanding of the interaction or role of a particular biochemical process.[128,129] Automation is a key element in HTS's efficacy, using robotics, data processing, control software, liquid handling devices, sensitive detectors, readout or detection, HTS robots can test up to 100 000 compounds per day.[130] Nevertheless, when seeking changes or defects that a computer may not easily determine manual measurements are necessary.[131] Automatic strain pickers select thousands of microbial strains for high throughput genetic screening.[132] Around 2008 *ultra-high-throughput screening* (uHTS) technologies, patented by Queeney and Hughes,[133] made possible the screening of more than 100 000 compounds per day.[133–135] Possessing the ability to rapidly screen diverse NPs to identify bioactive compounds, HTS has led to an eruption in the rate of generated data in recent years.[136] One of the most vital challenges in HTS experiments is to glean biochemical significance from an enormous amount of data, which relies on the development and adoption of appropriate experimental design and analytical methods for both quality control and hit selection.[137] Consequently, high-quality HTS assays are critical in HTS experiments. The development of high-quality HTS assays requires the integration of both experimental and computational approaches for quality control (QC). Effective analytical QC methods serve as gatekeepers for excellent quality assays.[138–142] In 2010, Weitz and coworkers reported an HTS process performing 1000 times faster screening (100 million reactions in 10 hours) at 1-millionth the cost (using $10^{-7}$ times the reagent volume) of conventional techniques using drop-based microfluidics.[143] A silicon sheet of lenses placed over microfluidic arrays with fluorescence measurement of 64 different output channels simultaneously with a single camera, analyzes 200 000 drops per second.[144] Standard HTS is now being tied to cell biology, mainly using technology such as high-content screening (HCS). High throughput cell biology requires methods that can take routine cell biology from low scale research to the speed and accuracy that allows the entire genome to be looked at, very rapidly. It will have its most significant impact in exploring biology towards cell models as a system progress rather than isolated pathways.[132] Advances in the field of tissue bioengineering aim to enhance the success of drug candidates through pre-clinical optimization. Models that are most amenable to HTS with emphasis on detection platforms and data modeling, 3D microorganoid systems will play an increasing role in drug testing and therapeutics over the next decade. Nevertheless, important hurdles remain before these models are fully developed for HTS.[145]

Despite the broad array of possible HTS tests, accelerating data-collection and hit selection processes, to avoid the re-

discovery of known or undesirable chemical compounds makes dereplication an imperative step to efficiently run NPs discovery programs. To meet the demand for rapid analytical characterization of biologically active samples identified by HTS, dereplication strategies using tandem analytical techniques and database searching to determine the identity of an active compound at the earliest possible stage in the discovery process have been refined.[52] For example, the LC-NMR technique used in NPs HTS programs, resulted in the identification of the marine alkaloid aaptamine **1** (Fig. 6), isolated from *Aaptos* sp. sponge, exhibiting inhibitory activity $IC_{50} = 120$ nM, against the enzyme glutamine:fructose-6-phosphate amidotransferase (GFAT).[51]

The application of the high-performance liquid chromatography-electrospray ionization MS (HPLC-ESI-MS) technique specifically for the integration of NPs sample mixtures into modern HTS, also reveals noteworthy impact upon several procedures associated with the HTS of NPs, including, among others, assessment of the extract sample diversity, dereplication, structure elucidation, preparative isolation and affinity-based biological activity evaluation.[11] Multiple approaches used in biodiversity- and taxonomy-guided microbial NPs libraries construction, combined with HTS assays plus liquid chromatography-mass spectrometry (LC-MS) have been highlighted as efficient dereplication processes.[12] Overall, emerging advances in MS, MS-MS, NMR and other technologies are making it possible to overcome the challenges encountered in screening NPs libraries in today's drug discovery environment. In fact, the success of any HTS campaign depends on the quality of the chemical library. The construction and maintenance of a high quality NPs library, whether based on microbial, plant, marine or other sources is a costly endeavor.[10] The NPs programs based on screening of extract libraries, bioassay-guided isolation, structure elucidation and subsequent scale-up production are challenged to meet the rapid cycle times that are characteristic of the modern HTS.[10] As we apply these technologies and develop them even further, we can look forward to increased impact of NPs in HTS based drug discovery.[10] Recently various attempts have been made to increase the efficacy and precision of chemical libraries used in HTS drug discovery approaches. One such approach is ChemGPS[144–148] developed at the Backlund lab, which provides a defined chemical space for pre-screening evaluation of chemical compounds properties or virtual dereplication. However, the need for space expansion in ChemGPS for NPs[144] was recently proposed, since several studied NPs to a large extent fell outside the defined ChemGPS chemical space.[148] Continually having the objective of keeping scientific accomplishments moving forward new issues are addressed, a recent review, concerning HTS methodologies for NPs samples, demands alterations in assay design as well as in sample preparation to increase the yield of meaningful hit structures.[8]



1

Fig. 6 Marine alkaloid, aaptamine **1**.

The HTS operation is still highly specialized and expensive. Although some Universities have one of their own, most of the labs resort to using the services of an existing HTS facility.[149]

The process of finding a new drug against a chosen target for a particular disease usually involves HTS, wherein large libraries of chemicals are tested for their ability to modify the target. Another important function of HTS is to show how selective the compounds are for the chosen target. The idea is to find a molecule which will interfere with only the chosen target, but not with other related targets. To this end, other screening runs will be made to see whether the "hits" against the chosen target will interfere with other related targets – this is the process of cross-screening.[150] Cross-screening is important because the more unrelated targets a compound hits, the more likely that off-target toxicity will occur with that compound once it reaches the clinic.[151] While HTS is a commonly used method for novel drug discovery, it is not the only method. It is often possible to start from a molecule which already has some of the desired properties. Such a molecule might be extracted from a natural source or even be a drug on the market which could be improved upon (so-called "me too" drugs). Other methods, such as virtual high throughput screening (VHTS), where screening is done using computer-generated models and attempting to "dock" virtual libraries to a target are also often used.[152,153]

According to Makley and Gestwicki, HTS sequencing is a methodology that will shape the dereplication future and expand the number of "druggable" targets.[154]

# 6 Analytical technologies

Analytical techniques are interconnected in such a way that the separation line is very narrow, which sometimes hampers their division into topics. Nevertheless to make a clear general overview, these will be distributed by subsections, including: (1) separation techniques; (2) detection methods and (3) hyphenated techniques.

## 6.1 Separation techniques

Compound isolation from NPs complex mixtures, such as crude extracts, can be performed using several developed separation techniques. The most usual are; Thin-Layer Chromatography (TLC), Gas Chromatography (GC), Capillary Electrophoresis (CE), Solid Phase Extraction (SPE), Column Chromatography (CC), Flash Column Chromatography (FCC), High Performance Liquid Chromatography (HPLC) and Ultra High Performance Liquid Chromatography (uHPLC).

High-Performance Thin-layer chromatography (HPTLC) is the result of improvements made to the original TLC method, including automation, increasing the resolution attained and enhancing quantitative analysis accuracy. This method is not used as commonly as the others and its use is usually related to

plant NPs.[146,147] GC possesses very high chromatographic resolution. Although only volatile chemical compounds that vaporize without decomposing may be separated and analyzed. Higher molecular weight (MW) and polar metabolites cannot be analyzed by this method.[148] Modern instruments allow '2D' chromatography (GC × GC), promoting an additional resolution increase.[149] CE is suitable for use with a wider range of compound classes than GC and has a higher theoretical separation efficiency than HPLC. As for all electrophoretic techniques, it is most appropriate for charged analytes. For NPs structure elucidation MS and/or NMR are essential.[150] Explorative Solid-Phase Extraction (E-SPE) and HPLC-PDA-MS-SPE-NMR were described in several reports as accelerating microbial and plant NPs discovery, purification and dereplication methodologies, respectively.[68,81,151] Column chromatography, for example CC and FCC, often times is used before using HPLC, the resulting fractions being analyzed by the latest technique. The HPLC method has been extensively used in the isolation of a broad range of NPs, and has become a very powerful, versatile and standardized chromatographic technique. It has the advantage of being able to be coupled with an endless list of detection methods, several of which discussed in further detail in subsection 6.3.[152]

uHPLC systems can work at up to 15 000 Psi. The higher pressures allow the use of much smaller particle sizes in the columns (~0.01 μm). Operating at very high pressures and using such small packing column particles and column sizes enables a remarkable decrease in analysis time, sample amount and eluent volume, increase in peak capacity, sensitivity and reproducibility compared to conventional HPLC. An important improvement of the overall performance was achieved, for numerous applications. This technology has rapidly been widely accepted by the analytical community and is being gradually applied to various fields of NPs analysis such as QC, profiling and fingerprinting, dereplication and metabolomics.[153] Similarly to HPLC this separation method can be used with several column types and chemical phases, as well as being coupled with several detection methods, according to the specifications of the NPs of interest. As an example, using UV detection – photodiode array detector (DAD) and Acquity BEH $C_{18}$ chromatographic column due of its universality, selectivity, efficiency and robustness a fingerprinting method for chemical screening of microbial metabolites, potential antibiotics, in spent cultivation broths was accomplished.[103]

### 6.2 Detection methods

MS is used to identify and to quantify metabolites after separation by GC, LC, CE, HPLC or uHPLC. There are several studies which use MS as a stand-alone technology: the sample is infused directly into the mass spectrometer with no prior separation, the MS serves both to separate and detect metabolites.[3] MS is a highly selective and high throughput analytical technique, which is ideally suited to the identification and purity determination of large numbers of compounds prepared using organic synthesis/combinatorial chemistry or for NPs dereplication. Major improvements in MS hardware and methodologies that are particularly relevant to NPs research and dereplication fields are highlighted in Fig. 7.[15,154]

Compounds may be characterized based on molecular weight, elemental composition and structural features based on fragmentation patterns. Surface-based mass analysis has seen a resurgence in the past decade, with new MS technologies focused on increasing sensitivity, minimizing background, and reducing sample preparation. The ability to analyze metabolites directly from biofluids and tissues continues to challenge current MS technology, largely because of the limits imposed by the complexity of these samples. Among the technologies being developed to address this challenge is the Nanostructure-Initiator MS (NIMS) desorption/ionization approach which does not require the application of a matrix and thereby facilitates small-molecule (*i.e.*, metabolite) identification. MALDI is also commonly used, yet the application of a MALDI matrix can add significant background at <1000 Da that complicates analysis of low-mass range metabolites. In addition, the size of the resulting matrix crystals limits the spatial resolution that can be achieved in tissue imaging. Because of these limitations, several other matrix-free desorption/ionization approaches have been applied to the analysis of biofluids and tissues. Secondary ion mass spectrometry (SIMS) was one of the first matrix-free desorption/ionization approaches used to analyze metabolites from biological samples. SIMS uses a high-energy primary ion beam to desorb and generate secondary ions from a surface. The primary advantage of SIMS is its high spatial resolution, a powerful characteristic for tissue imaging with MS. However, SIMS has yet to be readily applied to the analysis of biofluids and tissues because of its limited sensitivity at >500 Da and analyte fragmentation generated by the high-energy primary ion beam. Desorption electrospray ionization (DESI)[155,156] and Direct Analysis in Real Time (DART)[157] are matrix-free techniques for analyzing biological samples that use a charged solvent spray to desorb ions from a surface. Advantages in both DESI and DART are that no special surface is required and the analysis is performed at ambient pressure with full access to the sample during acquisition.[115] A limitation of DESI is spatial resolution because "focusing" the charged solvent spray is difficult.[115] However, a recent development termed Laser Ablation ESI (LAESI),[158] is a promising approach to circumvent this limitation, followed by nanoDESI upgrade.[159] For more detailed literature related with MS techniques see Carter[115] and Dorrestein and co-workers[15] reviews.

NMR is the only detection technique that allows the analysis of all kinds of small metabolites and in which the samples can thus be recovered for further analyzes. The great majority of NPs chemists consider that NMR is close to being a universal detector for structure elucidation. The main advantages of NMR are high analytical reproducibility and simplicity of sample preparation. In practice it is relatively insensitive compared to mass spectrometry-based techniques. In fact, recent advances in MS enable this method to have has as much structural elucidation potential as NMR. However it is not yet commonly used in this perspective by NPs researchers. While NMR and MS are the most widely used techniques, other methods of detection that have been used include ion-mobility spectrometry,

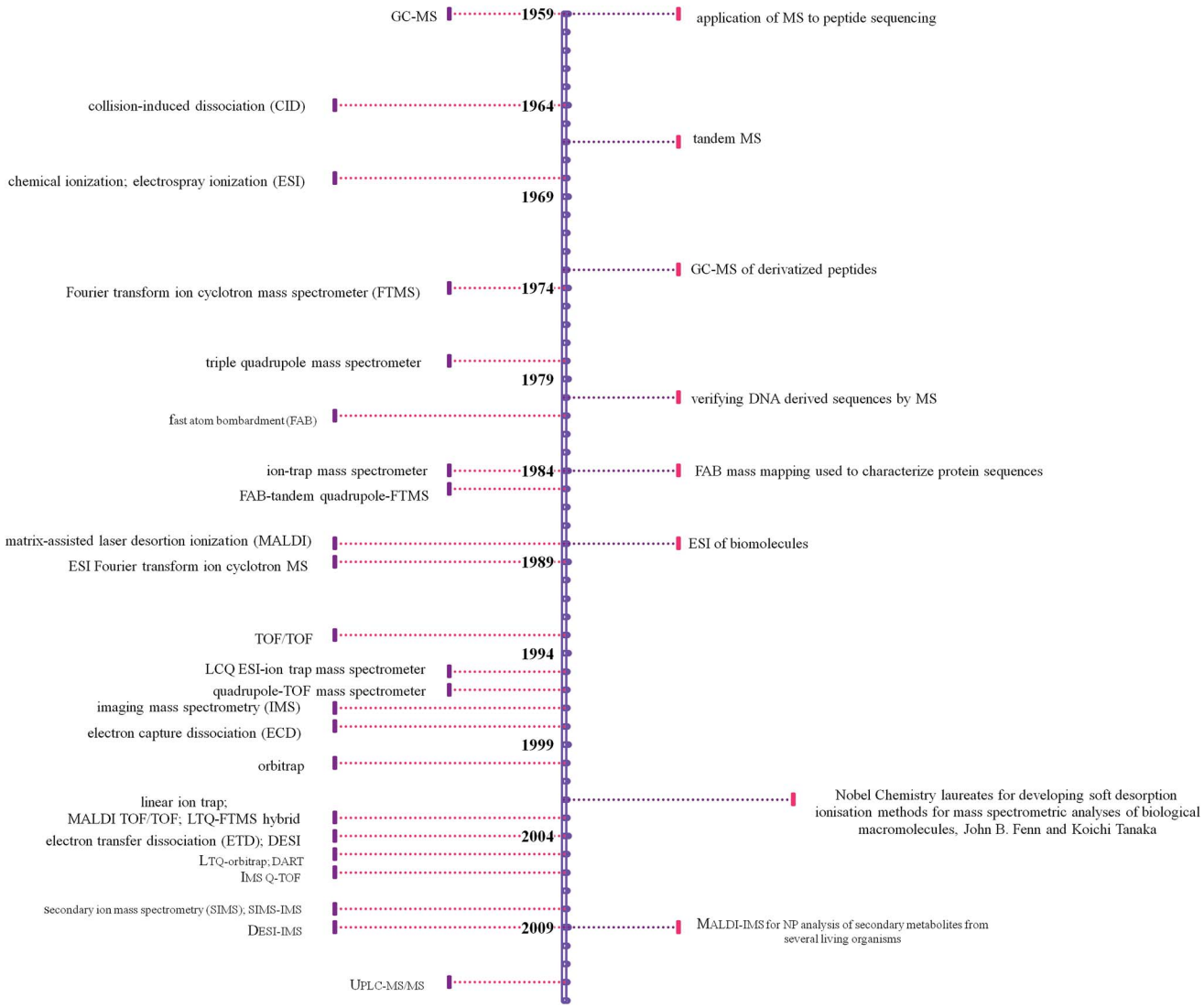



Fig. 7 Timeline illustrating the major advances in MS hardware and methodologies, period 1959–2012.

electrochemical detection (coupled to HPLC), radiolabel (when combined with TLC) and X-ray crystallography. Due to their enormous significance as dereplication methods MS, X-ray single crystal diffraction (SCD) and NMR will be focused in extra detail in Sections 6.3 and 7, 8 and 9, respectively.

### 6.3 Hyphenated techniques

Dereplication strategies rely on analytical techniques and database searching to determine the identity of an active compound at the earliest possible stage in the discovery process. This prevents wasted efforts on samples with no potential for development and allows resources to be focused on the most promising leads. In the past few years, advances in technology have permitted the development of tandem analytical techniques. Hyphenated methods use a combination of two (or more) separation and detection methods. Advances in technology have allowed the development of tandem, modern, sophisticated hyphenated analytical techniques which are commonly used for NPs dereplication achieving rapid lead identification,[4] examples are GC-MS, LC-PDA, LC-MS,[36,160] LC-FTIR, LC-NMR, LC-NMR-MS, and CE-MS.[161] LC-UV-DAD, LC-MS, LC-MS-MS, LC-NMR[146] and LC-NMR-MS. Modern spectroscopic methods have largely revolutionized compound identification and tremendously accelerated the speed at which isolated compounds can be identified.[4] Hyphenated techniques can be used as metabolomics, genomic and proteomics dereplication tools. These "omic" strategies will be further described in Section 12. The previously mentioned tandem methods, that have been reported as successful tools for dereplication purposes, as well as the most recent and exotic hyphenated trends will be described below, in this review. These will give information for NPs researchers to set their own dereplication workflow according to compound sources, characteristics, budget and potential collaborations.

