Briefings in Bioinformatics, 17(2), 2016, 309–321

doi: 10.1093/bib/bbv042
Advance Access Publication Date: 7 July 2015
Paper

Current status and prospects of computational resources for natural product dereplication: a review

Ahmed Mohamed, Canh Hao Nguyen and Hiroshi Mamitsuka

Corresponding author: Ahmed Mohamed, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Japan. Tel.: +81-774-38-3023; Fax.: +81-774-38-3037. E-mail: mohamed@kuicr.kyoto-u.ac.jp

Abstract

Research in natural products has always enhanced drug discovery by providing new and unique chemical compounds. However, recently, drug discovery from natural products is slowed down by the increasing chance of re-isolating known compounds. Rapid identification of previously isolated compounds in an automated manner, called dereplication, steers researchers toward novel findings, thereby reducing the time and effort for identifying new drug leads. Dereplication identifies compounds by comparing processed experimental data with those of known compounds, and so, diverse computational resources such as databases and tools to process and compare compound data are necessary. Automating the dereplication process through the integration of computational resources has always been an aspired goal of natural product researchers. To increase the utilization of current computational resources for natural products, we first provide an overview of the dereplication process, and then list useful resources, categorizing into databases, methods and software tools and further explaining them from a dereplication perspective. Finally, we discuss the current challenges to automating dereplication and proposed solutions.

Key words: natural products; dereplication; NMR; compound identification

Introduction

Natural products have been a precious resource for drug discovery and lead identification [1–3]. In all, 75% of all Food and Drug Administration-approved small molecules are either natural compounds or derivatives therefrom [4]. The potential of natural products in drug discovery can be attributed to their unique structural scaffolds and high complexity, creating diverse biological screening libraries [5]. Besides being attractive drug leads, the complexity of natural products and high content of stereogenic atoms increase protein binding selectivity [6], allowing natural products to be used in ligand design, particularly fragment-based drug design [7].

Despite the potential of natural products, there are two main factors that limit their role in recent drug discovery and lead identification research: i) time-consuming identification of active compounds: The general manner of experimental design for identifying natural products remained unchanged throughout the past decades. That is, it requires time-consuming purification and inefficient manual interpretation of compound nuclear magnetic resonance (NMR) spectra by experts. ii) Repetitive effort for identifying known compounds. While it is estimated that more than 250,000 natural compounds have already been isolated [8, 9], incorporation of such knowledge to enhance drug discovery is still not fully exploited.

To overcome these two factors, one promising approach is dereplication, which is the early identification of known compounds without time-consuming manual structure elucidation [10, 11]. Putative compounds are obtained by comparing preliminary spectral data to spectral databases of known compounds (This review mainly focuses on NMR spectra, while methods, software and databases of NMR spectra can be applied

Ahmed Mohamed obtained his MSc in natural products chemistry and is currently a PhD student at the Bioinformatics Center in Kyoto University. His current research focuses on developing software applications for biology and chemistry.
Canh Hao Nguyen is an assistant professor at the Bioinformatics Center, Institute for Chemical Research, Kyoto University, working on machine learning for graph data, with applications to biological networks.
Hiroshi Mamitsuka is a professor at the Bioinformatics Center, Institute for Chemical Research, Kyoto University, working on machine learning, data mining and application to biology and chemistry.
Submitted: 27 April 2015; **Received (in revised form):** 27 May 2015

to other types of spectra, such as mass spectrometry [MS]). Early detection of known compounds and their reported and potential biological activities help researchers to focus their efforts toward novel findings [12]. While the idea of dereplication is decades old [13], it has gained more attention recently with the increased sensitivity in analytical instruments [11], which allows structure elucidation at nanomole scales [14–16]. In addition, coupling of ultrasensitive instrument such as capillary NMR and high-resolution MS with chromatography allows preisolation compound identification [17–19], which significantly reduces time and effort.