**GC-MS.** GC-MS was the first hyphenated technique to be developed, it is one of the most widely used and potent

methods. GC-MS of derivative components of lipophilic extracts is the first step before any bioassay-guided studies, this technique being the method of choice for dereplication of fatty acids.[149]

**Liquid Chromatography (LC).** Liquid Chromatography (LC) metabolite profiling using hyphenated techniques such as LC-MS and more recently LC-NMR are the most frequently used methods in the NPs community. LC-MS has been used since the early days of dereplication. It quickly provides plenty of structural information, leading to a partial or a complete on-line *de novo* structure determination of the NPs of interest when combined with MS databases. As a complement to this approach, bioassays performed after LC/microfractionation of the extracts grant efficient localization of the bioactive LC-peaks in the chromatograms. The combination of metabolite profiling and LC/bioassays provides the possibility of distinguishing between already known bioactive compounds (*i.e.* dereplication) and new molecules directly in plants, bacteria and fungi crude extracts. Also several examples of rapid localization of bioactive compounds, based on post-chromatographic bioautographic testing of LC-NMR microfractions and subsequent on-line identification were illustrated by Hostettmann and coworkers.[54,58]

The LC-SPE-NMR approach introduced the solid-phase extraction (SPE) interface between the chromatography and NMR by on-line multiple trapping of the constituents to achieve accumulation for generating high quality 1D and 2D NMR spectra. This technique has greatly improved sensitivity and reduced NMR acquisition time. The LC-SPE-NMR method provides a highly valuable and efficient tool for NPs drug discovery research by performing it in an automated manner. The application of an LC-SPE-NMR-MS approach has proven to be the most effective combination for compound identification without traditional purification of individual components in the crude extract. The recent development of cryogenic flow and micro-coil (nL quantity) NMR probes set a dramatic increase in sensitivity to accomplish *de novo* structure elucidation of complex NPs in 10–50 μg quantities, and also makes on-line NMR data acquisition possible. Multiple hyphenation techniques, such as LC-SPE-NMR-MS-FTIR, represent the future direction of a comprehensive and robust method for the rapid dereplication of NPs extracts.[5]

NPs discovery is far from being at the ideal efficiency and productivity and can always be further improved; evolutions in hyphenation techniques are welcome. These might include the combination of an innovative on-line bioassay system, perfectly coupled with automated database searching capabilities (data libraries consist of LC, UV, NMR, MS, IR and other searchable parameters) to accelerate the entire process.[5] This has already been confirmed by several reports on on-line biochemical detection coupled to mass spectrometry (LC-BCD-MS), which has been shown to profoundly accelerate the time required for compound description and identification.[162]

**Liquid chromatography electrospray ionization-mass spectrometry and matrix-assisted laser desorption ionization-time-of-flight-mass spectrometry (LC-ESI-MALDI-TOF).** Liquid chromatography electrospray ionization-mass spectrometry and matrix-assisted laser desorption ionization-time-of-flight-mass spectrometry (LC-ESI-MALDI-TOF) have been revealed to be an improved efficiency approach over more traditional schemes utilizing off-line fraction collection and conventional ionization methods, which can be explained by several factors. First, the superior sensitivity of ESI and MALDI means that less material is required for a successful analysis. Second, on-line LC-MS optimizes the efficiency of sample transfer, saving both time and monotonous labor. Furthermore, the concentration dependence of ESI allows the majority of the LC-MS injected material to be recovered for biological testing without compromising the signal available for molecular weight determination.[163]

**Electrospray ionisation time-of-flight mass spectrometry (ESI-TOF-MS).** Electrospray ionisation time-of-flight mass spectrometry (ESI-TOF-MS) has been used for detection and identification as a highly sensitive and accurate method, it has proved to be very powerful for the analysis and dereplication of NPs in complex mixtures.[59] Hyphenated LC-DAD-SPE-NMR and LC-UV-ESI-MS techniques applied for the separation and structure verification of the major known constituents present in plant extracts revealed them to be worthy dereplication methodologies.[164]

LC-DAD-TOFMS spectral data is becoming easier to use for dereplication with advances in analytical equipment and better compound databases.[83] LC-MS-ELSD (Evaporative Light Scattering Detector) analysis allows the creation of a peak library, which can be used for different data mining strategies: (1) the dereplication of previously isolated NPs; (2) clustering/ranking of extracts for the creation of highly diverse NPs libraries; (3) a selection tool for the focused isolation of bioactive NPs; and (4) to search for alternative sources of a targeted NP. It also has the advantage of showing the predominant compounds for isolation in a complex mixture.[78,90] LC-MS-MS became a very important tool for the on-line identification of NPs in crude extracts. For an efficient use of this technique in NPs dereplication, a careful study of the parameters to generate informative MS-MS spectra is necessary. CID MS-MS spectra of ubiquitous NPs constituents have been systematically studied using hybrid quadrupole time of flight (Q-TOF) and ion trap (IT) mass analyzers under various CID energy conditions. For example, these guidelines and on-line characterization were applied to study C-glycosidic flavonoids by LC-MS-MS or LC-multiple-stage MS.[165]

**HPLC.** HPLC can be coupled to several detection methods, such as UV, DAD, FD, ECD, RID, FID, CL, ESLD, CAD, MS, MS-MS and many others. According to the diversity and specific features of the NPs, it allows optimization for the most efficient detection and isolation conditions in a personalized way for each specific case.[152] HPLC has lower chromatographic resolution than GC, but has the advantage of being able to analyze and/or separate a much wider range of compounds.[166] It also has the advantage of separating both polar and nonpolar compounds, using reverse or normal phase solvents and columns, respectively. High performance liquid chromatography-electrospray ionization mass spectrometry (HPLC-ESI-MS) specifically applied for the integration of NPs sample

mixtures into modern HTS, had significant impact upon a variety of procedures associated with the HTS of NPs, including extract sample diversity evaluation, dereplication, structure elucidation, preparative isolation, and affinity-based biological activity evaluation.[11] This improvement is due to the high resolution provided by reversed-phase HPLC coupled with the moderate and quite universal ionization aided by electrospray method.[11] The hybrid HPLC-DAD-MS-SPE-NMR technique must be highlighted as particularly promising due its versatility and opportune time-saving. After initial HPLC separation with protonated solvents and detection, a small fraction of the solution is sent for MS, with the remainder sent to an SPE cartridge. If the MS results suggest that further investigation is necessary, the stored sample can be conveniently washed off the cartridge with a deuterated solvent into an NMR tube or flow cell for NMR measurements.[68]

**uHPLC.** uHPLC is an improved methodology for the analysis of NPs crude extracts, using new MS analyzers with increased resolving power and accuracy such as the orbital trap (Orbitrap) HR-MS and HR-MS-MS, which hugely facilitates the identification of complex compound matrices.[167] uHPLC-DAD-QTOF and uHPLC-DAD-QTOF-MS screening applied to secondary metabolites of fungi extracts has proved to be an accurate dereplication tool.[105,168] Furthermore, uPLC-MS-ELSD-PDA databases produced during fractionation may be used as powerful dereplication tools to facilitate compound identification from small-molecule NPs libraries.[107] uHPLC-MS-MS, relying on molecular networking, is also an emerging technique to dereplicate related molecules.[95]

**Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS).** Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) performs fast and cheap analyzes for microorganism dereplication. It exhibits promising high-throughput automated analysis, and is consequently becoming a supreme dereplication tool.[88] MALDI-TOF-MS is an efficient dereplication device as it can be used to discriminate between bacterial isolates at the species level. Further studies corroborate that large scale MALDI-TOF-MS based on bacteria taxa identification is appropriate as a dereplication methodology, performing an efficient screening of microbial culture collections containing pigments with potential novel properties.[113]

**MALDI-TOF-imaging (npMALDI-I).** MALDI-TOF-imaging (npMALDI-I) characterizes the spatial distribution of NPs from intact organisms of differing complexities. It has been tested in cyanobacteria and sponges. In addition to identifying known NPs, it determines unknown ions co-localized in the different matrices, proving to be a suitable dereplication method for drug discovery programs.[87] The latest applications of Imaging MS (IMS)[117] to NPs research, technological challenges, prospects and improvements have been reviewed by Yang and co-workers,[14] and Dorrestein and co-workers.[15] Including in the latest, the most recent high-tech advances, such as 3D MALDI-IMS, DESI-IMS and SIMS-IMS[117] describing their various advantages over MALDI-IMS and their downsides as emergent approaches for NPs discovery and for the understanding of molecular interactions regulating and guiding ecology in complex biological systems. A major intention was to resolve the shortcomings presented by the newest mentioned approaches and predict the continuous development of experimental techniques with more efficient ionization sources and more sensitive detectors to refine structure elucidation to the visualization of intact NPs at the subcellular level and mapping NPs at a global level.[15]

**Intact-Cell MALDI-TOF (ICM).** Intact-Cell MALDI-TOF (ICM) mass spectrometry to achieve a rapid proteometric clustering of a subset of a microbial strain collection has been described.[86] In the reported study, cluster analyses of mass spectra resolved microbial strains into 11 groups corresponding to several species belonging to different genera; the results were verified by 16S rDNA analysis. This approach permits the rapid identification of isolates for dereplication, and the selection of strains, that represent rare species for subsequent characterization.[86]

**Collision-induced MS-MS technique.** The collision-induced MS-MS technique is used to fragment a precursor ion into several product ions, and individual product ions are selected and subjected to collision-induced MS-MS-MS analysis. This method enables the identification of the fragmentation pathway of a precursor molecule from its first-generation fragments (MS-MS), through to the $n^{th}$ generation product ions (MS$n$). It also allows the identification of the corresponding neutral products released (neutral losses). Elements used in the molecular formula analysis include C, H, N, O, and S, since most NPs are constituted by these five elements. High-resolution mass separation and accurate mass measurements afforded the unique identification of molecular formulas of small neutral products. Through sequential addition of the molecular formulas of the small neutral products, the molecular formula of the precursor ion and its ion products were uniquely determined. Using a reverse process the molecular formula of the precursor molecule was used to identify or confirm the molecular formulas of the neutral products and their ion products. The molecular formulas of the neutral fragments permitted the identification of substructures, leading to a rapid and efficient characterization of NPs precursors. The method was applied to paclitaxel **2** (Taxol™, $C_{47}H_{51}NO_{14}$; 853 amu), Fig. 8, to identify its molecular formula and substructures, and to characterize its potential fragmentation pathways. The method was further validated by correctly identifying the molecular formula of minocycline ($C_{23}H_{27}N_3O_7$; 457 amu) **3** and piperacillin ($C_{23}H_{27}N_5O_7S$; 517 amu) **4**, Fig. 8.[77]

Thus, a simple and sensitive mass spectrometric method has been developed for the dereplication of NPs. The method provides information about the molecular formula and substructure of a precursor molecule and its fragments, which are invaluable aids in dereplication of NPs at their early stages of purification and characterization.[77]

**Multistage MSn *de novo* sequencing dereplication.** Multistage mass spectrometry (MSn) generates so-called spectral trees, and is a powerful tool in the annotation and structural elucidation of metabolites and is increasingly used in the area of accurate mass LC/MS-based metabolomics to identify unknown, but biologically relevant compounds. As a

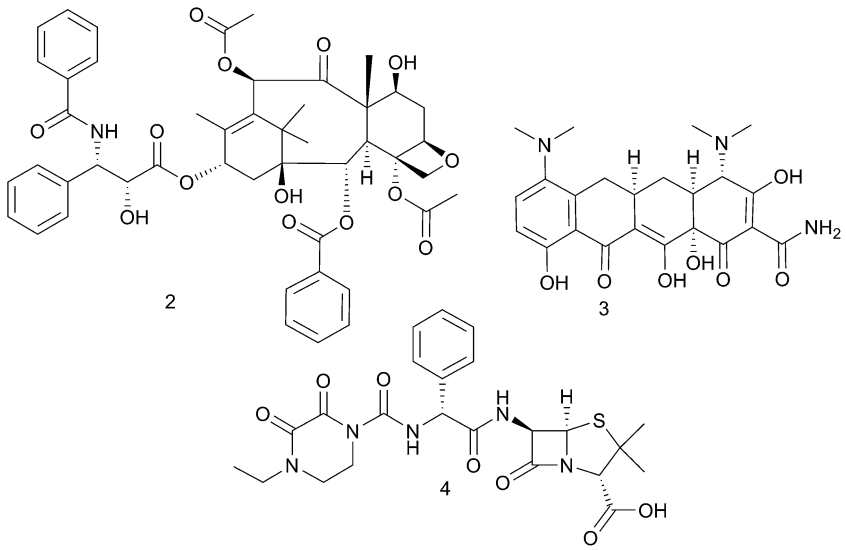



Fig. 8 Chemical structures of paclitaxel 2, minocycline 3 and piperacillin 4.

consequence, there is a growing need for computational tools specifically designed for the processing and interpretation of MSn data. A novel approach to represent and calculate the similarity between high-resolution mass spectral fragmentation trees was created.[112] This approach can be used to query multiple-stage mass spectra in MS spectral libraries. Additionally the method can be used to calculate structure–spectrum correlations and potentially deduce substructures from spectra of unknown compounds. Both the dereplication and *de novo* identification functionalities of the comparison approach were discussed by Reijmers and co-workers.[112] This novel MSn spectral processing and comparison method increases the probability of assigning the correct identity to experimentally obtained fragmentation trees. Ultimately, this tool may pave the way for constructing and populating large MSn spectral libraries that can be used for searching and matching experimental MSn spectra for annotation and structural elucidation of unknown metabolites detected in untargeted metabolomics studies.[112]

**Flow injection electrospray mass spectrometry (FIE-MS).** Flow injection electrospray mass spectrometry (FIE-MS) metabolite fingerprinting is widely used as a first pass screening for compositional differences, where discrimination between samples can be achieved without any preconceptions. Powerful data analysis algorithms can be used to select and rank FIE-MS fingerprint variables of the biological problem under investigation. Species-specific FIE-MS-MSn metabolite database creation and understanding how to query the database to predict identity of highly significant variables within FIE-MS fingerprints are required. Draper and co-workers[101] developed a protocol applicable to any bioscience research area, involving FIE-MS fingerprinting. It details how to interpret *m*/*z* signals within the explanatory variable list based on a correlation analysis in conjunction with an investigation of mathematical relationships regarding (de)protonated molecular ions, salt adducts, neutral losses and dimeric associations routinely observed in FIE-MS fingerprints. Although designed for use by biologists and analytical chemists, data-mining expertise is an added valued.[101] Using multistage mass spectrometry followed by spectral alignment algorithms made possible the development of technology for high-throughput nonribosomal peptide (NRPs) dereplication and sequencing. The algorithm, developed by Pevzner and co-workers, for comparative NRP dereplication establishes similarities between newly isolated and previously identified similar but non-identical NRPs which substantially reduces dereplication efforts. However, the weakness of the method is that it is only suitable for cyclic NRPs.[96] Successful applications of this novel technique combined with NMR for cyanobacteria and actinomycetes NPRs were attained.[169,170] Development of a chemoinformatic library-based and informatics search strategy for NPs (iSNAP) has been constructed and applied to NRPs, and it proved useful for true non-targeted dereplication across a spectrum of NRPs within NPs extracts.[123]

# 7 Computational mass spectrometry tools for dereplication

## 7.1 Ligand-guided approach – *small molecules*

Presently, the chemical structures of many thousands of NPs are known, *e.g.* the Chapman & Hall/CRC Dictionary of NPs contains a comprehensive database of 170 000 NPs, although the vast majority of metabolites still remain unknown.[171] Moreover, the structural diversity of metabolites is extraordinarily large, when compared with biopolymers, such as proteins. Identification of secondary metabolites poses a problem, unlike proteins these small molecules are usually not made from building blocks, and the genomic sequence does not reveal, in almost all cases, information about their structure. Consequently, a huge number of metabolites remain uncharacterized with respect to their structure and function.[172] The identification of small molecules from MS data continues to be

a major data interpretation challenge. Computational aspects for identifying small molecules range from searching for a reference spectral library to structural elucidation of an unknown compound. Single-stage MS does not provide information beyond the compound molecular mass, in order to obtain such information the analyte must be fragmented, frequently using fragmentation methods such as collision-induced dissociation (CID) for tandem MS, and fragmentation during electron ionization (EI). Therefore the most common dereplication approach using MS is to search for similar fragmentation spectra in a library.[173] The aim of library searching is either to obtain a correct structure hit of an already known compound or partial structural insights from novel compounds that nearly match. Unfortunately, the size of public and commercial available MS-MS libraries is still small compared with electron ionization libraries. Thus, the searching in MS-MS libraries is often unsuccessful.

After some initial progress as part of the DENDRAL project[174] throughout the 1970s and subsequent decades, not much progress has been accomplished regarding the development of computational methods for analyzing fragmentation patterns of small molecules, besides spectral libraries searching. In a recent review, Böcker *et al.*[175] reported several new ideas and approaches dealing with MS detection and investigation of small molecules that have surfaced over the last five years, relying on established computational methods such as combinatorial optimization (*MetFrag*,[176] *MetFusion*,[177] and *FT-BLAST*[178]) and machine learning (*ISIS*[179] and *FingerID*[180]) techniques. The crucial leap forward in drug discovery from natural sources can only be achieved, in our opinion, with an essential change in dereplication MS methodology, including predicted spectral data. In this sense four (MetFrag, MetFusion, ISIS and FingerID) out of the five above referenced methods, are not supported by spectral library searching, but instead they rely on more comprehensive molecular structure database searching. In fact, spectral libraries are obviously several orders of magnitude smaller than molecular structure databases. For example, the PubChem database currently contains about 50 million compounds, while the two largest (commercial) spectral libraries, National Institute of Standards and Technology-NIST (version 14) and Wiley Registry ($10^{th}$ edition) enclose mass spectra data (MS) of 250 000 and 680 000 compounds, respectively. Currently, NIST comprises 51 216 MS-MS spectra from 42 126 different precursor ions out of merely 8171 compounds. This limitation can be overcome by an accurate prediction of fragments and their respective abundances from compounds' molecular structures using computational methods. Consequently, searching in spectral libraries can be replaced and/or complemented by searching in an *in silico* mass spectra database. This line of attack has been very successfully used in proteomics for many years. Although it is necessary to realize that the prediction of peptide fragmentation patterns are comparatively easier than for small molecules. The ISIS[179] and FingerID[180] tools deal with metabolite fragmentation data, using a rule-based *in silico* fragmentation spectra prediction approach and a predicting structural features and compound classes approach, respectively. The DENDRAL project[174] was the first attempt to generate structural candidates and predict their fragmentation mass spectra using a rule-based approach. However, it failed in its major purpose of performing automatic structure elucidation using mass spectral data, and the research was discontinued.[181] Unlike the rules learned during the DENDRAL project, Kangas *et al.*[179] did not claim these predictions, which were obtained by simulating the behavior of ions in a linear ion trap using a kinetic Monte Carlo simulation, to achieve true fragmentation rules. Their ISIS tool built using an Artificial Neural Network machine learning technique worked well but only for lipids, 40 out of 45 lipids of the test set were correctly identified.[179] On the other hand, in the fingerprint approach the query spectrum of an unknown compound is transformed into a vector feature that was given to the substructure classifiers to predict the fingerprint of the molecule, using the same training data transformations. Heinonen *et al.*[180] used the predicted fingerprints from targeted LC-MS and CID fragmentations carrying out a Kernel-based approach, to retrieve and score candidate molecules from large molecular databases, such as PubChem.

The original ideas associated with mapping fragmentation spectra with compound structures (combinatorial optimization) were found in the literature in 1980,[182] but it was necessary to wait over 30 years for the development of a suitable combinatorial fragmentation approach to process a complete database.[175–177] Whereas *MetFrag*[176] compares *in silico* mass spectra obtained by a bond dissociation approach, with experimental mass spectra from PubChem or Kyoto Encyclopedia of Genes and Genomes (KEGG) assigning a score to all results, MetFusion[177] combines search results from a molecular structure database and from a spectral library, taking advantage of both resources. Although, MetFrag achieves promising results, it was clearly outperformed by its successor MetFusion. Another method, FT-BLAST is based on the calculation of fragmentation trees (FTs) and FT alignments. FT-BLAST searches in a FT library comprising the basic idea that fragmentation pattern similarities are correlated with the chemical similarity of the corresponding compounds.

All the discussed approaches are heavily focused on comprehensive molecular structure databases but these data can be exponentially amplified by employing fully comprehensive molecular structures generated by computational approaches to aid in chemical space exploration – Small Molecule Universe Database.[183–186] For example, Kerber *et al.*[187] reported that there are more than 109 million possible molecular structures for the molecular formula $C_8H_6N_2O$ with mass 146 Da, but only 413 hits matched in PubChem database.

### 7.2 Genome-guided approach

Tens of thousands of sequenced microbial genomes or drafts of genomes are available from the International Nucleotide Sequence Database Collaboration (INSDC) database, and this number is predicted to exponentially increase over the next decades. The huge sequence space that has been built can be used for the discovery of small bioactive molecules through a genome mining process.[188–192] In spite of the high rate by which

genome sequences are being obtained, the process of mining genetically encoded small molecules is performed one gene cluster at a time and requires many person-years efforts to annotate a single molecule, making computational approaches to automatically connect a molecule to its biosynthetic signature valuable tools. A few *in silico* methods have been published thus far to automate the analysis of secondary metabolism in bacterial and fungal genomes.[193–201] The first of these was ClustScan[193] (Cluster Scanner) which is designed for rapid, semi-automatic annotation of DNA sequences encoding modular biosynthetic enzymes *e.g.*, polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS) and hybrid (PKS/NRPS) enzymes in microbes, and invertebrate metagenomics datasets. Additionally, more tools were published such as SBSPKS toolbox[194] and NPs searcher web server.[195] Unfortunately, these tools are largely limited to analyzing the core genes for type I PKS and NRPS biosynthesis. Recently, a more comprehensive pipeline capable of identifying biosynthetic loci covering the whole range of known secondary metabolite classes (polyketides, non-ribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, beta-lactams, butyrolactones, siderophores, melanins and others), the antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) was technologically advanced by Takano and co-workers.[201]

In conjunction with these *in silico* methods, developed in the last years, a unique and impressive MS-based strategy that enables the genome mining of small-molecule families from living microbial populations has been recently reported.[159,202,203] In 2012, Dorrestein *et al.*[159] described a powerful integration of two methodologies, nanospray desorption electrospray ionization (nanoDESI) MS and the generation of molecular networks, which, together, allow the direct chemical analysis of secreted microbial exchange factors in live colonies. This technology was validated when metabolic profiling a *Pseudomonas* sp. strain by the detection and partial characterization of thanamycin, a chlorinated non-ribosomal peptide synthetase-derived with antifungal activity. The results are in accordance with previous predictions which attributed the antifungal activity of *Pseudomonas* sp. to the presence of thanamycin.[204] Moreover, the concepts of molecular families (MFs) and gene cluster families (GCFs) were introduced by the same group.[202] They used MS-MS networking as a tool to map the molecular network of more than 60 organisms, most of which unsequenced, and located their non-ribosomal peptide MFs, that were structurally related molecules, based on their mass spectral fragmentation patterns. Recently, the same group identified at least four non-ribosomal peptide synthetase-derived molecular families *e.g.* napsamycin **5**, arylomycin **6**, daptomycin **7**, stenothricin **8** (Fig. 9) and their gene subnetworks with different modes of action from *Streptomyces roseosporus*.[203] Through MS-MS mapping networking, a number of previously unreported analogs that were produced by *S. roseosporus* involving truncation, glycosylation, hydrolysis and biosynthetic intermediates and/or shunt products were captured and visualized.