Despite instrumental advances that are useful for compound identification, computational tools for dereplication are still at a developing stage. Fortunately, natural products and metabolomics share common compound identification techniques, and they are said to be 'two sides of the same coin' [20]. Focusing on detecting dynamic metabolite changes in biological fluids, research in metabolomics spurred simultaneous development of accurate computational methods for fast and high-throughput identification of compounds from complex biological mixtures. However, the small but significant differences between natural products and metabolomics prevent the direct cross-utilization of computational resources.

Table 1 shows the differences between compound identification in natural products and metabolomics. From data perspectives, there are particularly three key differences: (1) Natural products reference libraries are larger in size than those of metabolomics, increasing the computational demand to search through these libraries, and the lower quality of spectral data poses concern on the reliability of results. (2) Compound identification in metabolomics relies on 'landmark' peak detection [27], often obtainable from proton-based NMR spectra such as proton nuclear magnetic resonance (^{1}H) and total correlation spectroscopy [28, 29]. However, owing to structural diversity and spectral complexity of natural products, the identification of natural products often requires inclusion of carbon-based NMR measurements, such as carbon-13 nuclear magnetic resonance (^{13}C) and heteronuclear single quantum coherence spectra [20, 22, 23]. (3) Metabolomics samples are complex biological mixtures where the goal is to both identify and quantify metabolites. However, quantitative analysis of mixtures is not the current focus of dereplication.

We review the current status of computational resources that are or could be used as building blocks to automate dereplication and how they can fit in the current experimental design. We discuss the overlaps and differences in computational demands of dereplication and compound identification in metabolomics. We start by a brief overview of the experimental design of dereplication, followed by detailed discussion on three computational aspects of dereplication: databases, methods and software. We finally conclude with future perspectives.

Overview of natural products compound identification

Figure 1 shows compound identification in natural products without and with dereplication. The standard experimental design for natural product identification starts with purification of bioactive compounds using bioassay-guided fractionation from natural extracts (Figure 1IA, IB). Measured full spectral data of the purified compounds are manually interpreted for deducing the compound structure (Figure 1IC, ID), which is then used for literature inquiry (Figure 1IE). With the increasing chance of isolating known compounds, the time and cost are becoming unacceptable. Dereplication utilizes prior knowledge of previously isolated compounds for early identification to minimize human intervention. Ideally, preliminary experimental data, such as source organism, bioactivity and measured spectra, are used to filter compounds that are either previously reported or lacking drug-like characteristics.

For researchers to integrate dereplication in their experimental design, they need a full software suite for automatic NMR processing and analysis that is linked to a reference database for dereplication. The reference database should provide a wide coverage of previously isolated natural compounds with their source organisms and reported/predicted bioactivities. A database query should be carried out with a sophisticated method for compound matching integrating different types of spectral information.

Three components are needed to develop a complete dereplication software: i) databases to act as reference libraries, ii) spectral processing and searching methods to query databases and iii) software tools for spectral preprocessing and analysis. We discuss each component, identifying available resources and their current shortcomings where further research is needed. In the next section, we introduce available databases, discussing their coverage, deposited data and relevant query options. In the fourth section, we describe different methods as well as software tools for spectral preprocessing and compound identification.

Databases

The integration of chemoinformtics modeling in drug design motivated the development of numerous databases listing chemical compounds with their biological and physical properties. Databases relevant to natural products are already reviewed [30–32], while we discuss them here from a dereplication perspective. We divide available databases into general and natural product-specific databases (Tables 2 and 3, respectively), and score each database with seven criteria that are important for dereplication: (1) coverage of known natural compounds, (2) availability of bioactivity data, (3) availability of

Table 1. Differences between compound identification in natural products research and metabolomics

Comparison criterion	Natural products research	Metabolomics
Reference library size	Large (>250,000) [8]	Small (few 1,000 s) [21]
Quality of reference spectra	Low [20]	High [20]
Types of spectra	Both proton and carbon-based [22, 23]	Mainly proton-based [24]
Structural complexity	Complex [25, 26]	Simple
Sample purity	Purified or semi-purified compounds [20]	Complex biological fluid mixtures [20, 24]
Spectral comparison	Pairwise	Pairwise or multiple (time-series)
Overall goal	Compound identification	Compound identification and quantification

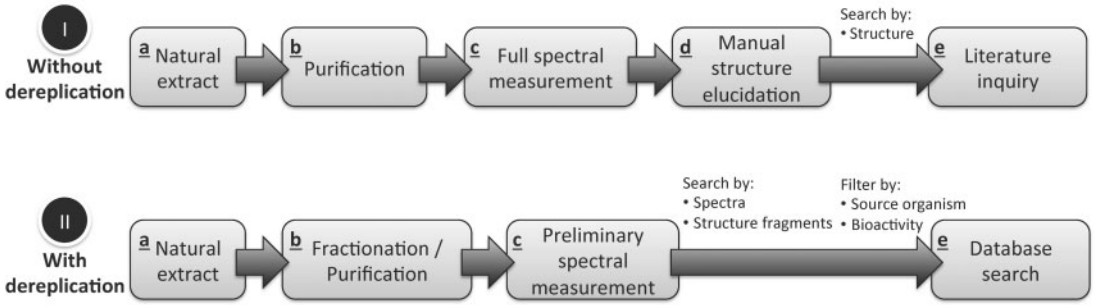

Figure 1. Compound identification in natural products without and with dereplication.

Table 2. General chemical databases

Database	Web site (*http://*)	Coverage		Data content		Spectral searchability	Programmatic access	Free?		Score
		# NPs	# Compounds	Bioactivity (type)	Source organism			Use	Download	
BindingDB [33]	*www.bindingdb.org*	NA	>450 k	•(protein binding)			•	•	•	5
ChEBI [34]	*www.ebi.ac.uk/chebi/*	>25 k	>42 k	•(all)	•		•	•	•	5
ChemBank [35]	*chembank.broadinstitute.org*	NA	>800 k	•(all)			•	•	•	5
Chembl [36]	*www.ebi.ac.uk/chembl/*	24K	>600 k	•(all)			•	•	•	5
ChemIDplus	*chem.sis.nlm.nih.gov/chemidplus/*	>9 k	>400 k	•(all)				•		2
ChemSpider [37]	*www.chemspider.com*	>660 K	>14 M	•(all)			•	•	•	5
CSEARCH [38]	*nmrpredict.orc.univie.ac.at/*	NA	>450 k			•		•		3
NCI	*cactus.nci.nih.gov/ncidb2.2/*	NA	>250 k	•			•	•	•	5
NIAID ChemDB	*chemdb.niaid.nih.gov*	>9 k	>130 k	•(allergy, infectious diseases)				•		2
NMRShiftDB [39]	*nmrshiftdb.nmr.uni-koeln.de*	NA	>42 k			•		•	•	4
PubChem [40]	*pubchem.ncbi.nlm.nih.gov*	NA	>30 M	•(all)			•	•	•	5
Reaxys [41]	*www.reaxys.com/reaxys*	>200 k	>10 M	•(all)	•		•			4
SciFinder	*scifinder.cas.org*	NA	>90 M	•(all)			•			3
SpecInfo [42]	*www.wiley-vch.de/stmdata/specinfo.php*	3.5 k	>500 k			•				1
ZINC [43]	*zinc.docking.org*	>180 k	>20 M					•	•	3

#NPs: Number of natural product compounds.