The MS-MS molecular network approach tackles the problem from new and different directions, emerging as a powerful instrument for dereplication of compounds originating from natural sources and identification of novel drug leads.[95] We agree with the authors, who described these approaches as a step forward to achieve the "holy grail" in microbiology,[159] an ideal methodology integrating governing chemistry with genomics and phenotypes of microbial colonies.

# 8 X-ray crystallography

X-ray crystallography enables the identification of the atomic and molecular structure of a crystal. For comparison, the nearest competing methods in terms of structures analyzed are NMR spectroscopy and mass spectrometry, which provide limited molecular structure information.[205] Moreover, crystallography can solve structures of large molecules, whereas solution-state NMR is restricted to relatively small ones (less than 70 kDa). X-ray crystallography is now used routinely by scientists to determine how a pharmaceutical drug interacts with its protein target and what changes might improve it. However, intrinsic membrane proteins remain challenging to crystallize because they require detergents or other means to solubilize them in isolation, and such detergents often interfere with crystallization. Such membrane proteins are a large component of the genome and include many proteins of great physiological importance, such as ion channels and receptors. The determination of the absolute configuration of chiral compounds is one of the most difficult analyses of molecular structures. NMR and spectrometric methods can determine in principle only relative stereochemistry. X-ray crystallography is the only method that can determine the absolute configuration of chiral molecules, on the basis of the anomalous scattering effects of heavy atoms. The X-ray technique provides direct structural information on molecules at the atomic level and is recognized as a reliable structure determination method.[206,207] However, as its name implies, the technique has a limitation, the sample needs to be available as a single crystal, the growth of which can be a time consuming process of trial-and-error, and is often not possible. Recently, fantastic advances to this method have been accomplished, in 2013 Fujita and co-workers reported a novel X-ray protocol for single-crystal diffraction (SCD) analysis that does not require sample crystallization. In this method, tiny crystals of porous complexes are soaked in a solution of the target, such that the complexes can absorb the targeted molecules avoiding crystallization of the sample itself.[208] The real and intrinsic problems of X-ray crystallography are thus solved and transformed into a rapid and convenient method for the analysis of molecular structures using only a trace amount of sample. The following features are worthy of special mention: (1) the crystallization step, which is the bottleneck of the X-ray analysis protocol, becomes unnecessary. Therefore, the crystallographic study of molecular structures is drastically accelerated and is now applicable to the analysis of liquid or even volatile compounds; (2) crystallographic analysis can be performed on trace amounts, on the nanogram–microgram scale. Thus, in terms of sensitivity, X-ray analysis overwhelmingly dominates NMR analysis and is even comparable to mass spectrometry; and (3) the determination of the absolute

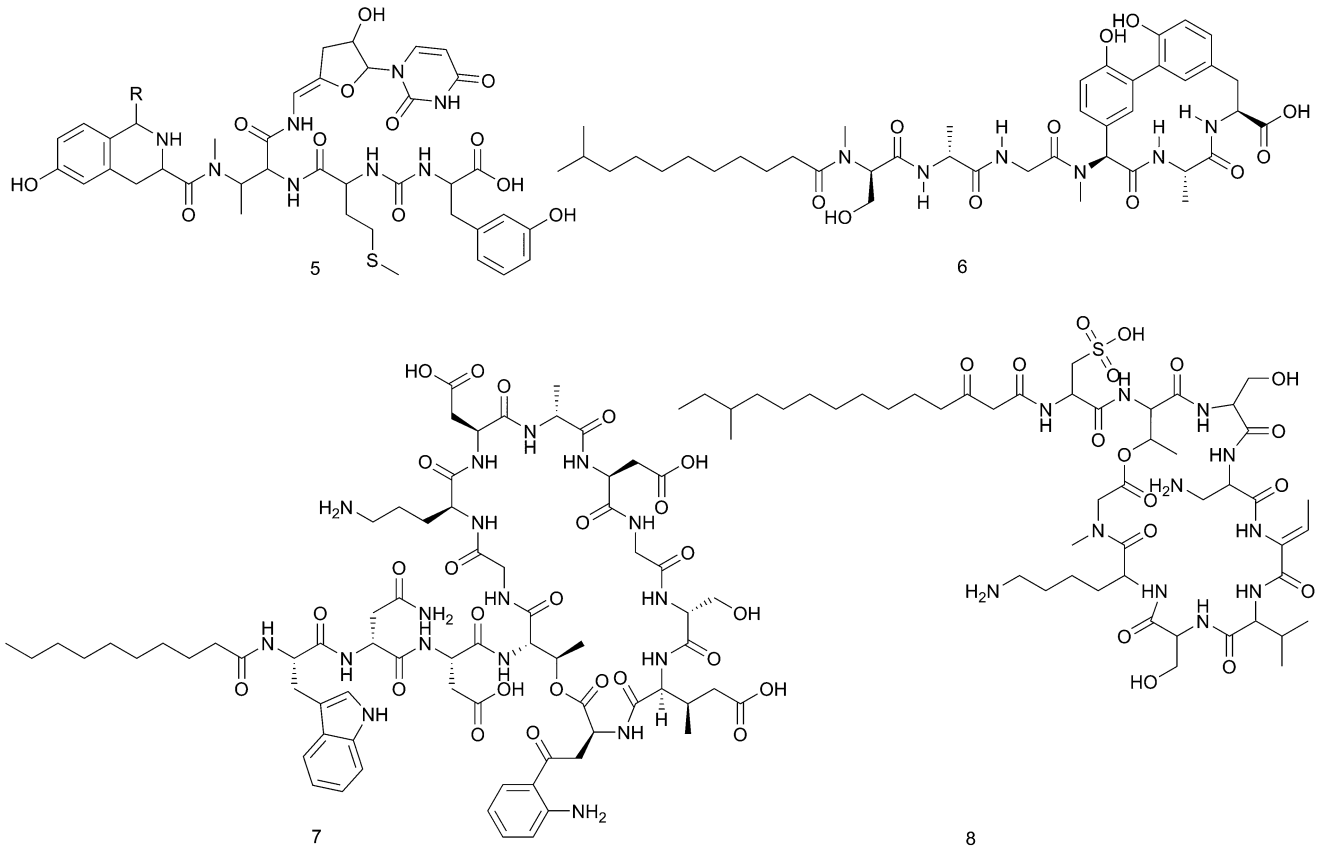


Fig. 9 Four non-ribosomal peptide synthetase derivatives **5–8**, from *Streptomyces roseosporus*.

configuration of chiral molecules can be easily carried out without any chemical modification of the sample molecules. Moreover, it was possible to successfully determine the structure of a scarce marine natural product, miyakosyne A **9**, isolated from a marine sponge *Petrosia* sp., including the absolute configuration of its chiral center, which could not be determined by conventional chemical and spectroscopic methods, Fig. 10.[208]

In combination with HPLC (LC-SCD), this new protocol allows the direct characterization of multiple fractions, establishing a prototypical means of liquid chromatography SCD analysis. It is expected that this methodology will be applicable to microanalysis in NPs chemistry and we predict that many NPs that researchers have given up hope and left behind will be easily and precisely characterized and their structures determined.[208] This technological advent brings added value to several industrial and commercial applications. It has the power to totally shape the future of compound structure elucidation research, and it will open up a new era for drug discovery when routinely applied. Although the authors consider the applicable range of their X-ray protocol, the LC-SCD analysis will be a powerful tool for the rapid characterization of multiple components with much higher structural reliability than LC-MS and LC-NMR techniques, having a considerably extended scope that can include polycyclic, non-aromatic and non-planar molecules. This revolutionary method works perfectly well for several compound types, although there have been some drawbacks in which some molecular characterizations were initially flawed because of atom miss-assignment, symmetry problems and guest disorder.[208] Nevertheless, this problem can be easily fixed and the incorrect structures can be correctly elucidated using only the mass spectrometric data (molecular weight information). Like common crystallographic analysis, the refined crystal structures often times have to be supported by MS and NMR spectroscopy. As a consequence, these techniques may not be discarded with the advent of LC-SCD technique and still need to be used together.

# 9 NMR

A long way has been travelled by NMR spectroscopy from a little used technique by NPs chemists 50 years ago, to currently being an extremely potent and very extensively used method for structure elucidation. Over the last decades several thousands of publications describing the use of 2D NMR to identify and

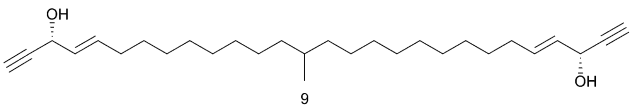


Fig. 10 Miyakosyne A **9**, isolated from a marine sponge *Petrosia* sp.

characterize NPs have been reported. During this period of time, the amount of sample needed for elucidation purposes has decreased from the 20–50 mg range to less than 1 mg. Therefore, it is a very important accomplishment for NPs research, considering that typically it involves small amounts of newly isolated compounds. Major improvements in NMR hardware and methodologies, which are particularly relevant to NPs research and dereplication fields are highlighted in Fig. 11.[18,209,210]

As high- and ultra-high-field NMR instrumentation (400–1000 MHz) becomes increasingly available, the quantitative $^{1}$H NMR (qHNMR) method has been converted into a valuable and unbiased analytical tool for NPs analysis, particularly focused on bioactive NPs, covering all small molecules <2000 Da.[211] The interest in this method within the dereplication context is supported mainly by two approaches: (1) qHNMR-based purity–activity relationships (PARs),[212] a powerful tool for recognizing "hidden" mechanisms that might involve single chemical entities and their interactions with (residually) complex natural matrices using non chromatographic methodologies; and (2) qNMR as an alternative to LC-based quantitation of NPs. The qHNMR experiments based on the most recent development of microcryoprobes (1.7 mm i.d. and smaller) present lower limits of detection and quantification from microgram to nanogram ranges.[211,213–217]

Nevertheless, major sensitivity improvements in modern spectrometers would still obviously be convenient in order to improve NPs dereplication, reducing the time required for data acquisition, particularly for multidimensional NMR of small molecules. Herein, below we discuss two relatively novel developments, which appear to be particularly promising for the NPs investigation field. One of these is the ASAP-HMQC approach (acceleration by sharing adjacent polarization),[218] which uses homonuclear Hartmann–Hahn mixing (similar to that used in TOCSY sequences) during the relaxation delay. It transfers magnetization from protons bonded to $^{12}$C to those bonded to $^{13}$C, enhancing $^{13}$C–$^{1}$H peaks that consequently allow a shorter relaxation delay. Even with an inferior resolution than HSQC, the shorter experiment time makes ASAP-HMQC a potentially useful technique for rapid compound screening and dereplication.[18,218] A second approach corresponds to a non-uniform sampling (or sparse sampling)[18] methodology, which consists in replacing the regular time increments used during the evolution period of a 2D experiment by irregularly spaced intervals. In a recent work, Rovnyak and co-workers[219] reported that a non-uniformly sampled HSQC spectrum of a NP could be obtained in one quarter of the time required to obtain a regular HSQC spectrum with no loss in resolution, using non uniform sampling.

The complete structure elucidation of complex NPs by analysis of 2D NMR data is subject to a fundamental limitation, which is a sufficient number of assignable $^{1}$H signals relative to $^{13}$C or $^{15}$N nuclei of the underlying molecular structure.[220]

Taking into account the so-called *Crews rule*[221] (a guideline for successful 2D NMR analysis) which states that for an easier chemical structure elucidation of a given molecule the ratio of H/C must be greater than 2, otherwise it can be a backbreaking or an inaccurate process (*e.g.*, polycyclic alkaloids with high heteroatom content and many sp$^2$ carbons). For NPs with the right H/C ratio (*i.e.* $\geq$2), the dipolar coupling (1D and 2D NOESY and ROESY) and $^{1}$H–$^{1}$H scalar coupling are great methods for identifying pairs of protons which are spatially close, even if separated by a large number of bonds. Thus, the NOESY sequence, is widely exploited for solving the relative configuration of small molecules.[220] Though, it presents limitations for large molecules such as proteins.[18] Moreover, NOE analysis is an excellent method for conformation and configuration assignments of cyclic compounds as seen in numerous examples of chemical structure elucidation studies of NPs. However, for acyclic compounds with highly flexible carbon chains and multiple conformers, it was necessary to develop innovative and imaginative methods for connecting isolated "islands of stereochemistry"[220] within complex molecules. In fact, it is one of the extreme difficulties in NPs structure elucidation. One creative solution, the J-Based configurational analysis (Murata's method),[222] was developed in 1999, which exploits both $^{1}$H–$^{1}$H and $^{1}$H–$^{13}$C coupling constants in order to assign anti or gauche relationships of vicinal substituted chains. Another solution developed by Kishi *et al.*[223] relies on the observation of several examples of configurational assignments in complex polyketides prepared by synthesis. They realized that small systematic patterns of $^{1}$H NMR and $^{13}$C NMR chemical shift differences are associated with different diastereomers. Furthermore, expanding on this observation, a *universal database* (UDB) to assign the relative absolute configuration of contiguous stereogenic units of complex polyketides was set up.[223–225] Later an extension of UDB was reported using overlapping contiguous triads of $^{1}$H–$^{1}$H coupling constants with those of synthetic diastereomers with defined configuration for assignment of polyol and polyacetoxy compounds purposes.[226,227] Application of the UDB approach was illustrated,[228,229] in the configurational assignment of sagittamides A, **10** and B, **11** (Fig. 12), polyacetoxy long chain α,ω-dicarboxylic acids terminated as amides of ornithine and valine, that were previously reported by Lievens and Molinski from an unidentified tunicate collected in Micronesia.[230]

Diffusion-ordered spectroscopy (DOSY)[231–240] is a family of NMR experiments used in mixture analysis to allow signals belonging to a given organic compound to be correlated through their rate of diffusion. Molecules of different sizes and shapes will often present different diffusion coefficients. Gerwick *et al.*[241] reported a diffusion-edited NMR approach using the DECODES (homonuclear $^{1}$H–$^{1}$H TOCSY or $^{1}$H–$^{1}$H COSY spectra)[241,242] and HETDECODES (heteronuclear $^{1}$H–$^{13}$C HMBC or $^{1}$H–$^{13}$C HSQC spectra)[241,243] experiments, to confirm whether the major compound in a given biologically active fraction of a natural extract was in fact symplostatin 1, a known cytotoxic peptide derivative from the dolastatin class. One of the major glitches of DOSY approaches is spectral overlap. Although the extension of the 2D DOSY experiment to 3D DOSY (*e.g.*, using HMQC or COSY) contributed significantly to reducing overlapping, further advances are needed to resolve complex mixtures. Numerous processing techniques have been developed over the past years to enable resolution of the diffusion

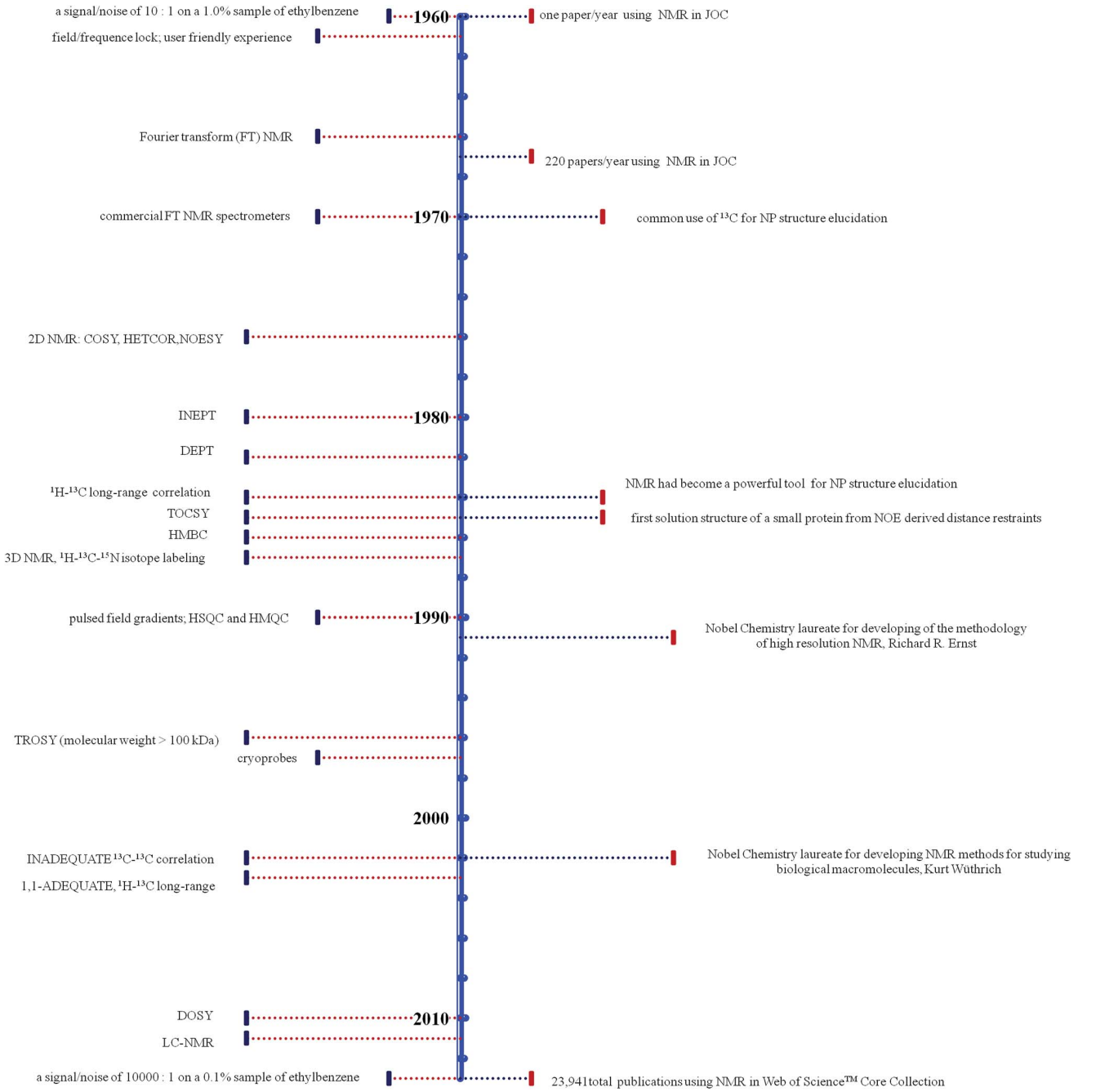



Fig. 11 Timeline illustrating the major advances in NMR hardware and methodologies, period 1960–2013.