Table 3. Natural products-specific databases

Database	Web site (*http://*)	Coverage	Data content		Spectral searchability	Programmatic access	Free?		Score
		# Compounds	Bioactivity	Source organism			Use	Download	
AntiBase [44]	*www.wiley-vch.de/stmdata/antibase.php*	>40 k	•(all)	•	•				4
BACTIBASE [45]	*bactibase.pfba-lab-tun.org*	220	•(all)	•			•	•	4
CamMedNP [46]	NA	2.5 k		•			•	•	3
ConMedNP [47]	NA	3.2 k		•			•	•	3
Dictionary of marine NP	*dmnp.chemnetbase.com*	>30 k	•(all)	•					3
Dictionary of NP	*dnp.chemnetbase.com*	>250 k	•(all)	•					3
HeteroCycles	*www.heterocycles.jp/newlibrary/natural_products/structure*	>58 k	○(anti-microbial)	•			•		4
Marinlit	*pubs.rsc.org/marinlit/*	>24 k	•(all)	•	•				4
NAPROC-13 [48]	*c13.usal.es*	>20 k			•		•		3
NPACT [49]	*crdd.osdd.net/raghava/npact/*	1574	•(anti-cancer)				•		2
NuBBE [50]	*nubbe.iq.unesp.br/portal/nubbedb.html*	640	•(anti-microbial)	•			•	•	4
PhytAMP [51]	*phytamp.pfba-lab-tun.org*	273	•(anti-microbial)	•			•	•	4
SuperNatural [52, 53]	*bioinformatics.charite.de/supernatural*	>350 k	•(all)				•		3
TCM database [54]	*tcm.cmu.edu.tw*	>20 k	•(traditional Chinese medicine)	•			•	•	5
UDNP [55]	*pkuxxj.pku.edu.cn/UNPD*	230 k		•			•	•	4

○: Limited data.

source organism data, (4) searchability over compounds by measured compound spectra, (5) programmatic access through Web services or application programming interfaces (APIs), (6) free availability to use and (7) free availability to download. Tables 2 and 3 demonstrate that no available databases satisfy all seven criteria for an ideal dereplication database. Below, we discuss these databases in terms of coverage, data content, spectral searchability and access.

General databases

We include 15 chemical databases as general databases according to the following criteria: (i) Cover more than 10% of already isolated natural products; around 20,000 compounds. (ii) Contain at least 40,000 entries including both synthetic and natural compounds. (iii) Contain information useful in dereplication, such as bioactivity, source organism or spectra (Table 2).

Regarding coverage, five databases contain more than 10 million entries. General databases provide wide coverage of natural compounds, with 11 databases containing more than 20,000 natural compounds (roughly 10% of already isolated compounds). Despite their wide coverage, searching is not easy to use for dereplication because synthetic compounds are among search candidates. Seven databases have natural compounds annotation, which allows users to limit their search to natural products only.

Dereplication-relevant data-contents are two: bioactivity and source organism. Eleven databases include biological activity. PubChem [40], ChEBML [36] and BindingDB [33] databases contain detailed bioactivity information such as biological mechanism and protein targets, which can be used, in conjunction with spectral information, to enhance compound identification [56]. Regarding source organism, only two databases, ChEBI [34] and Reaxys [41], contain this information.

While spectral searchability is important in dereplication, searching compounds by spectral data is not the focus of general databases, and only NMRShiftDB [39], CSEARCH [38] and SpecInfo [42] have this ability. Compounds in all 15 general databases are searchable by similarity of structures or substructures; however, this search has strong limitations for dereplication, where molecular structures are unknown.

There are three ways to access general databases: (1) manual access, (2) access via database download or (3) programmatic access. Twelve databases can be accessed manually for free and nine of them are freely downloadable. Ten databases provide APIs to access the data though programs, which enable their integration to user-customized analysis flows. However, programmatic access has limitations for dereplication because either necessary data or query options are lacking.

Natural products-specific databases

We raise 15 databases that catalog molecules isolated from natural origins only, excluding those limited to primary metabolites, as those are relevant only to metabolomics (Table 3). In terms of coverage, nine specific databases exceed 20,000 entries. Because of the coverage limitation, it is better to use multiple specific databases for reliable dereplication. Some specific databases have limited coverage because they focus on: i) particular compound features such as compound class (PhytAMP [51] and BACTIBASE [45]) or bioactivity (NPACT [49]), and ii) particular compound origins such as compounds from a particular family of source organisms (CamMedNP[46] and ConMedNP [47]) or geographic location (NuBBE [50] and TCM [54]).