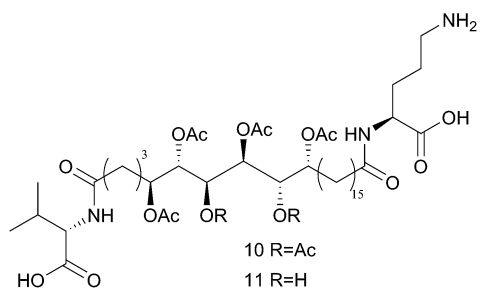



Fig. 12 Sagittamides A **10** and B **11** chemical structures.

dimension that include fitting signal decays to a sum of exponentials,[244] continuous distributions[245] or iterative thresholding.[246] Even when these methods were applied to the very best quality experimental data, it was still required that the components in a mixture were well-separated as regards diffusion coefficient (*i.e.* diffusion coefficient difference > 30%) and of limited number (2–4 components).[247] Recently, new methods were reported to avoid those limitations, specifically LOCOSY[248] an influential method that allows the resolution of a considerably larger number of mixture components, and OUTSCORE[247] showing a cleaner spectra resolution of two

components with diffusion coefficients differing by only 17%. Additionally, an interesting approach used to improve resolution is matrix-assisted DOSY, in which a mixture of compounds with similar diffusion constants (*e.g.* isomers) is manipulated by adding a co-solute (*e.g.* sodium dodecyl sulphate micelles)[249] or a solid phase (*e.g.* stationary phase material used for chromatographic columns).[250] Undoubtedly DOSY approaches added new tools to the "bag of tricks" available for dereplication of known or nuisance compounds in NPs drug discovery processes. Moreover, they are currently recognized as valuable complements to LC-MS for dereplication purposes.[241]

## 10 Computer assisted structure elucidation (CASE)

CASE is by its nature a very complex process, which cannot afford to ignore any available information that can possibly be used for solving the structure of an unknown compound. Therefore, such a system needs to integrate all existing computational methods *e.g.* spectroscopy databases, knowledge and rule collections as well as deterministic and stochastic structure generators, carefully choosing the right method for the structure elucidation problem in question. There are many different ways to approach a particular CASE issue; the common steps of a prototype CASE process are outlined in Fig. 13.[22,251]

The CASE field in regard to NPs starts with a dereplication procedure supported by structure-spectra databases.[252–254] Therefore, spectroscopic databases (*e.g.* 1D $^{1}$H and $^{13}$C NMR, infrared spectroscopy-IR and MS) are the first selected methods when confronted with a structure elucidation question and can thus be seen as some kind of fingerprint method for fast identification of a chemical compound (recorded within minutes if the compound is available in the database). Since early attempts to create databases from all types of spectroscopic data, the largest efforts have been made for NMR databases.[23] Examples of NMR databases are SpecInfo,[252] CSearch,[253] NMRShiftDB,[66] and ACD/Labs NMR (http://www.acdlabs.com/products/dbs/nmr_db/).

NMRShiftDB is an open-access, open-submission database of organic compounds and their NMR data. NMRShiftDB is available to the public *via* a web-interface at http://www.nmrshiftdb.org and through an alpha-quality standalone client. This database contains about 42 000 organic molecules and their associated recorded experimental 1D NMR spectra can be used for searching and spectrum prediction. NMRShiftDB's functionality includes (sub-) spectra and (sub-) structure searches as well as shift prediction of $^{13}$C spectra based on the database support.

ACD/Labs NMR is a commercial database containing more than 1 400 000 experimental $^{1}$H chemical shifts and $^{1}$H–$^{1}$H 450 000 coupling constants.[255] Databases are available for $^{1}$H, $^{13}$C, $^{15}$N, $^{19}$F, and $^{31}$P corresponding to over 210 000; 200 000; 16 780; 9200; and 27 000 chemical structures, respectively. Each standalone database contains chemical shifts, original literature references, molecular formula, molecular weight, and the IUPAC name for each structure record.

Historically, the CASE area has been researched for over 40 years. One of the most established CASE systems, SESAMI,[256] was developed by Munk and co-workers. In SESAMI, two distinct lines of structure elucidation approaches were used, one based on the principle of structure assembly, the other based on structure reduction. Later, a structure generator with a user-friendly interface, termed Assemble 2.0,[243] was created by this group. The generated structures can be ranked automatically according to the agreement of their predicted proton or carbon NMR shifts with the experimental spectrum of the unknown compound. In addition, Assemble 2.0 as a pure structure generator only possesses the most basic knowledge about organic chemistry and does not make any attempt to perform spectra interpretations.[257] In 1988, COCOA[258] a structure generator based on structure reduction was also incorporated into SESAMI. The same group using an independent method from the above mentioned approach, developed structure reduction based methods that were implemented in the SESAMI and COCOA programs.[251] These methods were complemented by spectra interpreters such as INTERPRET and INFER2D.[258,259] More recently, Munk and co-workers reported the improvement of a new structure generator HOUDINI[260] with significantly enhanced performance when compared to the previously used COCOA program. HOUDINI[260] is based on two central data structures: (1) a square matrix of atoms constructed upon input of the molecular formula; and (2) a data structure called substructure representation, which consists of substructures in the form of atom-centred fragments. The performance of COCOA-based and HOUDINI-based SESAMI were compared using a set of seven complex naturally occurring compounds, as a test set of unknowns (between 16 and 76 heavy atoms, non-hydrogen atoms).[261] These comparative tests clearly revealed faster execution times and more efficient processing of ambiguous structural information for HOUDINI.[261] Another approach developed by Köck *et al.*[262] using the CASE system COCON was focused on the integration of new NMR experiments (*e.g.* $^{1}$H–$^{15}$N-HMBC and 1,1-ADQUATE) to overcome known problems in automated structure elucidation. Therefore, COCON has proven to solve CASE technical hitches for molecules as large as the macrolide ascomycin **12** (Fig. 14), using both 1,1-ADEQUATE and $^{1}$H–$^{15}$N-HMBC data.[263] A web version of the program, called WebCocon, is available at http://cocon.nmr.de.

As highlighted by Steinbeck[23] in an excellent NPR review, the *Structure Elucidator* expert system[64,65,264] *StrucEluc* was the first commercial system (Canadian company ACDLabs) with general applicability that presented the most promising achievements in terms of practical application of a CASE system.[265,266] Elyashberg and co-workers, in recent years, have described two generations of the *StrucEluc*: (1) the first generation system, StrucEluc-1 (ref. 267) enabling structure elucidation of organic molecules with 1D $^{13}$C NMR spectra; and (2) the second system, StrucEluc-2,[64,65,268] which is capable of elucidating the chemical structure of large NPs (to date, systems up to 1515 amu mass and 106 skeleton atoms)[98] with 2D NMR spectral data. Several examples have been reported in the literature documenting the successful

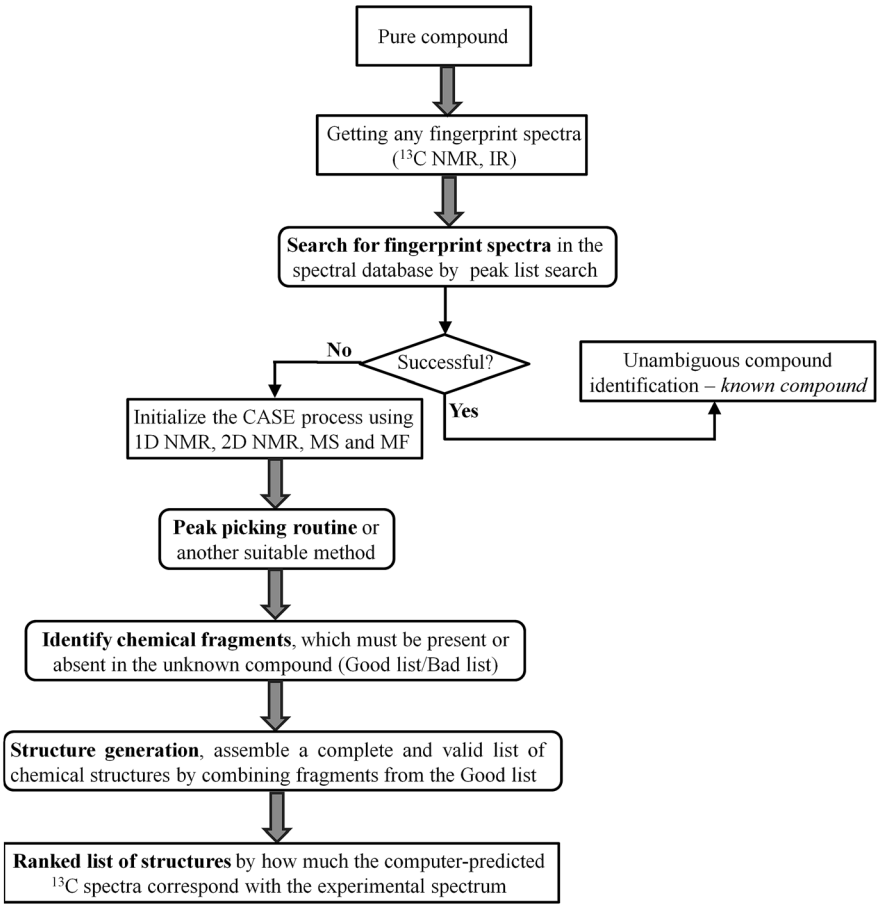
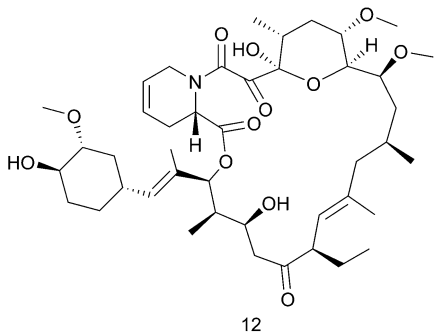



Fig. 13 Components of a CASE process prototype.

application of *StrucEluc* for elucidation of complex NPs as well as for structure revision purposes.[21,269] The NP elucidated structure asperjinone **13**, isolated from thermophilic *Aspergillus terreus*, proposed by Liao *et al.*,[270] was recently revised using the expert system *StrucEluc* and it was suggested that structure **14** is the correct structure (see Fig. 15).[271] The authors stated that it was the first example of a reliable structure revision being performed with the assistance of CASE system only, without additional experiments and quantum chemical NMR shift calculations.[271]



12

Fig. 14 Chemical structure of macrolide ascomycin **12**.

Steinbeck and co-workers suggested and developed various components for CASE over the last years.[254,272] The efficient platform for end users, named SENECA[273,274] was created by them to integrate those CASE systems. The current version of SENECA incorporates the evolutionary algorithm and is available as a GUI client or as a stand-alone command-line executable for free download under artistic license from SourceForge at http://sourceforge.net/projects/seneca/. This version of SENECA was completely refactored and is currently in the Chemistry Development Kit (CDK), a notorious open-source library for chemoinformatics.[275,276]

SENECA performs a stochastic search, using an evolutionary algorithm of constitution space (space made up with all chemical compounds with the same molecular formula), guided by fitness or scoring functions. Two fitness evaluators, NMRShiftDBjudge[273] and AntiBredtJudge,[277] have already been presented in previous versions of SENECA. In addition, a third fitness evaluator, NPLikenessJudge, was incorporated within the last version of SENECA.[274] To test the performance of the SENECA system to predict the correct structures using only $^{13}$C spectral data, Jayaseelan and Steinbeck collected 41 test cases (with heavy atom counts $\leq 15$) from recently published articles in *J. Nat. Prod.*[274] The best performance of the CASE system was revealed with the application of the three fitness evaluators,

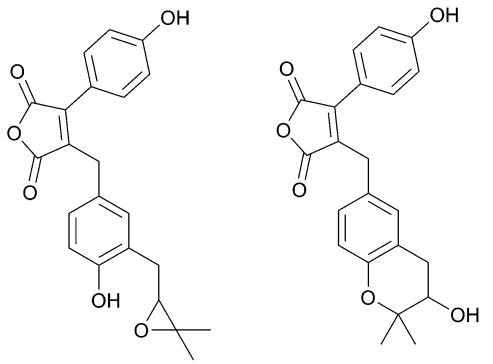



Fig. 15 Proposed structure of asperjinone **13** and the corresponding revised structure **14**.

correct structures were retrieved in the solution set for 36 out of 41 cases. Therefore the authors concluded that natural product-likeness can contribute to a better ranking of correct structure result lists for solving unknown NPs structures.[274]

In addition to CASE systems, there are a number of spectral databases and programs to calculate chemical shifts from structures, which are potentially useful for the NPs dereplication process. For example, the two commercial softwares: ACD/CNMR (Canadian company ACDLabs) and ChemNMR (CambridgeSoft) and the two open-source softwares: NMRShiftDB[66] and SPINUS (http://joao.airesdesousa.com/spinus).[278]

In the last decade the CASE research field has been completely consolidated. We highlighted particularly the recent developments in the SENECA platform which can lead to significant improvements in CASE systems for NPs dereplication applications. In addition, we believe that open-source and open-data implementation strategies surveyed by research groups in this field can make possible, within a few years, the use of CASE systems at NPs laboratories on a daily basis.

## 11 NPs databases

At the present time, databases and the manipulation of databases (data mining) are standard features of chemistry research. Only a fraction of these large knowledge databases are immediately applicable in the NPs field. Several public domain, private domain and commercial databases have been developed, that can assist NPs chemists in the dereplication process. The most relevant databases for NPs dereplication as well as their searchable attributes are listed in Table 1.

Chemical Abstracts Service's Registry File, available at Scientific and Technical Network (http://www.cas.org/products/scifinder) a commercial database, comprises the largest online repository of NPs structures. Other commercially available databases are: NaprAlert (http://www.napralert.org/), which represents a significant resource for the terrestrial sources NPs chemists, Chapman & Hall/CRC Dictionary NPs (http://dnp.chemnetbase.com/tour/) and MNPs (http://www.crcpress.com/product/isbn/9780849382161), MarinLit (http://pubs.rsc.org/marinlit), AntiBase (http://eu.wiley.com/WileyCDA/WileyTitle/productCd-3527338411.html) and AntiMarin.

The remaining databases listed in Table 1 are from public domain (freely available for consultation without fees). These are CSLS (http://cactus.nci.nih.gov/), ChemSpider (http://www.chemspider.com), PubChem (http://pubchem.ncbi.nlm.nih.gov/), ZINC (http://zinc.docking.org/), and spectral databases NAPROC-13 (http://c13.usal.es/c13/usuario/views/inicio.jsp?lang=es&country=ES), NMRShiftDB (http://nmrshiftdb.nmr.uni-koeln.de/), Massbank[279] (http://www.massbank.jp/index.html), Metlin[280] (http://metlin.scripps.edu/index.php), and ReSpect[281] (http://spectra.psc.riken.jp/), a MSn spectral database for phytochemicals. The GNPS: Global Natural Products Social Molecular Networking (http://gnps.ucsd.edu/) is a bacterial network that includes selected MS-MS data of ions below *m/z* 2000 from HMDB (http://www.hmdb.ca), LipidMaps (http://www.lipidmaps.org/), MassBank and NIST.[95] GNPS is a global social analysis infrastructure and has the largest database of worldwide contributed MS-MS spectra to data. None of the public databases include taxonomic resources information. However, PubChem, the largest chemical biology public available database, is closely integrated with other literature and biomedical databases such as PubMed, Protein, Gene, Structure and Taxonomy.[282] See Blunt and Munro[283] and Yuliana *et al.*[284] for more comprehensive reviews of NPs available literature, Cheng *et al.*[282] for applications of PubChem database in drug discovery, and Irwin *et al.*[285] for ZINC database as virtual screening. An extensive route has been traversed with respect to database advances for NPs dereplication since Corley and Durley published a review in 1994,[32] especially regarding the systematic inclusion of MS and NMR spectral data and sub-structure searching. The emerging of open-source and open-data implemented databases (*e.g.*, PubChem, NMRShiftDB and GNPS) are of great significance because they can be continuously developed and improved by the entire community, including researchers, funding agencies and open access journals.

## 12 The "omics" revolution

Several developments and scientific advances have improved the analysis of biological systems. Rapidly expanding important tools and research fields, such as: (1) genomics: DNA sequencing and its related research. Genetic fingerprinting and DNA microarray; (2) proteomics: protein concentrations and modifications analysis, especially in response to various parameters; and (3) metabolomics: analogous to proteomics, but dealing with metabolites. Imperatively bioinformatics research efforts had to parallel the development of these fields to enable processing the huge data information provided by the mentioned "omics" topics. All these multi and interdisciplinary research areas play an important and increasing role in NPs dereplication, especially in the discharge of species that have already been studied or do not reveal potential research interest, targeting the most promising for investigation, *i.e.* organism/microorganism biotechnological dereplication.

Table 1 Essential features of selected databases for NPs dereplication

| | Compounds[a] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Database | Total | NPs | Period | MW | MF | UV[b] | NMR[c] | MS[d] | Bioactivity | Taxonomy | SSS[e] |
| CAS/SciFinder | $8.9 \times 10^7$ | >283 000 | Current | + | + | – | – | – | + | + | + |
| CSLS | $4.6 \times 10^7$ | Extracts | ~2010 | + | + | – | – | – | + | – | + |
| ChemSpider | $3.2 \times 10^7$ | >7800 | Current | + | + | + | – | – | + | – | + |
| PubChem | $5.1 \times 10^7$ | >438 00 | Current | + | + | – | – | – | + | – | + |
| ZINC | $3.4 \times 10^7$ | >19 000 | Current | + | – | – | – | – | + | – | + |
| NAPROC-13 | | >6000 | ~2007 | + | + | – | +[c1,c2,c3] | – | – | – | + |
| NMRShiftDB | 42 000 | ?[f] | Current | + | + | – | +[c1,c2,c3] | – | – | – | + |
| Massbank | 13 000 | >2500 | Current | + | + | – | – | +[d1,d2,d3] | – | – | + |
| ReSpect | | >3595 | Current | + | + | – | – | +[d1,d2,d3] | – | – | + |
| Metlin | | 64 000 | Current | + | + | – | – | +[d1,d3] | – | – | + |
| GNPS | $1.6 \times 10^5$ | $>1.4 \times 10^5$ | Current | + | + | – | – | +[d1,d3] | – | – | + |
| NaprAlert | | >150 000 extracts | ~2003[g] | + | + | +[h] | – | – | + | + | – |
| Dictionary NP | | >260 000 | Current | + | + | + | – | – | + | + | + |
| Dictionary MNP | | 25 000 | Current | + | + | + | – | – | + | + | + |
| MarinLit | | 23 500 | Current | + | + | + | +[c1,c2,c3] | – | + | + | +[h] |
| AntiBase | | 42 950 | Current | + | + | +[h] | +[c1,h] | – | + | + | + |
| AntiMarin | | 53 000 | 2013[i] | + | + | +[h] | +[c1,c2,c3,h] | – | + | + | +[h] |

[a] When possible an estimate number of NPs in the database is given. [b] $\lambda$ UV data values. [c] Three NMR data options have been used: [c1] $\delta$ values (experimental or calculated), [c2] spectra or [c3] $^1$H NMR structural features ($^1$H-SF). [d] Three MS data options have been used: [d1] positive, negative, and neutral MSn $m/z$-value, [d2] spectra or [d3] fragment ion ($m/z$). [e] Sub-structure searching. [f] NPs reported in the database, without numbers. [g] Only includes *ca.* 15% of the literature from 2004 to present time. [h] Partial data only. [i] Is the result of a merger between AntiBase (a database of all terrestrial and marine microbial natural products) and MarinLit (a database of marine natural products) that finished in 2013.

### 12.1 Genomics

Genome sequencing is rapidly changing the field of NPs research by providing opportunities to assess the biosynthetic potential of strains prior to chemical analysis or biological testing. Ready access to sequence data is driving the development of new bioinformatics tools and methods to identify the products of silent or cryptic pathways. While genome mining fast became a useful approach to NPs discovery it also became clear that identifying pathways of interest is much easier than finding the associated products. This led to bottlenecks in the dereplication process that must be overcome, to fully realize the potential of genomics-based NPs discovery.[286]

Sequence-based analysis of secondary metabolite biosynthesis using primer sets employed to specifically target biosynthetic types, such as adenylation domains associated with nonribosomal peptide synthetases (NRPS) and ketosynthase (KS) domains associated with type I modular, iterative, hybrid, and enediyne polyketide synthases (PKSs) is a strategy that provides an estimate of a pathways diversity and assesses the biosynthetic richness of individual strains. Bioinformatics evaluation of secondary-metabolite biosynthetic potential that can be applied in the absence of fully assembled pathways or genome sequences can be performed. The rapid identification of strains that possess the greatest potential to produce new secondary metabolites along with those that produce known compounds can be used to improve the process of NPs dereplication by providing a method to prioritize strains for fermentation studies and chemical analysis. Nevertheless the indication of the presence of a certain biosynthetic pathway, for example PKS, does not totally guarantee the production of unknown polyketide-derived secondary metabolites. For example, in a recent work[287] the production of polyketides from kijanimicin and tautomycin classes was predicted based on the PKS analyzes of a *Streptomyces tendae* strain. However, only the known compound kijanimicin 53, **15** (Fig. 16), a synthetic derivative of kijanimicin family, was in fact isolated from this strain. The authors reported several culture efforts to obtain tautomycin **16** (Fig. 16), originally reported in *Streptomyces* sp., or any known/unknown derivatives from tautomycin family which turned out to be unsuccessful.[287]

The term dereplication is also applied for the selection of the most biotechnology-based interesting microorganisms for detailed studies. These approaches are focused on diversifying microbial NPs producing strains and extract libraries, while decreasing genetic and chemical discharge,[50] avoiding their re-testing, isolation and consequent associated costs. Ribosomal 16S DNA sequences, used for phylogenetic studies, is an essential tool for identifying and classifying microorganisms. The discrimination of distinct cultures among morphologically similar strains (dereplication) and the detection of specific biosynthetic pathways in these strains are important steps in the selection of microorganisms to include in NPs libraries. The Basic Local Alignment Search Tool (BLAST),[288] that finds regions of local similarity between sequences, is a forceful bioinformatics tool. BLAST (available at http://blast.ncbi.nlm.nih.gov/Blast.cgi) compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. The discovery of novel species or novel phylogeny raises the probability of finding new bioactive compounds.

There are several analytical and/or computational approaches available for 16S genes profiling and subsequent microbial dereplication, which are described below.

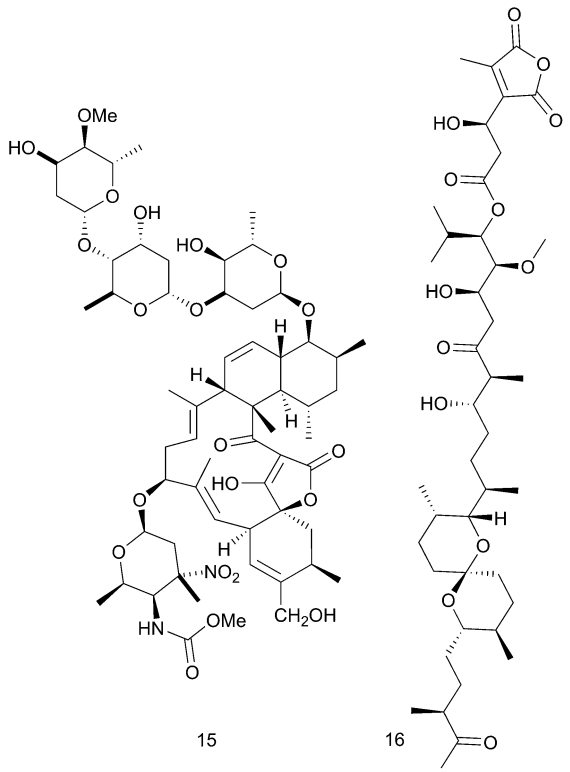



Fig. 16 Polyketide-derived secondary metabolites, kijanimicin 53 **15** and tautomycin **16**.

The use of high-throughput DNA sequencing to produce very large datasets of 16S rDNA sequences in short time periods is now economically affordable. Dereplication of 16S rDNA sequence libraries, removing duplicate sequences from a library and preparing the raw sequences for subsequent analysis, genomics involves: (1) comparing all the sequences in a data set to each other; (2) group similar sequences together; and (3) output a representative sequence from each group. This is possible with the advent of new computer analysis tools like the Java program FastGroup.[47] The described optimal strategies for dereplicating sequences are: (1) trim ambiguous bases from the 5′ end of the sequences and all sequence 3′ of the conserved Bact517 site; (2) match the sequences from the 3′ end; and (3) group sequences equal or higher than 97% identical match to each other.[47]

Pyrolysis mass spectrometry (PyMS)[48] is a fully automated whole-cell fingerprinting technique that enables the rapid and reproducible sorting and profiling of 16S rRNA microorganism genes, using small samples and evidencing discriminatory capacity at the infraspecies level. The congruence found between the clusters defined by the chemometric and molecular fingerprinting techniques was very high and demonstrated the effectiveness of PyMS as a rapid sorting and dereplicating procedure for putatively novel strains. This was outlined by performing polymerase chain reaction-restriction fragment length polymorphism-single-strand conformational polymorphism (FIRS) studies to compare chemometric fingerprinting *vs.* ribotyping fingerprinting methods, in mycolic acid containing actinomycetes strains.[48]

RiboPrinter[289] is a microbial characterization system, an automated instrument that performs ribotyping on bacterial samples, with multiple applications in a NPs research program, which was initially developed for actinomycetes analysis. This system is able to identify closely related isolates and to discriminate between morphologically similar isolates with unique genetic, fatty acid and fermentation profiles. For the detection of biosynthetic genes, a 1006-bp probe containing a portion of an adenylation domain of a non-ribosomal peptide synthetase (NRPS) was employed. Using this alternate probe in place of the standard ribosomal probe, the RiboPrinter was able to detect NRPS genes in several microbial strains.[289]

FT-IR spectroscopy, is also a rapid and reliable whole-organism fingerprinting method, that can be applied as a very useful dereplication tool to indicate which environmental isolates have been previously cultured.[290]

Direct metabolite profiling techniques such as direct injection MS or NMR can easily be used for chemotyping/metabolomics of strains from microbial and fungi collections, using modern informatics tools. BOX and ERIC fingerprinting which are rapid and reproducible can be applied as robust dereplication procedures.[291]

Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) is a powerful tool for microbial dereplication. Moreover, it was approved for diagnostic use by FDA Bruker and bioMérieux platforms. It has higher reproducibility than repetitive element sequence based polymerase chain reaction (rep-PCR), has high-throughput potential, it is faster and less expensive. Its taxonomic resolution was situated at the strain species level. Experiments based on concatenated 16S rRNA, gyrB, and recA gene sequences indicated that phylogeny clustering of bioactive species has also the potential to be a useful dereplication tool in biodiscovery.[292]

Full genome sequencing, a vital tool to disclose complex secondary metabolomes, is becoming affordable and accessible, being already used as a HTS strategy. It allows the analysis of all identifiable secondary NPs gene clusters and determines the percentage of the microorganism's genome dedicated to NPs assembly which can be compared with genome sequences from several other NPs producing microorganisms. As an example for *Salinispora tropica*, a marine actinomycete, genome sequencing exposed the novelty of the majority of the 17 biosynthetic loci. In addition, bioinformatic analysis not only was critical for the structure elucidation of the polyene macrolactam salinilactam A **17** (Fig. 17), but its structural analysis aided the genome assembly of the highly repetitive slm loci. Thus firmly establishing the genus *Salinispora* as a rich source of drug-like molecules and importantly revealing the strong interplay between genomic analysis and traditional NP isolation studies.[293]

### 12.2 Metabolomics

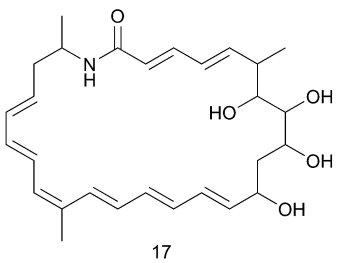
The great majority of protocols for taxonomic dereplication of microbial strains mostly use molecular tools which do not take into consideration the ability of these selected bacteria to produce secondary metabolites. As the identification of novel

chemical entities is one of the key elements driving drug discovery programs, novel methodologies to dereplicate microbial strains by metabolomics approaches needed to be brought to light. Metabolite profiling is important for functional genomics and metabolomics methods, being utilized to screen diverse biological sources of potentially novel and sustainable sources of pharmacologically-active drugs.

Metabolomics has become an effective tool in systems biology, allowing insight to be gained into the potential of natural isolates for the synthesis of significant quantities of promising new agents which enables the environment within fermentation systems to be manipulated in a rational manner to select a desired metabolome.[123]

Due to the important advances registered in analytical techniques, profiling methods for the analysis of crude extracts from biological sources have evolved into powerful tools for dereplication, quality assessment and metabolomics. Metabolite profiling of crude extracts represents a challenging analytical task since these mixtures are composed of hundreds of NPs. Depending on the type of study the focus can be put on major bioactive constituents or minor significant biomarkers. In many cases, a rapid on-line or at-line identification of the compound(s) of interest and in some cases of all detected constituents (metabolomics) is required. The most common techniques for these types of analyses consist of a detection method hyphenated to HPLC, such as LC-PDA, LC-MS or LC-NMR. With the evolution of multivariate data analysis (MVDA) methods, profiling extracts may also rely on direct NMR or MS analysis without prior HPLC separation, which requires high resolution instruments.[102,294]

Adding LC-HR-MS to the stratagem improves a process that provides the ability to identify putative novel chemical entities as NPs discovery leads. In more detail, to process large and complex three dimensional LC-HR-MS datasets, the reported method uses a bucketing and presence–absence standardization strategy in addition to statistical analysis tools including principal component analysis (PCA) and cluster analysis. LC-MS-PCA is effective for strain prioritization in a drug discovery program, supporting drug discovery in the search for unique NPs and for rapid assessment of regulation of NPs production. Demonstrating that grouping bacteria according to the chemical diversity of the produced metabolites is reproducible and provides great improved resolution for the discrimination and prioritization of microbial strains compared to current molecular dereplication techniques.[122,295]




17

Fig. 17 Salinilactam A **17**, from *Salinispora tropica*.

Additionally, using the metabolomics tools through the employment of high resolution Fourier transform mass spectrometry coupled to liquid chromatography (LC-HRFTMS) and NMR spectroscopy, one can establish the chemical profile of endophytic and/or endozoic microbial extracts and their plant or animal sources, resulting in a very efficient dereplication approach. Identifying the compounds of interest at an early stage will aid in the isolation of the bioactive components.

Metabolomic profiling has found its application in screening extracts of macroorganisms as well as in the isolation and cultivation of microorganisms that produce bioactive natural NPs. Metabolomics is being applied to identify and biotechnologically optimize the production of pharmacologically active secondary metabolites. The links between metabolome evolution during optimization and processing factors can be identified through metabolomics. Information obtained from a metabolomics dataset can efficiently establish cultivation and production processes at a small scale which will be finally scaled up to a fermenter system, while maintaining or enhancing synthesis of the desired compounds. MZmine[124] and SIEVE[296] software are employed to perform differential analysis of sample populations to find significant expressed features of complex biomarkers between variable parameters. Metabolomes are identified with the support of existing high resolution MS and NMR records from public or commercial databases (see databases section for details) and further validated through available reference standards and NMR experiments.[123]

### 12.3 Proteomics

The use of proteomics approaches for dereplication purposes, is not as widely described as for the other "omics" fields. Nevertheless it presents a unique opportunity to look at the relationships between the bioactivity profile of the microorganisms under study and their genome, proteome and metabolome characteristics. It will definitively contribute to the understanding of the molecular pathways that control the synthesis of targeted bioactive metabolites and the future optimization of their production. While virtually describing all proteins present in the cell and providing information on their abundance and post-translational modifications, proteomics has emerged as an indispensable strategy to decipher the molecular mechanisms underlying cell metabolisms.

Characterization of the proteins involved in the biosynthesis of bioactive compounds can be an important step in the drug discovery process. Such biosynthetic pathways are usually unveiled using a differential proteomic approach, based on 2D-electrophoresis (2DE) and advanced MS techniques. Automated molecular formula determination by tandem high-resolution MS-MS spectroscopic data is a valuable method for the rapid screening and identification of small molecules such as the dereplication of NPs, characterization of drug metabolites, and identification of small peptide fragments in proteomics.[80]

In addition to MALDI, typing an alternative diagnostic method based on the knowledge of specific NRPs and other metabolite structures, is an approach that can be used in microorganism mixture analysis representing a similar benefit

from proteomics, by going from peptide mapping to peptide sequencing and aiding better identification rates.[297]

## 13 *In silico* dereplication

The inspection of new approved drugs during 1981–2010 shows that there is a downward trend line since the highpoint in 1987 (68 new chemical entities (NCEs) per year) and a minimum of 24 NCEs per year in 2004.[1] Therefore, one of the most important questions in drug discovery still remains unanswered: where in chemical structural space are biologically relevant compounds to be found? The answer to this question is not easy, however as Stockwell stated "The mapping of biological-activity space using small molecules is akin to mapping the stars – uncharted territory is explored using a system of coordinates that describes where each new feature lies".[298] From these coordinates it is possible to identify some of the central criteria in designing compound libraries to modulate the functions of proteins: (1) diversity; (2) drug-likeness; and (3) biological relevance. The last criterion is fulfilled by the NPs which have been optimized in a very long natural selection process for optimal interaction with biological macromolecules. Two distinct but complementary computer-driven drug discovery approaches can be applied: ligand-based methodology and structure-based methodology.

In the ligand-based virtual screening process, the most effective biologically active lead molecule is detected using a structural, topological or pharmacophoric similarity search. The Lipinski rule-of-five is widely used as a filter in drug discovery to evaluate drug-likeness and to verify if a compound with some activity has properties that would make it a possible orally active drug in humans.[299] Waldmann *et al.*[300] suggested from a statistical analysis of the structural classification of NPs that more than half of all NPs have just the right size to serve as a starting point for hit and lead discovery. Ertl and co-workers[301] developed NP-likeness score, a Bayesian measure which permits the determination of how molecules are similar to the structural space covered by NPs. Recently, we reported that the topological descriptor MDEO-12,[302] the electronic descriptor TopoPSA,[302] the quantum-chemical descriptor and the energy of the highest occupied molecular orbital ($\varepsilon_{HOMO}$)[303] have a remarkable performance in discriminating antitumor, antibiotic and overall biological lead-like compounds, respectively. These approaches can be used in virtual screening, prioritization of compound libraries and design of building blocks towards NPs lead-like libraries. One of the most widely used virtual screening approaches is Quantitative Structure Activity Relationship (QSAR) modeling.[304] Tropsha *et al.*[305] reported the discovery of novel tylophrine derivatives as anticancer agents using combined approaches of validated QSAR modeling and virtual screening. They experimentally tested ten structurally diverse hits that were predicted and eight of these were confirmed to be active with the highest experimental EC(50) of 1.8 μM implying an exceptionally high hit rate (80%).[305] More recently, a diverse set of flavonoids structures that were isolated from plants, used in traditional medicine, were investigated by Wiese and co-workers.[306] The flavones retusin and ayanin which possess a rare C-methylated structure were found to be highly potent inhibitors of breast cancer resistance protein, showing only slightly less potency than the most potent inhibitor known so far (ABCG2). Through 2D and 3D QSAR modeling, the authors were able to identify the structural features which significantly influence the inhibitory potency.[306]

In the structure-based virtual screening process, the 3D structure of the target of interest must be known from X-ray crystallography, NMR spectroscopy or molecular modeling. Furthermore, if an X-ray structure of the protein with a ligand is available, the binding mode of the ligand can be analyzed as well. Structure-based methodology involves automated and fast docking of a large number of chemical compounds against a protein-binding or active site, taking advantage of the growing number of protein 3D-structures.[307,308] The therapeutic targets of nine medicinal plant ingredients (*e.g.* genistein, ginsenoside Rg1, quercetin, acronycine, baicalin, emodin, allicin, catechin, camptothecin) were predicted using an extended ligand–protein docking method, INVDOCK,[309] and the results obtained were in accordance with available experimental findings.[310] Rollinger *et al.*[311] reported a structure-based pharmacophore model utilizing an *in silico* filtering experiment for targeting the selection of acetylcholinesterase (AChE) inhibitors from more than 110 000 NPs. Two coumarin derivatives, scopoletin **18** and glucoside scopolin **19** (Fig. 18), were proposed as promising AChE inhibiting hits from the virtual screening procedure and successfully tested with respect to their anticholinesterase potential.[311]

Recently, ten NPs from an *in-house* NPs database were successfully identified using structure-based virtual screening, as cysteine protease falcipain-2 (FP-2) inhibitors, one of the most promising targets for antimalarial agents discovery.[312] Moreover, these ten caffeate, flavonoid and flavonoid glycoside derivatives have showed moderate inhibitory activities against FP-2 with $IC_{50}$ values ranging from 3.18 to 68.19 μM.[312] See Langer and Krovat,[313] Hou and Xu,[314] and Rollinger *et al.*,[315] for more comprehensive reviews of the available literature.

## 14 Combined techniques

Combined techniques embody an added value to avoid the rediscovery of NPs. Besides the use of hyphenated techniques coupling any other described methodologies in this review is conceivable. Examples are, dereplication strategies that use analytical techniques and database searching to determine the identity of an active compound at the earliest possible stage in the discovery process. This prevents wasted efforts on samples with no potential for development and allows resources to be focused on the most promising lead.[4] MS and HPLC-MS spectrometry together with spectral databases are powerful tools in the chemometric profiling of bio-resources for NPs production. High throughput techniques, high sensitivity LC-NMR and LC-SCD are also emerging tools in this area. Additionally, the structures of breitfussin A **20** and B **21** (Fig. 19), comprising a rare molecular framework, with the combination of an indole, an oxazole and a pyrrole, were recently supported by a combination of atomic-force microscopy (AFM), CASE, and DFT

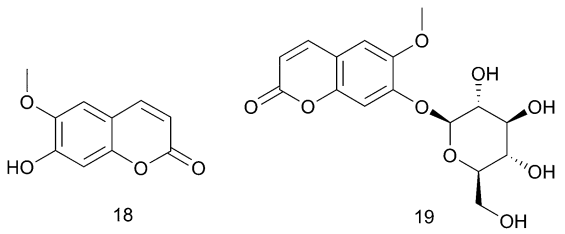



Fig. 18 Sopoletin **18** and glucoside scopolin **19**, isolated from the medicinal plant *Scopolia carniolica* Jaqc.

calculations. However, none of these was able to propose a unique solution individually.[316]

Screening of NPs extract libraries continues to furnish novel lead molecules for further drug development, despite the analysis challenges and the prioritization of potential NPs hits.[10] Bioassay guided isolation is broadly used to prioritize crude extracts, fractions or pure compounds. The development of simple and rapid bioassays is decisive for an efficient localization of active principles enabling a direct screening of bioactive constituents either by refined HTS automation robots or in small scale facilities. Nevertheless the development of modern LC hyphenated techniques, such as LC-UV, LC-MS, LC-MS-MS, LC-NMR, and LC-MS-MS-NMR, combined with any of the described HTS systems, are potent procedures that are becoming routinely used in NPs studies. Hence, techniques that can be applied *in situ*, or in real-time monitoring of NPs in biological samples still need to be improved to comprehend the biological functions of secondary metabolites.[14] Characterization of small molecules and their interaction with natural proteins (*e.g.* receptors, ion channels) also include well-known and widely used standard analytical methods, such as HPLC-MS, NMR, calorimetry and X-ray diffraction, combined with newer and more specialized analytical methods (*e.g.* biosensors), biological systems (*e.g.* cell lines and animal models), and *in silico* approaches.[317] From our point of view implementing routine bioassays on healthy cells at the same time that compound libraries are being screened, for example screening of human and veterinary microbial, fungal and viral pathogens, cancer types as well as several other diseases, would give a push moving forward the selection process of drug-like *vs.* lead-like hits.