Despite their limited coverage, specific databases contain bioactivity and source organism information, useful in dereplication. Eleven specific databases contain bioactivity data. Typical examples are NuBBE [50] and NPACT [49], which provide effective compound concentrations of different bioactivities for each entry. All specific databases have source organism information, except for SuperNatural [52, 53], NPACT [49] and NAPROC-13 [48].

Spectral searchability is limited in specific databases because of the scarcity of spectral data. Only three databases have spectral searchability, and only one database, NAPROC-13 [48], is freely accessible but limited to ^{13}C spectra only.

Regarding database access, 11 specific databases can be manually searched, seven of which are freely downloadable. Specific databases are usually in-house developed and all of them do not provide programmatic access to the data, limiting automatic search and integration to other software.

Methods and software

This section describes computational methods and software tools used as parts of natural product dereplication process. Table 4 summarizes two main steps of dereplication: spectral preprocessing, and compound identification. First, spectral preprocessing involves reformatting and denoising of the acquired spectra to alleviate the instrumental and experimental discrepancies [27, 57]. Second, compound identification uses preprocessed spectra and compares them to a reference database. To realize automatic and fast dereplication, each step needs to be carried out efficiently with minimal human intervention. Table 5 lists, to the best of our knowledge, currently available software tools for these steps, comparing the tools according to functionalities.

Note that while we focus here on software for spectral preprocessing and compound identification, natural product dereplication needs additional tools to manage and visualize chemical structures and spectra. For example, structures of chemical compounds are usually represented as SDF or MOL files, and software tools, such as Open Babel toolbox [69] and ChemmineR [70], rcdk [71] and Rcpi [72] R packages, are needed to handle these files and pass the data to the dereplication software for processing or visualization. For result visualization, Java and javascript libraries, such as JSpecView [73], JSME [74] and MarvinJS [75], can offer in-browser chemical structure and spectral visualization for Web applications.

Spectral preprocessing

We categorize preprocessing methods into three main steps: file format conversion, baseline correction and alignment. We first discuss each step, and demonstrate baseline correction and alignment on example spectra (^1H NMR spectrum of stigmasterol in Figure 2). We finally summarize available software tools.

File format conversion

While the acquired spectra are initially stored as proprietary data formats that are specific to each instrument, converting them to a common instrument-independent format ensures easier data exchange and wider compatibility. For NMR spectra, JCAMP-DX [76], NMRPipe [65] and Sparky [77] are among the most used file formats for describing spectral information of small molecules. JCAMP-DX [76] provides a simple and human-readable format, and allows additional *labels* to describe

Table 4. Analysis flow of spectra from acquisition to compound identification

Spectral preprocessing	File format conversion	• JCAMP-DX • NMRPipe • Sparky
	Baseline correction	1. Baseline recognition • Derivative functions • Wavelet-based 2. Baseline modeling • Polynomial • Regression • Smoothing 3. Baseline subtraction
	Alignment	• FFT alignment • Multiple-dimension
Compound identification	Data reduction	❖ Peak lists • Peak picking ○ Numerical vectors • Binning • Feature extraction ○ Sliding window ○ PCA ➢ Trees
	Spectral comparison	❖ Peak lists • Tanimoto coefficient • Jaccard similarity ○ Numerical vectors • Correlation-based ○ Dot product ○ Pearson's correlation ○ Spearman's correlation ○ Weighted cross-correlation ○ Partial and semi-partial correlation • Distance-based ○ Absolute value distance ○ Euclidean distance ➢ Trees • Tree-based comparison
	Database search	• Identity search • Ranking search • Interpretative search

experimental conditions and parameters. However, representation of multi-dimensional NMR spectra in JCAMP-DX is not standardized. NMRPipe [65] and Sparky [77] have been used in Web applications [78, 79] for their strong standardization and the ability to represent multi-dimensional NMR spectra.

Current NMR file formats mainly have the following three limitations for dereplication. First, current file formats do not contain structures of measured compounds, which prevent assigning spectral peaks to corresponding atoms. We have to include additional files for compound structure and peak assignment information [80, 81], which cannot be linked easily with spectral files. Second, one-dimensional (1D) NMR and two-dimensional (2D) NMR spectral data of the same sample cannot be linked with each other in current file formats. Third, current file formats are still insufficient to fully represent measurements and experimental parameters in high-throughput studies [80]. CCPN [82, 83] and STAR [84–86] provide different formats that can be used for high-throughput studies, but are tailored for protein NMR experiments. A suitable file format for natural product dereplication is still needed to overcome the above three limitations.

Baseline correction

Removal of baseline drifting is crucial to remove noise and artifacts resulting from different measurement conditions. Generally, baseline correction has three steps, baseline recognition, modeling and subtraction. First, baseline recognition distinguishes peak regions from baseline points, exploiting the fact that peak regions have higher variation in intensity. Higher variation regions are detected using spectrum derivatives [87] or wavelet transformation [88–91]. Second, baseline modeling estimates a curve based on baseline points, by linear interpolation or non-linear approximations like polynomial fitting [92, 93], LOcally Weighted Scatterplot Smoothing (LOWESS) and quantile regressions [94–98] and Whittaker smoother [88, 99]. Finally, in baseline subtraction, the estimated baseline curve is subtracted from the spectrum, leaving only the peak signals.

In natural product dereplication, baseline correction is a minor step compared to metabolomics because of two main differences: (i) As dereplication currently focuses on compound identification rather than quantification, accurate baseline estimation is less significant [20]. (ii) Dereplication is usually performed on purified compounds where spectra are less crowded than those of biological mixtures. Therefore, simple polynomial fitting is usually preferred for baseline correction, instead of more computationally demanding techniques such as LOWESS and quantile regressions and Whittaker smoother. In our example, the baseline is estimated as a third-order polynomial function (Figure 2).

Alignment

Alignment of spectra is a process to alleviate the effect of experimental conditions on peak positions by shifting data points to match a reference spectrum [24, 57]. Spectral alignment and relevant software tools are already reviewed in detail [57], and so we only describe alignment here briefly. Alignment is performed for quantitative comparison between multiple spectra of different samples that have similar chemical compositions, and therefore, it is a standard manner for time-series NMR spectra in metabolomics. Using the same concept, alignment can be applied in dereplication when spectra for different fractions of the same extract are compared [100]. Figure 2 shows how alignment removes subtle chemical shift differences in the spectra of two structurally similar compounds, cholesterol and stigmasterol, increasing the overall similarity between the two spectra.

Software summary for spectral preprocessing

Table 5 shows that out of 16 currently available software, 3 steps in spectral preprocessing, i.e. file format conversion, baseline correction and alignment, are implemented in 12, 13 and 9 software tools, respectively, meaning that baseline correction is the most implemented. Six software tools, ACD Labs, Automics [58], Chenomx NMR suite, MestreNova, MVAPACK [62] and PERCH, implement all three steps, of which Automics and MVAPACK are freely available, making them most useful for spectral preprocessing. Six other tools implement two steps, and the remaining four tools (all are R packages) specialize in only one step.