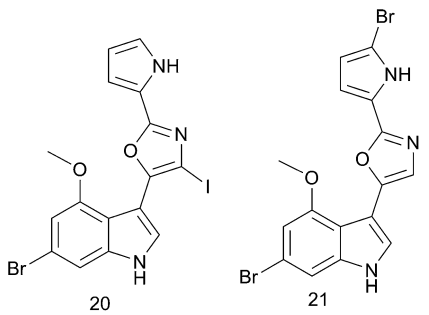



Fig. 19 Highly modified halogenated dipeptides, breitfussin A **20** and B **21**, from *Thuiaria breitfussi*.

## 15 Conclusions and future prospects

"Chemoprospection" as we like to name the search for treasurable novel drug-like agents, embraces an extensive number of dereplication strategies that can make the NCEs discovery process faster. The publication boom that we have been assisting since 2012, on the topic of dereplication is the consequence of researchers' versatility and aptitude to take advantage of a vast array of multidisciplinary advents that can meet the purpose of pursuing novel NPs with potential biotechnological and industrial applications. This occurrence is directly related to the extraordinary technological advances accomplished *via* an impressive number of methodologies tackling the dereplication blockage, from small molecules to species-picking strategies, in such different investigation fields, directions and successful ways, allowing the selection of the most prolific species and exploring the NPs chemical and genomic space. A long way has been travelled by dereplication in 36 years, switching from a little used approach by NPs chemists in the early 90's to currently being an extremely potent and indispensable used manifold procedure. We highlight some examples, such as HTS screening, analytical technology for separation (particularly hyphenated and combined techniques), detection methods (*e.g.* DAD, MS and NMR) and the achievement of comprehensive molecular structural/spectral databases. Progress in MS, MS-MS, MS-MS molecular networking,[159,202,203] IMS,[117] followed by computational MS tools such as combinatorial optimization,[176–178] machine learning,[179,180] and *in silico* techniques and genome mining methods,[194,195,201] have profoundly contributed to advance dereplication methods, with the consequential leap forward in drug discovery from natural sources. Enhanced sensitivity in modern NMR spectrometers definitely developed NPs dereplication, reducing data acquisition time and sample amount, which is a crucial parameter in the NPs field, giving credit to the advent of capillary cryoprobes and ASAP, -HMQC[218] methods, non-uniform sampling (or sparse sampling)[18,219] and UDB[228,229] methodologies. Likewise, DOSY[233] experiments, complementary to LC-MS, in particular LOCODOSY[248] and matrix-assisted DOSY methods[249,250] are additional tools to the "bag of tricks", available for boosting dereplication of known or nuisance NP compounds.

The use of proteomics for dereplication purposes is not as widely described as the other "omics". However, with regard to drug discovery, knowing the targets by characterization of the proteins involved in the biosynthesis of the bioactive compounds can be a key factor, especially from the chemoinformatics (receptor-based computational drug discovery approach), NMR and X-ray points of view. The use of open-source docking simulators containing cohesive updated information about protein 3D-structures, protein-binding, active sites, diversity, drug-likeness and biological importance, will allow the dereplication concept to expand from the discharge of known NPs to the discharge of chembiochemical non valuable NPs.

Dereplication will adapt and keep following the pace of new scientific inventions as far as our imagination can go. As a

result, we are going to highpoint approaches that, in our opinion, are the most auspicious and in addition we are going to provide several insights that will take some of the reviewed approaches to a whole new level.

The recent developments in the SENECA open-source and open-data platform[273,274] carrying out search in a fully comprehensive chemical space, comprising all possible chemical compounds generated by computational approaches (*i.e.* with the same molecular formula), will lead to major progress in CASE systems for NPs dereplication applications. We believe that CASE is going to be advanced to a level that will allow the systematic assessment of biological-activity space.

Thinking about the methodical inclusion of NMR and MS spectral data and sub-structure searching, the emergence of open-source and open-data implemented databases, *e.g.*, PubChem, NMRShiftDB, GNPS have relevant impact and can be continuously developed and upgraded by the entire NPs "prospectors" community.

HTS full genome sequencing will make the number of sequenced genomes exponentially increase in the next decade. Together with bioinformatics tools it will allow the analysis of all identifiable secondary NP gene clusters and determine the percentage of the genome dedicated to NPs assembly. Currently starting to be widely applied to microorganisms, imagine the power of expanding it similarly to macroorganisms (*e.g.* plants, invertebrates) and additionally to build a database with all these data. Enabling the comparison of our organism of study to known genome sequences of NPs producing organisms, it will definitely accelerate the discovery of novel drug-like NPs.

Despite developments of databases, the lack of integrated databases is a major hindrance for rapid dereplication. Like DNA sequences are stored at GenBank, in our opinion, building a universal open-source and open-data all-inclusive chemical database will solve the fundamental database bottleneck, revolutionizing dereplication as it is performed today. We strongly believe that worldwide open-source, open-data database implementation policies, and the merger of widespread available information in a unique library, a "NPsDataBank", including information provided by CASE, NMR, MSn and full genome sequencing, *etc.*, with standard features, enabling its use in any kind of analytical software, regardless of the equipment brand, as well as in a platform in a website, will map NPs at a global scale and will have the power of boosting dereplication as we never seen before. Such a worldwide "NPsDataBank" could be achieved with data generated through advanced computational spectra prediction and *in silico* building NPs lead-like library approaches and additionally by uploading and registering NPs data. Submitting and providing all the data as it is in peer review of publications/patents (*e.g.* HR-MS, 1D and 2D NMR, MS and MSn, *etc.*), which is currently an under-exploited data source. Combining these records and creating this "NPsDataBank" would permit by uploading experimental compound data (*e.g.* 1D NMR or/and 2D NMR spectra, NP chemical structure, *etc.*) immediate access to matches, identical structures and unknown compounds.

In addition to what has been mentioned above, we also consider of vital importance the implementation of toxicity screening in healthy cells in an early stage of NPs downstream analysis as a way of taking the fastest lane to get NCEs. A large percentage of novel lead-like drugs end up failing phase I of clinical trials due to toxicity issues. If the objective is "digging" for drug-leads instead of bioactive compounds only, this procedure, using either *in vitro*, *in vivo*, *virtual* or cross screenings, will save considerable research investment.

The outstanding LC-SCD technique[208] brings tremendous value to several industrial and commercial applications. It has the power of totally modeling compound structure elucidation research as it is now, inaugurating a new era for drug discovery when routinely applied. Hybrid techniques such as HPLC-DAD-BDC-MS-SPE-NMR-SCD will be a dream come true for NPs chemists. This will allow researchers not only to perform efficient and ultrafast dereplication, but will also advance and accelerate structure elucidation, by directing to SCD promising compounds obtained in small amounts, whose structure cannot be fully justified by NMR.

The tendency is to minimize sample preparation and avoid compound separation; we predict that the evolution of multivariate data analysis methods will proceed, with new NMR and MS technologies focused on increasing sensitivity and minimizing background. Taking into account, the amazing MS-based strategy that enables genome mining families of small-molecules in living microbial populations, integrating leading chemistry with genomics and phenotypes of microbial colonies, we can conceive escalating it to more complex organisms, which will make this approach very close to an ideal methodology. Analytical equipment miniaturization, to carry out and perform *in loco* analysis, will exponentially expand the variety of uses in the NPs area.[15]

Recent cutting-edge developments in adaptive optics (AO) microscopy have dramatically improved resolution, proving particularly promising for applications that require images from deep within biological tissue specimens.[318] Additionally, the "one-Ångström barrier" plus the "deep sub-Ångström" resolution regions were successfully exceeded. Although the use of microscopy for dereplication objectives is still in a very early stage, we have confidence that microscopy will effectively transition from a precision laboratory instrument to a rugged, frontline tool for dereplication and structure elucidation, in the near future. For example, 3D *in loco* or *in situ* atomic-force microscopy (AFM)[316] will increase the speed and productivity of drug discovery, changing dereplication as it is performed nowadays and in the years to come. It is obvious that a NPs researcher's ultimate goal is the development of *in loco* and *in situ* dereplication and structure elucidation approaches. Though extremely challenging it is also absolutely superb and convenient, we envision that in the near future we will get to the precision and sensitivity level of analyzing pure NPs, and in a further advanced stage NPs sources themselves, in a microscope and actually perceive the molecular structures, using the same principle to realize chemical reactions. Performing structure elucidation through visualization of the NPs at the subcellular level is what the future holds.

We have no doubt that the output of such fantastic research evolution described in the present review and what is yet to

come will lead to a cumulative percentage of novel NCEs in the forthcoming years, meeting vital societal challenges related to a broad range of diseases.

## 16 Funding information

Fundação para a Ciência e a Tecnologia (FCT), Ministério da Educação e Ciência, Portugal and FEDER through grant no. PTDC/QUI-QUI/119116/2010, PEst-C/EQB/LA0006/2013 and PEst-OE/BIA/UI0457/2011-CREM, and the European Union 7th Framework Programme (FP7/2007–2013) under grant agreement no. PCOFUND-GA-2009-246542.

## 17 References

1 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2012, **75**, 311–335.
2 W. H. Gerwick and B. S. Moore, *Chem. Biol.*, 2012, **19**, 85–98.
3 Y. G. Shin and R. B. van Breemen, *Biopharm. Drug Dispos.*, 2001, **22**, 353–372.
4 K. V. Sashidhara and J. N. Rosaiah, *Nat. Prod. Commun.*, 2007, **2**, 193–202.
5 D. D. Baker, M. Chu, U. Oza and V. Rajgarhia, *Nat. Prod. Rep.*, 2007, **24**, 1225–1244.
6 J.-L. Wolfender, E. F. Queiroz and K. Hostettmann, *Expert Opin. Drug Discovery*, 2006, **1**, 237–260.
7 O. Potterat and M. Hamburger, *Nat. Prod. Rep.*, 2013, **30**, 546–564.
8 C. J. Henrich and J. A. Beutler, *Nat. Prod. Rep.*, 2013, **30**, 1284–1298.
9 D. J. Hook, *J. Biomol. Screening*, 1998, **3**, 78.
10 F. E. Koehn, *Prog. Drug Res.*, 2008, **65**, 177–210.
11 M. A. Strege, *J. Chromatogr. B: Biomed. Sci. Appl.*, 1999, **725**, 67–78.
12 E. J. Ashforth, C. Fu, X. Liu, H. Dai, F. Song, H. Guo and L. Zhang, *Nat. Prod. Rep.*, 2010, **27**, 1709–1719.
13 F. Hufsky, K. Scheubert and S. Boecker, *Nat. Prod. Rep.*, 2014, **31**, 807–817.
14 C.-J. Shih, P.-Y. Chen, C.-C. Liaw, Y.-M. Lai and Y.-L. Yang, *Nat. Prod. Rep.*, 2014, **31**, 739–755.
15 A. Bouslimani, L. M. Sanchez, N. Garg and P. C. Dorrestein, *Nat. Prod. Rep.*, 2014, **31**, 718–729.
16 D. Krug and R. Mueller, *Nat. Prod. Rep.*, 2014, **31**, 768–783.
17 T. O. Larsen, J. Smedsgaard, K. F. Nielsen, M. E. Hansen and J. C. Frisvad, *Nat. Prod. Rep.*, 2005, **22**, 672–695.
18 R. C. Breton and W. F. Reynolds, *Nat. Prod. Rep.*, 2013, **30**, 501–524.
19 T. Ito and M. Masubuchi, *J. Antibiot.*, 2014, **67**, 353–360.
20 M. Halabalaki, K. Vougogiannopoulou, E. Mikros and A. L. Skaltsounis, *Curr. Opin. Biotechnol.*, 2014, **25**, 1–7.
21 M. Elyashberg, A. J. Williams and K. Blinov, *Nat. Prod. Rep.*, 2010, **27**, 1296–1328.
22 M. Jaspars, *Nat. Prod. Rep.*, 1999, **16**, 241–247.
23 C. Steinbeck, *Nat. Prod. Rep.*, 2004, **21**, 512–518.
24 O. Genilloud, I. Gonzalez, O. Salazar, J. Martin, J. Ruben Tormo and F. Vicente, *J. Ind. Microbiol. Biotechnol.*, 2011, **38**, 375–389.
25 T. Roemer, D. Xu, S. B. Singh, C. A. Parish, G. Harris, H. Wang, J. E. Davies and G. F. Bills, *Chem. Biol.*, 2011, **18**, 148–164.
26 A. Agarwal, P. D'Souza, T. S. Johnson, S. M. Dethe and C. V. Chandrasekaran, *Curr. Opin. Biotechnol.*, 2014, **25**, 39–44.
27 O. Sticher, *Nat. Prod. Rep.*, 2008, **25**, 517–554.
28 F. Bucar, A. Wube and M. Schmid, *Nat. Prod. Rep.*, 2013, **30**, 525–545.
29 J. Berdy, *J. Antibiot.*, 2005, **58**, 1–26.
30 L. J. Hanka, S. L. Kuentzel, D. G. Martin, P. F. Wiley and G. L. Neil, *Recent Results Cancer Res.*, 1978, **63**, 69–76.
31 J. A. Beutler, A. B. Alvarado, D. E. Schaufelberger, P. Andrews and T. G. McCloud, *J. Nat. Prod.*, 1990, **53**, 867–874.
32 D. G. Corley and R. C. Durley, *J. Nat. Prod.*, 1994, **57**, 1484–1490.
33 J. Antonio and T. F. Molinski, *J. Nat. Prod.*, 1993, **56**, 54–61.
34 J. H. Cardellina, M. H. G. Munro, R. W. Fuller, K. P. Manfredi, T. C. McKee, M. Tischler, H. R. Bokesch, K. R. Gustafson, J. A. Beutler and M. R. Boyd, *J. Nat. Prod.*, 1993, **56**, 1123–1129.
35 H. L. Constant and C. W. W. Beecher, *Nat. Prod. Lett.*, 1995, **6**, 193–196.
36 S. L. Zhou and M. Hamburger, *J. Chromatogr. A*, 1996, **755**, 189–204.
37 J. L. Wolfender, C. Terreaux and K. Hostettmann, *Pharm. Biol.*, 2000, **38**, 41–54.
38 A. Tsipouras, J. Ondeyka, C. Dufresne, S. Lee, G. Salituro, N. Tsou, M. Goetz, S. B. Singh and S. K. Kearsley, *Anal. Chim. Acta*, 1995, **316**, 161–171.
39 C. Peng, S. G. Yuan, C. Z. Zheng and Y. Z. Hui, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 805–813.
40 C. Peng, S. G. Yuan, C. Z. Zheng, Z. S. Shi and H. M. Wu, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 539–546.
41 M. Kock, J. Junker, W. Maier, M. Will and T. Lindel, *Eur. J. Org. Chem.*, 1999, 579–586.
42 T. Lindel, J. Junker and M. Kock, *Eur. J. Org. Chem.*, 1999, 573–577.
43 G. A. Cordell, C. W. W. Beecher, A. D. Kinghorn, J. M. Pezzuto, H. L. Constant, H. B. Chai, L. Q. Fang, E. K. Seo, L. Long, B. L. Cui and K. Slowing-Barillas, *Struct. Chem.*, 1997, **19**, 749–791.
44 G. A. Cordell and Y. G. Shin, *Pure Appl. Chem.*, 1999, **71**, 1089–1094.
45 G. A. Cordell, *Phytochemistry*, 2000, **55**, 463–480.
46 J. A. Colquhoun, J. Zulu, M. Goodfellow, K. Horikoshi, A. C. Ward and A. T. Bull, *Antonie van Leeuwenhoek*, 2000, **77**, 359–367.
47 V. Seguritan and F. Rohwer, *BMC Bioinf.*, 2001, **2**, 9.
48 P. F. B. Brandao, M. Torimura, R. Kurane and A. T. Bull, *Appl. Microbiol. Biotechnol.*, 2002, **58**, 77–83.
49 S. Donadio, P. Monciardini, R. Alduina, P. Mazza, C. Chiocchini, L. Cavaletti, M. Sosio and A. M. Puglia, *J. Biotechnol.*, 2002, **99**, 187–198.
50 V. Knight, J. J. Sanglier, D. DiTullio, S. Braccili, P. Bonner, J. Waters, D. Hughes and L. Zhang, *Appl. Microbiol. Biotechnol.*, 2003, **62**, 446–458.