Table 5. Software tools with a potential role in dereplication

Software	Software type	Spectra type	GUI	Spectral preprocessing			Compound identification			Free?	Score
				File Format Conversion	Baseline Correction	Alignment	Peak Picking	Binning	Feature Extraction		
ACD Labs	Desktop	NMR (1D, 2D), MS	•	•	•	•	•				4
Automics [58]	Desktop	NMR	•	•	•	•	•	•	•	•	7
BATMAN [59]	R package	NMR			•	•		•		•	4
ChemoSpec [60]	R package	Any			•		•			•	3
Chenomx NMR suite	Desktop	NMR (1D, 2D)	•	•	•	•	•	•			5
cuteNMR	Desktop	NMR	•	•	•		•			•	4
MestreNova	Desktop	NMR (1D, 2D), MS	•	•	•	•	•				4
mSPA [61]	R package	Any				•				•	2
MVAPACK [62]	Octave package	NMR (1D, 2D)		•	•	•	•	•	•	•	7
mylims.org [63]	Web	NMR, MS		•	•		•			•	4
Nmrglue [64]	Python package	NMR (1D, 2D)		•	•		•			•	4
Nmrpipe [65]	Desktop	NMR		•	•		•	•		•	5
NMRS [66]	R package	NMR		•						•	2
PERCH	Desktop	NMR (1D, 2D)	•	•	•	•	•				4
rnmr [67]	R package	NMR (2D)	•	•	•		•	•		•	5
speaq [68]	R package	NMR				•	•			•	2

GUI: Graphical User Interface.

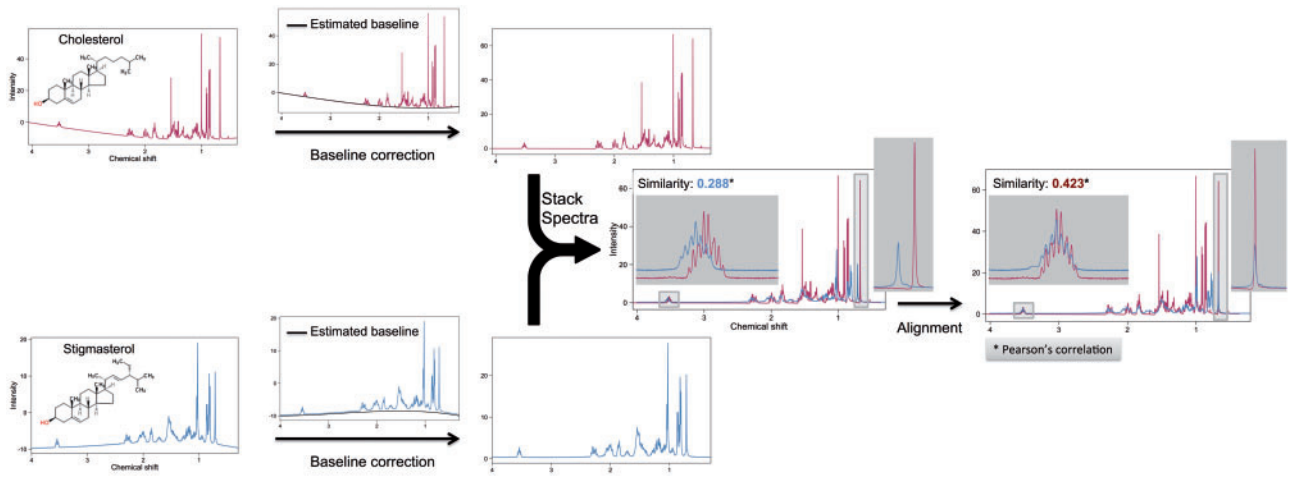

Figure 2. Spectral preprocessing. ^{1}H NMR spectra of cholesterol and stigmasterol, two common and structurally similar natural compounds, are used for demonstration. The raw NMR files were downloaded from HMDB [21] and converted to JCAMP-format DX using Mestrenova. Baseline estimation was performed in R using third-order polynomial fitting. The baseline-corrected spectra were stacked, and then aligned using Mestrenova, showing higher similarity (Pearson's correlation of 0.423) than before alignment (0.288). A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

Compound identification

For compound identification, preprocessed spectra are converted into different representations to be compared against reference spectra in a computationally efficient manner, to find compounds with the highest spectral similarity. To carry out compound identification, three steps are required: data reduction, spectral comparison and searching databases. We explain each of these three below.

Data reduction

Spectral comparison of raw spectra needs long computation time because each spectrum has a large number of data points (more than 20,000 points for ^{1}H NMR [24]), where each point has a position (chemical shift) and an intensity. To reduce computation time, we need methods to reduce data size without substantial loss of information. Data reduction transforms spectral data into peak lists, numerical vectors or trees. We describe the characteristics of each of these three representations.