51 S. C. Bobzin, S. T. Yang and T. P. Kasten, *J. Chromatogr. B: Biomed. Sci. Appl.*, 2000, **748**, 259–267.
52 S. C. Bobzin, S. Yang and T. P. Kasten, *J. Ind. Microbiol. Biotechnol.*, 2000, **25**, 342–345.
53 J. R. Gilbert, P. Lewer, D. O. Duebelbeis, A. W. Carr, C. E. Snipes and R. T. Williamson, in *Liquid Chromatography/Mass Spectrometry, Ms/Ms and Time-of-Flight Ms*, ed. I. Ferrer and E. M. Thurman, 2003, vol. 850, pp. 52–65.
54 J. L. Wolfender, K. Ndjoko and K. Hostettmann, *J. Chromatogr. A*, 2003, **1000**, 437–455.
55 K. Hostettmann, J. L. Wolfender and C. Terreaux, *Pharm. Biol.*, 2001, **39**, 18–32.
56 J. Bradshaw, D. Butina, A. J. Dunn, R. H. Green, M. Hajek, M. M. Jones, J. C. Lindon and P. J. Sidebottom, *J. Nat. Prod.*, 2001, **64**, 1541–1544.
57 A. D. Kinghorn, N. R. Farnsworth, D. D. Soejarto, G. A. Cordell, S. M. Swanson, J. M. Pezzuto, M. C. Wani, M. E. Wall, N. H. Oberlies, D. J. Kroll, R. A. Kramer, W. C. Rose, G. D. Vite, C. R. Fairchild, R. W. Peterson and R. Wild, *Pharm. Biol.*, 2003, **41**, 53–67.
58 K. F. Nielsen and J. Smedsgaard, *J. Chromatogr. A*, 2003, **1002**, 111–136.
59 O. Potterat, K. Wagner and H. Haag, *J. Chromatogr. A*, 2000, **872**, 85–90.
60 J. L. Wolfender, P. Waridel, K. Ndjoko, K. R. Hobby, H. J. Major and K. Hostettmann, *Analusis*, 2000, **28**, 895–906A.
61 G. E. Martin, C. E. Hadden, D. J. Russell, B. D. Kaluzny, J. E. Guido, W. K. Duholke, B. A. Stiemsma, T. J. Thamann, R. C. Crouch, K. Blinov, M. Elyashberg, E. R. Martirosian, S. G. Molodtsov, A. J. Williams and P. L. Schiff, *J. Heterocycl. Chem.*, 2002, **39**, 1241–1250.
62 C. C. Stessman, R. Ebel, A. J. Corvino and P. Crews, *J. Nat. Prod.*, 2002, **65**, 1183–1186.
63 B. Jaki, S. Franzblau and G. F. Pauli, *Phytochem. Anal.*, 2004, **15**, 213–219.
64 K. A. Blinov, D. Carlson, M. E. Elyashberg, G. E. Martin, E. R. Martirosian, S. Molodtsov and A. J. Williams, *Magn. Reson. Chem.*, 2003, **41**, 359–372.
65 M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin and E. R. Martirosian, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 771–792.
66 C. Steinbeck and S. Kuhn, *Phytochemistry*, 2004, **65**, 2711–2717.
67 A. D. Buss and M. S. Butler, *Drug Dev. Res.*, 2004, **62**, 362–370.
68 H. Tang, C. Xiao and Y. Wang, *Magn. Reson. Chem.*, 2009, **47**, S157–S162.
69 C. Clarkson, D. Staerk, S. H. Hansen and J. W. Jaroszewski, *Anal. Chem.*, 2005, **77**, 3547–3553.
70 A. Fredenhagen, C. Derrien and E. Gassmann, *J. Nat. Prod.*, 2005, **68**, 385–391.
71 J. W. Jaroszewski, *Planta Med.*, 2005, **71**, 691–700.
72 J. W. Jaroszewski, *Planta Med.*, 2005, **71**, 795–802.
73 L. Zhang, *Nat. Prod.*, 2005, 33–55.
74 J. L. Wolfender, E. F. Queiroz and K. Hostettmann, *Magn. Reson. Chem.*, 2005, **43**, 697–709.
75 L. L. Silver, in *Proceedings of the First International Symposium on Natural Preservatives in Food Systems*, ed. D. HavkinFrenkel, C. Frenkel and N. Dudai, 2006, pp. 115–123.
76 D. R. Appleton, A. D. Buss and M. S. Butler, *Chimia*, 2007, **61**, 327–331.
77 Y. Konishi, T. Kiyota, C. Draghici, J.-M. Gao, F. Yeboah, S. Acoca, S. Jarussophon and E. Purisima, *Anal. Chem.*, 2007, **79**, 1187–1197.
78 D. Wolf and K. Siems, *Chimia*, 2007, **61**, 339–345.
79 C. A. Motti, M. L. Freckelton, D. M. Tapiolas and R. H. Willis, *J. Nat. Prod.*, 2009, **72**, 290–294.
80 S. Jarussophon, S. Acoca, J.-M. Gao, C. Deprez, T. Kiyota, C. Draghici, E. Purisima and Y. Konishi, *Analyst*, 2009, **134**, 690–700.
81 D. Staerk, J. R. Kesting, M. Sairafianpour, M. Witt, J. Asili, S. A. Emami and J. W. Jaroszewski, *Phytochemistry*, 2009, **70**, 1055–1061.
82 V. Pieri, S. Sturm, C. Seger, P. Schneider and H. Stuppner, *Planta Med.*, 2009, **75**, 996.
83 K. F. Nielsen, M. Mansson, C. Rank, J. C. Frisvad and T. O. Larsen, *J. Nat. Prod.*, 2011, **74**, 2338–2348.
84 M. E. Hansen, J. Smedsgaard and T. O. Larsen, *Anal. Chem.*, 2005, **77**, 6805–6817.
85 G. Lang, N. A. Mayhudin, M. I. Mitova, L. Sun, S. van der Sar, J. W. Blunt, A. L. J. Cole, G. Ellis, H. Laatsch and M. H. G. Munro, *J. Nat. Prod.*, 2008, **71**, 1595–1599.
86 R. Dieckmann, I. Graeber, I. Kaesler, U. Szewzyk and H. von Dohren, *Appl. Microbiol. Biotechnol.*, 2005, **67**, 539–548.
87 E. Esquenazi, C. Coates, L. Simmons, D. Gonzalez, W. H. Gerwick and P. C. Dorrestein, *Mol. BioSyst.*, 2008, **4**, 562–570.
88 J. Ghyselinck, K. Van Hoorde, B. Hoste, K. Heylen and P. De Vos, *J. Microbiol. Methods*, 2011, **86**, 327–336.
89 G. G. Harrigan and G. H. Goetz, *Comb. Chem. High Throughput Screening*, 2005, **8**, 529–534.
90 J. Bitzer, B. Koepcke, M. Stadler, V. Heilwig, Y.-M. Ju, S. Seip and T. Henkel, *Chimia*, 2007, **61**, 332–338.
91 R. Dunkel and X. Wu, *J. Magn. Reson.*, 2007, **188**, 97–110.
92 J. L. Lopez-Perez, R. Theron, E. del Olmo and D. Diaz, *Bioinformatics*, 2007, **23**, 3256–3257.
93 M. I. Mitova, A. C. Murphy, G. Lang, J. W. Blunt, A. L. J. Cole, G. Ellis and M. H. G. Munro, *J. Nat. Prod.*, 2008, **71**, 1600–1603.
94 T. A. Johnson, J. Sohn, W. D. Inman, S. A. Estee, S. T. Loveridge, H. C. Vervoort, K. Tenney, J. Liu, K. K.-H. Ang, J. Ratnam, W. M. Bray, N. C. Gassner, Y. Y. Shen, R. S. Lokey, J. H. McKerrow, K. Boundy-Mills, A. Nukanto, A. Kanti, H. Julistiono, L. B. S. Kardono, L. F. Bjeldanes and P. Crews, *J. Nat. Prod.*, 2011, **74**, 2545–2555.
95 J. Y. Yang, L. M. Sanchez, C. M. Rath, X. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, W. R. Wong, R. G. Linington, L. Zhang, H. M. Debonsi, W. H. Gerwick and P. C. Dorrestein, *J. Nat. Prod.*, 2013, **76**, 1686–1699.

96 J. Ng, N. Bandeira, W.-T. Liu, M. Ghassemian, T. L. Simmons, W. H. Gerwick, R. Linington, P. C. Dorrestein and P. A. Pevzner, *Nat. Methods*, 2009, **6**, 596–599.
97 M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, A. J. Williams and G. E. Martin, *J. Chem. Inf. Model.*, 2007, **47**, 1053–1066.
98 M. E. Elyashberg, A. Williams and G. E. Martin, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2008, **53**, 1–104.
99 C. M. Farnet and E. Zazopoulos, *Nat. Prod.*, 2005, 95–106.
100 K. Boroczky, H. Laatsch, I. Wagner-Dobler, K. Stritzke and S. Schulz, *Chem. Biodiversity*, 2006, **3**, 622–634.
101 D. P. Overy, D. P. Enot, K. Tailliart, H. Jenkins, D. Parker, M. Beckmann and J. Draper, *Nat. Protoc.*, 2008, **3**, 471–485.
102 J.-L. Wolfender, G. Marti and E. F. Queiroz, *Curr. Org. Chem.*, 2010, **14**, 1808–1832.
103 Z. Kamenik, F. Hadacek, M. Mareckova, D. Ulanova, J. Kopecky, V. Chobot, K. Plhackova and J. Olsovska, *J. Chromatogr. A*, 2010, **1217**, 8016–8025.
104 F. Qiu, A. Imai, J. B. McAlpine, D. C. Lankin, I. Burton, T. Karakach, N. R. Farnsworth, S.-N. Chen and G. F. Pauli, *J. Nat. Prod.*, 2012, **75**, 432–443.
105 S. Kildgaard, M. Mansson, I. Dosen, A. Klitgaard, J. C. Frisvad, T. O. Larsen and K. F. Nielsen, *Mar. Drugs*, 2014, **12**, 3681–3705.
106 J. J. J. van der Hooft, R. C. H. de Vos, L. Ridder, J. Vervoort and R. J. Bino, *Metabolomics*, 2013, **9**, 1009–1018.
107 J. Yang, Q. Liang, M. Wang, C. Jeffries, D. Smithson, Y. Tu, N. Boulos, M. R. Jacob, A. A. Shelat, Y. Wu, R. R. Ravu, R. Gilbertson, M. A. Avery, I. A. Khan, L. A. Walker, R. K. Guy and X.-C. Li, *J. Nat. Prod.*, 2014, **77**, 902–909.
108 S. D. Sarker and L. Nahar, *Methods Mol. Biol.*, 2012, **864**, 301–340.
109 C. S. Funari, P. J. Eugster, S. Martel, P.-A. Carrupt, J.-L. Wolfender and D. H. S. Silva, *J. Chromatogr. A*, 2012, **1259**, 167–178.
110 K. T. Johansen, S. G. Wubshet and N. T. Nyberg, *Anal. Chem.*, 2013, **85**, 3183–3189.
111 S. Bertrand, O. Schumpp, N. Bohni, A. Bujard, A. Azzollini, M. Monod, K. Gindro and J.-L. Wolfender, *J. Chromatogr. A*, 2013, **1292**, 219–228.
112 M. Rojas-Cherto, J. E. Peironcely, P. T. Kasper, J. J. J. van der Hooft, R. C. H. de Vos, R. Vreeken, T. Hankemeier and T. Reijmers, *Anal. Chem.*, 2012, **84**, 5524–5534.
113 M. H. Stafsnes, M. Dybwad, A. Brunsvik and P. Bruheim, *Antonie van Leeuwenhoek*, 2013, **103**, 603–615.
114 T. El-Elimat, M. Figueroa, B. M. Ehrmann, N. B. Cech, C. J. Pearce and N. H. Oberlies, *J. Nat. Prod.*, 2013, **76**, 1709–1716.
115 G. T. Carter, *Nat. Prod. Rep.*, 2014, **31**, 711–717.
116 P. C. Dorrestein, *Nat. Prod. Rep.*, 2014, **31**, 704–705.
117 E. Esquenazi, Y.-L. Yang, J. Watrous, W. H. Gerwick and P. C. Dorrestein, *Nat. Prod. Rep.*, 2009, **26**, 1521–1534.
118 G. F. Pauli, S.-N. Chen, D. C. Lankin, J. Bisson, R. J. Case, L. R. Chadwick, T. Godecke, T. Inui, A. Krunic, B. U. Jaki, J. B. McAlpine, S. Mo, J. G. Napolitano, J. Orjala, J. Lehtivarjo, S.-P. Korhonen and M. Niemitz, *J. Nat. Prod.*, 2014, **77**, 1473–1487.
119 M. R. Andersen, J. B. Nielsen, A. Klitgaard, L. M. Petersen, M. Zachariasen, T. J. Hansen, L. H. Blicher, C. H. Gotfredsen, T. O. Larsen, K. F. Nielsen and U. H. Mortensen, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E99–E107.
120 J. Rocha-Martin, C. Harrington, A. D. W. Dobson and F. O'Gara, *Mar. Drugs*, 2014, **12**, 3516–3559.
121 G. D. Wright, *Can. J. Microbiol.*, 2014, **60**, 147–154.
122 Y. Hou, D. R. Braun, C. R. Michel, J. L. Klassen, N. Adnani, T. P. Wyche and T. S. Bugni, *Anal. Chem.*, 2012, **84**, 4277–4283.
123 A. Ibrahim, L. Yang, C. Johnston, X. Liu, B. Ma and N. A. Magarvey, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 19196–19201.
124 A. F. Tawfike, C. Viegelmann and R. Edrada-Ebel, *Methods Mol. Biol.*, 2013, **1055**, 227–244.
125 D. A. Dias, S. Urban and U. Roessner, *Metabolites*, 2012, **2**, 303–336.
126 W. R. Wong, A. G. Oliver and R. G. Linington, *Chem. Biol.*, 2012, **19**, 1483–1495.
127 H. A. Kirst, *Expert Opin. Drug Discovery*, 2013, **8**, 479–493.
128 N. M. Nebane, T. Coric, K. Whig, S. McKellip, L. Woods, M. Sosa, R. Sheppard, L. Rasmussen, M.-A. Bjornsti and E. L. White, *J. Lab. Autom.*, 2013, **18**, 334–339.
129 M. R. Duffy, A. L. Parker, E. R. Kalkman, K. White, D. Kovalskyy, S. M. Kelly and A. H. Baker, *J. Controlled Release*, 2013, **170**, 132–140.
130 M. M. Hann and T. I. Oprea, *Curr. Opin. Chem. Biol.*, 2004, **8**, 255–263.
131 S. C. Desbordes, D. G. Placantonakis, A. Ciro, N. D. Socci, G. Lee, H. Djaballah and L. Studer, *Cell Stem Cell*, 2008, **2**, 602–612.
132 Farrelly *et al.*, Systems and methods for automated proteomics research, US2007184546-A1, CA2533219-A1, 2007.
133 Queeny *et al.*, Droplet detection system, US2008240542-A1, WO2008042960-A2, WO2008042960-A3, US2008240542-A1, WO2008042960-A8, 2008.
134 M. Weber, L. Muthusubramaniam, J. Murray, E. Hudak, O. Kornienko, E. N. Johnson, B. Strulovici and P. Kunapuli, *Assay Drug Dev. Technol.*, 2007, **5**, 117–125.
135 S. Fox, S. Filichkin and T. C. Mockler, *Methods Mol. Biol.*, 2009, **553**, 79–108.
136 D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White and S. Y. Rhee, *Nature*, 2008, **455**, 47–50.
137 M. Eisenstein, *Nature*, 2006, **442**, 1067–1070.
138 L. A. T. Cox, D. Popken, M. S. Marty, J. C. Rowlands, G. Patlewicz, K. O. Goyak and R. A. Becker, *Regul. Toxicol. Pharmacol.*, 2014, **69**, 443–450.
139 X. D. Zhang, *J. Biomol. Screening*, 2010, **15**, 1116–1122.
140 X. D. Zhang, *J. Biomol. Screening*, 2011, **16**, 775–785.
141 X. D. Zhang and Z. Zhang, *Bioinformatics*, 2013, **29**, 794–796.
142 T. Zhang, R. Omar, W. Siheri, S. Al Mutairi, C. Clements, J. Fearnley, R. Edrada-Ebel and D. Watson, *Talanta*, 2014, **120**, 181–190.

143 J. J. Agresti, E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths and D. A. Weitz, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 4004–4009.
144 E. Schonbrun, A. R. Abate, P. E. Steinvurzel, D. A. Weitz and K. B. Crozier, *Lab Chip*, 2010, **10**, 852–856.
145 L. Kimlin, J. Kassis and V. Virador, *Expert Opin. Drug Discovery*, 2013, **8**, 1455–1466.
146 K. Hostettmann, A. Marston and J. L. Wolfender, *Chimia*, 2005, **59**, 291–294.
147 D. Rakshith, P. Santosh, K. Tarman, D. M. Gurudatt and S. Satish, *J. Planar Chromatogr.–Mod. TLC*, 2013, **26**, 470–474.
148 N. Schauer, D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M. G. Forbes, L. Willmitzer, A. R. Fernie and J. Kopka, *FEBS Lett.*, 2005, **579**, 1332–1337.
149 M. Stavri, R. Schneider, G. O'Donnell, D. Lechner, F. Bucar and S. Gibbons, *Phytother. Res.*, 2004, **18**, 774–776.
150 T. Soga, Y. Ohashi, Y. Ueno, H. Naraoka, M. Tomita and T. Nishioka, *J. Proteome Res.*, 2003, **2**, 488–494.
151 M. M. Pedersen, J. C. Chukwujekwu, C. A. Lategan, J. van Staden, P. J. Smith and D. Staerk, *Phytochemistry*, 2009, **70**, 601–607.
152 J.-L. Wolfender, *Planta Med.*, 2009, **75**, 719–734.
153 P. J. Eugster, D. Guillarme, S. Rudaz, J.-L. Veuthey, P.-A. Carrupt and J.-L. Wolfender, *J. AOAC Int.*, 2011, **94**, 51–70.
154 J. R. Yates III, *Nat. Methods*, 2011, **8**, 633–637.
155 Z. Takats, J. M. Wiseman, B. Gologan and R. G. Cooks, *Science*, 2004, **306**, 471–473.
156 A. U. Jackson, A. Tata, C. Wu, R. H. Perry, G. Haas, L. West and R. G. Cooks, *Analyst*, 2009, **134**, 867–874.
157 R. B. Cody, J. A. Laramee and H. D. Durst, *Anal. Chem.*, 2005, **77**, 2297–2302.
158 G. Parsiegla, B. Shrestha, F. Carriere and A. Vertes, *Anal. Chem.*, 2012, **84**, 34–38.
159 J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1743–E1752.
160 H. L. Constant, K. Slowing, J. G. Graham, J. M. Pezzuto, G. A. Cordell and C. W. W. Beecher, *Phytochem. Anal.*, 1997, **8**, 176–180.
161 S. D. Sarker and L. Nahar, *Methods Mol. Biol.*, 2012, **864**, 1–25.
162 D. A. van Elswijk, U. P. Schobel, E. P. Lansky, H. Irth and J. van der Greef, *Phytochemistry*, 2004, **65**, 233–241.
163 B. L. Ackermann, B. T. Regg, L. Colombo, S. Stella and J. E. Coutant, *J. Am. Soc. Mass Spectrom.*, 1996, **7**, 1227–1237.
164 E. C. Tatsis, S. Boeren, V. Exarchou, A. N. Troganis, J. Vervoort and I. P. Gerothanassis, *Phytochemistry*, 2007, **68**, 383–393.
165 P. Waridel, J. L. Wolfender, K. Ndjoko, K. R. Hobby, H. J. Major and K. Hostettmann, *J. Chromatogr. A*, 2001, **926**, 29–41.
166 H. G. Gika, G. A. Theodoridis, J. E. Wingate and I. D. Wilson, *J. Proteome Res.*, 2007, **6**, 3291–3303.
167 J. Tchoumtchoua, D. Njamen, J. C. Mbanya, A.-L. Skaltsounis and M. Halabalaki, *J. Mass Spectrom.*, 2013, **48**, 561–575.
168 A. Klitgaard, A. Iversen, M. R. Andersen, T. O. Larsen, J. C. Frisvad and K. F. Nielsen, *Anal. Bioanal. Chem.*, 2014, **406**, 1933–1943.
169 H. Mohimani, W.-T. Liu, Y.-L. Yang, S. P. Gaudencio, W. Fenical, P. C. Dorrestein and P. A. Pevzner, *J. Comput. Biol.*, 2011, **18**, 1371–1381.
170 P. N. Leao, A. R. Pereira, W.-T. Liu, J. Ng, P. A. Pevzner, P. C. Dorrestein, G. M. Koenig, V. M. Vasconcelosa and W. H. Gerwick, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 11183–11188.
171 M. Baker, *Nat. Methods*, 2011, **8**, 117–121.
172 F. Hufsky, K. Scheubert and S. Bocker, *TrAC, Trends Anal. Chem.*, 2014, **53**, 41–48.
173 S. Stein, *Anal. Chem.*, 2012, **84**, 7274–7282.
174 R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum and J. Lederberg, *Artif. Intell.*, 1993, **61**, 209–261.
175 F. Hufsky, K. Scheubert and S. Böcker, *Nat. Prod. Rep.*, 2014, **31**, 807–817.
176 S. Wolf, S. Schmidt, M. Mueller-Hannemann and S. Neumann, *BMC Bioinf.*, 2010, **11**, 148.
177 M. Gerlich and S. Neumann, *J. Mass Spectrom.*, 2013, **48**, 291–298.
178 F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatos and S. Boecker, *Anal. Chem.*, 2012, **84**, 3417–3426.
179 L. J. Kangas, T. O. Metz, G. Isaac, B. T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R. R. Lewis and J. H. Miller, *Bioinformatics*, 2012, **28**, 1705–1713.
180 M. Heinonen, H. Shen, N. Zamboni and J. Rousu, *Bioinformatics*, 2012, **28**, 2333–2341.
181 J. Gasteiger, W. Hanebeck and K. P. Schulz, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 264–271.
182 N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White and L. Creary, *Anal. Chem.*, 1980, **52**, 1095–1102.
183 E. Luethi, K. T. Nguyen, M. Buerzle, L. C. Blum, Y. Suzuki, M. Hediger and J.-L. Reymond, *J. Med. Chem.*, 2010, **53**, 7236–7250.
184 K. T. Nguyen, S. Syed, S. Urwyler, S. Bertrand, D. Bertrand and J.-L. Reymond, *ChemMedChem*, 2008, **3**, 1520–1524.
185 Z.-L. Deng, C.-X. Du, X. Li, B. Hu, Z.-K. Kuang, R. Wang, S.-Y. Peng, H.-Y. Zhang and D.-X. Kong, *J. Chem. Inf. Model.*, 2013, **53**, 2820–2828.
186 A. M. Virshup, J. Contreras-Garcia, P. Wipf, W. T. Yang and D. N. Beratan, *J. Am. Chem. Soc.*, 2013, **135**, 7296–7303.
187 A. Kerber, R. Laue, M. Meringer and C. Rucker, *J. Comput. Chem., Jpn.*, 2004, **3**, 85–96.
188 C. Corre and G. L. Challis, *Nat. Prod. Rep.*, 2009, **26**, 977–986.
189 R. D. Kersten, Y. L. Yang, Y. Q. Xu, P. Cimermancic, S. J. Nam, W. Fenical, M. A. Fischbach, B. S. Moore and P. C. Dorrestein, *Nat. Chem. Biol.*, 2011, **7**, 794–802.

190 W. T. Liu, J. Ng, D. Meluzzi, N. Bandeira, M. Gutierrez, T. L. Simmons, A. W. Schultz, R. G. Linington, B. S. Moore, W. H. Gerwick, P. A. Pevsner and P. C. Dorrestein, *Anal. Chem.*, 2009, **81**, 4200–4209.
191 H. B. Bode and R. Muller, *Angew. Chem., Int. Ed.*, 2005, **44**, 6828–6846.
192 M. O. Maksimov, I. Pelczer and A. J. Link, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 15223–15228.
193 A. Starcevic, J. Zucko, J. Simunkovic, P. F. Long, J. Cullum and D. Hranueli, *Nucleic Acids Res.*, 2008, **36**, 6882–6892.
194 S. Anand, M. V. R. Prasad, G. Yadav, N. Kumar, J. Shehara, M. Z. Ansari and D. Mohanty, *Nucleic Acids Res.*, 2010, **38**, W487–W496.
195 M. H. T. Li, P. M. U. Ung, J. Zajkowski, S. Garneau-Tsodikova and D. H. Sherman, *BMC Bioinf.*, 2009, **10**, 185.
196 A. de Jong, A. J. van Heel, J. Kok and O. P. Kuipers, *Nucleic Acids Res.*, 2010, **38**, W647–W651.
197 A. J. van Heel, A. de Jong, M. Montalban-Lopez, J. Kok and O. P. Kuipers, *Nucleic Acids Res.*, 2013, **41**, W448–W453.
198 V. Mallika, K. C. Sivakumar, S. Jaichand and E. V. Soniya, *J. Integr. Bioinform.*, 2010, **7**, 143.
199 N. Khaldi, F. T. Seifuddin, G. Turner, D. Haft, W. C. Nierman, K. H. Wolfe and N. D. Fedorova, *Fungal Genet. Biol.*, 2010, **47**, 736–741.
200 T. Weber, C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D. H. Huson and W. Wohlleben, *J. Biotechnol.*, 2009, **140**, 13–17.
201 M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano and R. Breitling, *Nucleic Acids Res.*, 2011, **39**, W339–W346.
202 N. Don Duy, C.-H. Wu, W. J. Moree, A. Lamsa, M. H. Medema, X. Zhao, R. G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, J. Ballesteros, J. Sanchez, J. D. Watrous, V. V. Phelan, C. van de Wiel, R. D. Kersten, S. Mehnaz, R. De Mot, E. A. Shank, P. Charusanti, H. Nagarajan, B. M. Duggan, B. S. Moore, N. Bandeira, B. O. Palsson, K. Pogliano, M. Gutierrez and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E2611–E2620.
203 W.-T. Liu, A. Lamsa, W. R. Wong, P. D. Boudreau, R. Kersten, Y. Peng, W. J. Moree, B. M. Duggan, B. S. Moore, W. H. Gerwick, R. G. Linington, K. Pogliano and P. C. Dorrestein, *J. Antibiot.*, 2014, **67**, 99–104.
204 R. Mendes, M. Kruijt, I. de Bruijn, E. Dekkers, M. van der Voort, J. H. M. Schneider, Y. M. Piceno, T. Z. DeSantis, G. L. Andersen, P. A. H. M. Bakker and J. M. Raaijmakers, *Science*, 2011, **332**, 1097–1100.
205 L.-l. Ooi, *Principles of X-ray Crystallography*, Oxford University Press, 2014.
206 G. M. Sheldrick, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2008, **64**, 112–122.
207 *Models, Mysteries, and Magic of Molecules*, ed. J. C. A. Boeyens and J. F. Ogilvie, University of Pretoria (South Africa) and Universidad de Costa Rica (Costa Rica), Springer, 2008.
208 Y. Inokuma, S. Yoshioka, J. Ariyoshi, T. Arai, Y. Hitora, K. Takada, S. Matsunaga, K. Rissanen and M. Fujita, *Nature*, 2013, **495**, 461–466.
209 K. Wuthrich, *Angew. Chem., Int. Ed.*, 2003, **42**, 3340–3363.
210 R. R. Ernst, *Angew. Chem., Int. Ed.*, 2010, **49**, 8310–8315.
211 G. F. Pauli, T. Goedecke, B. U. Jaki and D. C. Lankin, *J. Nat. Prod.*, 2012, **75**, 834–851.
212 B. U. Jaki, S. G. Frazblau, L. R. Chadwick, D. C. Lankin, F. Zhang, Y. Wang and G. F. Pauli, *J. Nat. Prod.*, 2008, **71**, 1742–1748.
213 G. F. Pauli, B. U. Jaki and D. C. Lankin, *J. Nat. Prod.*, 2005, **68**, 133–149.
214 G. F. Pauli, *Phytochem. Anal.*, 2001, **12**, 28–42.
215 P. Krishnan, N. J. Kruger and R. G. Ratcliffe, *J. Exp. Bot.*, 2005, **56**, 255–265.
216 D. S. Wishart, *TrAC, Trends Anal. Chem.*, 2008, **27**, 228–237.
217 S. Heikkinen, M. M. Toikka, P. T. Karhunen and I. A. Kilpelainen, *J. Am. Chem. Soc.*, 2003, **125**, 4362–4367.
218 E. Kupce and R. Freeman, *Magn. Reson. Chem.*, 2007, **45**, 2–4.
219 M. R. Palmer, B. R. Wenrich, P. Stahlfeld and D. Rovnyak, *J. Biomol. NMR*, 2014, **58**, 303–314.
220 T. F. Molinski and B. I. Morinaka, *Tetrahedron*, 2012, **68**, 9307–9343.
221 K. N. White, T. Amagata, A. G. Oliver, K. Tenney, P. J. Wenzel and P. Crews, *J. Org. Chem.*, 2008, **73**, 8719–8722.
222 N. Matsumori, D. Kaneno, M. Murata, H. Nakamura and K. Tachibana, *J. Org. Chem.*, 1999, **64**, 866–876.
223 Y. Kobayashi, C. H. Tan and Y. Kishi, *J. Am. Chem. Soc.*, 2001, **123**, 2076–2078.
224 S. Fidanze, F. B. Song, M. Szlosek-Pinaud, P. L. C. Small and Y. Kishi, *J. Am. Chem. Soc.*, 2001, **123**, 10117–10118.
225 S. Higashibayashi and Y. Kishi, *Tetrahedron*, 2004, **60**, 11977–11982.
226 S. Higashibayashi, W. Czechtizky, Y. Kobayashi and Y. Kishi, *J. Am. Chem. Soc.*, 2003, **125**, 14379–14393.
227 H. Seike, I. Ghosh and Y. Kishi, *Org. Lett.*, 2006, **8**, 3861–3864.
228 H. Seike, I. Ghosh and Y. Kishi, *Org. Lett.*, 2006, **8**, 5177.
229 H. Seike, I. Ghosh and Y. Kishi, *Org. Lett.*, 2006, **8**, 3865–3868.
230 S. C. Lievens and T. F. Molinski, *Org. Lett.*, 2005, **7**, 2281–2284.
231 J. J. H. Ackerman and J. J. Neil, *NMR Biomed.*, 2010, **23**, 725–733.
232 J. Barbera, L. Puig, P. Romero, J. L. Serrano and T. Sierra, *J. Am. Chem. Soc.*, 2005, **127**, 458–464.
233 Y.-T. Chan, X. Li, J. Yu, G. A. Carri, C. N. Moorefield, G. R. Newkome and C. Wesdemiotis, *J. Am. Chem. Soc.*, 2011, **133**, 11967–11976.
234 Y. Cohen, L. Avram and L. Frish, *Angew. Chem., Int. Ed.*, 2005, **44**, 520–554.
235 T. Evan-Salem, I. Baruch, L. Avram, Y. Cohen, L. C. Palmer and J. Rebek Jr, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 12296–12300.
236 B. Fritzinger, I. Moreels, P. Lommens, R. Koole, Z. Hens and J. C. Martins, *J. Am. Chem. Soc.*, 2009, **131**, 3024–3032.
237 N. Giuseppone, J.-L. Schmitt, L. Allouche and J.-M. Lehn, *Angew. Chem., Int. Ed.*, 2008, **47**, 2235–2239.

238 K. F. Morris and C. S. Johnson, *J. Magn. Reson., Ser. A*, 1993, **101**, 67–73.
239 Q. Zhou, L. Li, J. Xiang, Y. Tang, H. Zhang, S. Yang, Q. Li, Q. Yang and G. Xu, *Angew. Chem., Int. Ed.*, 2008, **47**, 5590–5592.
240 R. Novoa-Carballal, E. Fernandez-Megia, C. Jimenez and R. Riguera, *Nat. Prod. Rep.*, 2011, **28**, 78–98.
241 R. T. Williamson, E. L. Chapin, A. W. Carr, J. R. Gilbert, P. R. Graupner, P. Lewer, P. McKamey, J. R. Carney and W. H. Gerwick, *Org. Lett.*, 2000, **2**, 289–292.
242 M. F. Lin and M. J. Shapiro, *J. Org. Chem.*, 1996, **61**, 7617–7619.
243 H. Barjat, G. A. Morris and A. G. Swanson, *J. Magn. Reson.*, 1998, **131**, 131–138.
244 M. Nilsson, M. A. Connell, A. L. Davis and G. A. Morris, *Anal. Chem.*, 2006, **78**, 3040–3045.
245 M. A. Delsuc and T. E. Malliavin, *Anal. Chem.*, 1998, **70**, 2146–2148.
246 M. Urbanczyk, D. Bernin, W. Kozminski and K. Kazimierczuk, *Anal. Chem.*, 2013, **85**, 1828–1833.
247 A. A. Colbourne, S. Meier, G. A. Morris and M. Nilsson, *Chem. Commun.*, 2013, **49**, 10510–10512.
248 A. A. Colbourne, G. A. Morris and M. Nilsson, *J. Am. Chem. Soc.*, 2011, **133**, 7640–7643.
249 C. F. Tormena, R. Evans, S. Haiber, M. Nilsson and G. A. Morris, *Magn. Reson. Chem.*, 2010, **48**, 550–553.
250 S. Viel, F. Ziarelli and S. Caldarelli, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9696–9698.
251 C. Steinbeck, Computer-Assisted Structure Elucidation, *Handbook of Chemoinformatics: from Data to Knowledge in 4 Volumes*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, pp. 1378–1406.
252 R. Neudert and M. Penk, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 244–248.
253 V. Schutz, V. Purtuc, S. Felsinger and W. Robien, *Fresenius' J. Anal. Chem.*, 1997, **359**, 33–41.
254 C. Steinbeck, S. Krause and S. Kuhn, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1733–1739.
255 M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov and E. D. Smurnyi, *J. Anal. Chem.*, 2008, **63**, 13–20.
256 M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 997–1009.
257 M. Badertscher, A. Korytko, K. P. Schulz, M. Madison, M. E. Munk, P. Portmann, M. Junghans, P. Fontana and E. Pretsch, *Chemom. Intell. Lab. Syst.*, 2000, **51**, 73–79.
258 B. D. Christie and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 87–93.
259 B. D. Christie and M. E. Munk, *J. Am. Chem. Soc.*, 1991, **113**, 3750–3757.
260 A. Korytko, K. P. Schulz, M. S. Madison and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1434–1446.
261 K. P. Schulz, A. Korytko and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1447–1456.
262 T. Lindel, J. Junker and M. Köck, *J. Mol. Model.*, 1997, **3**, 364–368.
263 J. Junker, W. Maier, T. Lindel and M. Kock, *Org. Lett.*, 1999, **1**, 737–740.
264 S. G. Molodtsov, M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. E. Martirosian, G. E. Martin and B. Lefebvre, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1737–1751.
265 K. A. Blinov, M. E. Elyashberg, S. G. Molodtsov, A. J. Williams and E. R. Martirosian, *Fresenius' J. Anal. Chem.*, 2001, **369**, 709–714.
266 M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. R. Martirosian and S. G. Molodtsov, *J. Nat. Prod.*, 2002, **65**, 693–703.
267 M. E. Elyashberg, K. A. Blinov and E. R. Martirosian, *Lab. Autom. Inf. Manage.*, 1999, **34**, 15–30.
268 M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov and G. E. Martin, *J. Chem. Inf.Model.*, 2006, **46**, 1643–1656.
269 A. J. Williams, M. E. Elyashberg, K. A. Blinov, D. C. Lankin, G. E. Martin, W. F. Reynolds, J. A. Porco Jr, C. A. Singleton and S. Su, *J. Nat. Prod.*, 2008, **71**, 581–588.
270 W.-Y. Liao, C.-N. Shen, L.-H. Lin, Y.-L. Yang, H.-Y. Han, J.-W. Chen, S.-C. Kuo, S.-H. Wu and C.-C. Liaw, *J. Nat. Prod.*, 2012, **75**, 630–635.
271 M. Elyashberg, K. Blinov, S. Molodtsov and A. J. Williams, *J. Nat. Prod.*, 2013, **76**, 113–116.
272 S. Kuhn, B. Egert, S. Neumann and C. Steinbeck, *BMC Bioinf.*, 2008, **9**, 400.
273 C. Steinbeck, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1500–1507.
274 K. V. Jayaseelan and C. Steinbeck, *BMC Bioinf.*, 2014, **15**, 234.
275 C. Steinbeck, Y. Q. Han, S. Kuhn, O. Horlacher, E. Luttmann and E. Willighagen, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 493–500.
276 C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E. L. Willighagen, *Curr. Pharm. Des.*, 2006, **12**, 2111–2120.
277 J. M. Nuzillard, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 723–724.
278 Y. Binev, M. M. B. Marques and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2007, **47**, 2089–2097.
279 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.
280 C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan and G. Siuzdak, *Ther. Drug Monit.*, 2005, **27**, 747–751.
281 Y. Sawada, R. Nakabayashi, Y. Yamada, M. Suzuki, M. Sato, A. Sakata, K. Akiyama, T. Sakurai, F. Matsuda, T. Aoki, M. Y. Hirai and K. Saito, *Phytochemistry*, 2012, **82**, 38–45.
282 T. Cheng, Y. Pan, M. Hao, Y. Wang and S. H. Bryant, *Drug Discovery Today*, 2010, **15**, 1052–1057.
283 J. W. Blunt and M. H. G. Munro, *Phytochem. Rev.*, 2013, **12**, 435–447.
284 N. D. Yuliana, M. Jahangir, R. Verpoorte and Y. H. Choi, *Phytochem. Rev.*, 2013, **12**, 293–304.
285 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.

286 P. R. Jensen, K. L. Chavarria, W. Fenical, B. S. Moore and N. Ziemert, *J. Ind. Microbiol. Biotechnol.*, 2014, **41**, 203–209.
287 E. A. Gontang, S. P. Gaudencio, W. Fenical and P. R. Jensen, *Appl. Environ. Microbiol.*, 2010, **76**, 2487–2499.
288 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
289 F. V. Ritacco, B. Haltli, J. E. Janso, M. Greenstein and V. S. Bernan, *J. Ind. Microbiol. Biotechnol.*, 2003, **30**, 472–479.
290 C. Hao, S. Huang, Z. Deng, C. Zhao and Y. Yu, *PLoS One*, 2014, **9**, e99077.
291 L. A. Maldonado, J. E. M. Stach, A. C. Ward, A. T. Bull and M. Goodfellow, *Antonie van Leeuwenhoek*, 2008, **94**, 289–298.
292 N. G. Vynne, M. Mansson and L. Gram, *Mar. Drugs*, 2012, **10**, 1729–1740.
293 D. W. Udwary, L. Zeigler, R. N. Asolkar, V. Singan, A. Lapidus, W. Fenical, P. R. Jensen and B. S. Moore, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 10376–10381.
294 C. Abbet, I. Slacanin, E. Corradi, M. De Mieri, M. Hamburger and O. Potterat, *Food Chem.*, 2014, **160**, 165–170.
295 D. Forner, F. Berrue, H. Correa, K. Duncan and R. G. Kerr, *Anal. Chim. Acta*, 2013, **805**, 70–79.
296 A. Kamleh, M. P. Barrett, D. Wildridge, R. J. S. Burchmore, R. A. Scheltema and D. G. Watson, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 1912–1918.
297 V. Havlicek and K. Lemr, in *Rapid Characterization of Microorganisms by Mass Spectrometry*, ed. C. Fenselau and P. Demirev, 2011, vol. 1065, pp. 51–60.
298 B. R. Stockwell, *Nature*, 2004, **432**, 846–854.
299 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
300 S. Wetzel, R. S. Bon, K. Kumar and H. Waldmann, *Angew. Chem., Int. Ed.*, 2011, **50**, 10800–10826.
301 P. Ertl, S. Roggo and A. Schuffenhauer, *J. Chem. Inf. Model.*, 2008, **48**, 68–74.
302 F. Pereira, D. A. R. S. Latino and S. P. Gaudencio, *Mar. Drugs*, 2014, **12**, 757–778.
303 F. Pereira, D. A. R. S. Latino and S. P. Gaudencio, *Front. Mar. Sci. Conference Abstract: IMMR | International Meeting on Marine Research*, 2014.
304 A. Tropsha, *Mol. Inf.*, 2010, **29**, 476–488.
305 S. Zhang, L. Wei, K. Bastow, W. Zheng, A. Brossi, K.-H. Lee and A. Tropsha, *J. Comput.-Aided Mol. Des.*, 2007, **21**, 97–112.
306 A. Pick, H. Mueller, R. Mayer, B. Haenisch, I. K. Pajeva, M. Weigt, H. Boenisch, C. E. Mueller and M. Wiese, *Bioorg. Med. Chem.*, 2011, **19**, 2090–2102.
307 C. Bissantz, G. Folkers and D. Rognan, *J. Med. Chem.*, 2000, **43**, 4759–4767.
308 A. Bender and R. C. Glen, *J. Chem. Inf. Model.*, 2005, **45**, 1369–1375.
309 Y. Z. Chen and C. Y. Ung, *Am. J. Chin. Med.*, 2002, **30**, 139–154.
310 X. Chen, C. Y. Ung and Y. Z. Chen, *Nat. Prod. Rep.*, 2003, **20**, 432–444.
311 J. M. Rollinger, A. Hornick, T. Langer, H. Stuppner and H. Prast, *J. Med. Chem.*, 2004, **47**, 6248–6254.
312 L. Wang, S. Zhang, J. Zhu, L. Zhu, X. Liu, L. Shan, J. Huang, W. Zhang and H. Li, *Bioorg. Med. Chem. Lett.*, 2014, **24**, 1261–1264.
313 T. Langer and E. M. Krovat, *Curr. Opin. Drug Discovery Dev.*, 2003, **6**, 370–376.
314 T. J. Hou and X. J. Xu, *Curr. Pharm. Des.*, 2004, **10**, 1011–1033.
315 J. M. Rollinger, T. Langer and H. Stuppner, *Curr. Med. Chem.*, 2006, **13**, 1491–1507.
316 K. O. Hanssen, B. Schuler, A. J. Williams, T. B. Demissie, E. Hansen, J. H. Andersen, J. Svenson, K. Blinov, M. Repisky, F. Mohn, G. Meyer, J.-S. Svendsen, K. Ruud, M. Elyashberg, L. Gross, M. Jaspars and J. Isaksson, *Angew. Chem., Int. Ed.*, 2012, **51**, 12238–12241.
317 P. Fechner, O. Bleher, M. Ewald, K. Freudenberger, D. Furin, U. Hilbig, F. Kolarov, K. Krieg, L. Leidner, G. Markovic, G. Proll, F. Proll, S. Rau, J. Riedt, B. Schwarz, P. Weber and J. Widmaier, *Anal. Bioanal. Chem.*, 2014, **406**, 4033–4051.
318 M. J. Booth, *Light: Sci. Appl.*, 2014, **3**, e165.