Peak lists: Spectra are reduced to peak lists by peak picking [101–103], which greatly simplifies the spectra to a handful of peak positions and their intensities. Limitations of peak picking arise if the spectrum contains broad or overlapped peaks, such as crowded ^{1}H NMR spectra, in which important peaks can be missed.

Numerical vectors: Spectra can be reduced to numerical vectors of the same size by binning, sliding window or principal component analysis. First, binning [104–106] divides the spectrum into intervals and the total intensity in each interval is

extracted. While binning keeps representative information about the spectrum, a peak may be split into two bins if the bin boundary lies on a peak center, which misrepresents the peak as shown in Figure 3-2A. So, adaptive binning changes bin boundaries to prevent overlap with peak centers [104, 106]. Second, sliding window divides the spectrum into fix-sized but overlapped intervals [107]. Third, principal component analysis reduces spectra by transforming the original data space into a lower dimension space [108].

Trees: A spectrum is transformed into a tree by assigning peaks to end nodes through recursively dividing the spectrum into subspectra at mass centers [109, 110] (Figure 3-3A, B). The resulting tree has spectra mass centers as branching nodes and peaks as end (leaf) nodes, which retains information about peak positions as well as their hierarchy.

Two factors are important in data reduction for natural product dereplication: i) Type of measured spectra: NMR spectra vary in how sharp peaks are and the propensity for peaks to overlap. Sharp peaks in ^{13}C NMR spectra are unlikely to overlap and so peak lists are suitable. In contrast, ^{1}H NMR peaks tend to heavily overlap, especially that of complex mixtures and in condensed methylene regions, and so binning or trees are preferred. ii) Spectral comparison measure suitable for the representation (described in the next section).

Spectral comparison

Spectra are compared using a similarity measure that reflects the structure similarity of the corresponding compounds. The choice of the similarity measure depends on the data representation, determined by the data reduction method (described in the previous section). We discuss available similarity measures for each representation.

Peak lists: When spectra are reduced to peak lists, which are of different sizes, they are represented as sets. Comparing two sets of peaks requires two steps: (1) Matching of set members, to produce one-to-one mappings between peaks of the query and reference sets. First, a list of matching candidates for each peak is narrowed to peaks whose positions lie within a defined threshold. A threshold can be either a fixed window (hard thresholding), which is chosen manually, or defined statistically using Bayesian [111, 112] and probability-based [113, 114] models (soft thresholding), which are more flexible. Second, matching peaks are chosen from the candidate list by either i) selecting the nearest peak, or ii) maximum bipartite matching [115], which maximizes the number of pairs between peaks of the two sets. (2) Measuring the overlap between two sets, which is computed by set similarity measures, typically Jaccard's similarity and Tanimoto's coefficient [116–118].

Numerical vectors: Numerical vectors have the same dimension and a typical way to compare them uses a correlation or distance-based similarity measure such as inner product [119–121], Euclidean distance [122–124] or difference in absolute value [125, 126]. Among the three measures, inner product was reported to outperform the other two measures [127]. Measures combining both correlation and distance-based similarities, such as partial correlation [128] and composite similarity measures [129], have been shown to perform better than a single measure [130, 131].

Trees: Trees are compared by taking into account both peak positions (node position) and their hierarchy (children nodes) [109, 110].

Computational efficiency: Applying spectral comparison to large databases requires efficient computation of similarity scores. The speed of computing similarity scores for each representation is affected by two factors: (1) number of data points in spectral representation and (2) computational complexity of comparing two spectra. First, number of data points varies between spectra owing to different spectral features or data reduction parameters. The number of data points in peak lists and trees depends on the number of peaks, while in numerical vectors, it is equal to the number of bins. The typical size of natural product compound spectra is tens for peak lists, and 250 for numerical vectors (chemical shift range: 0∼10 ppm, bin size: 0.04 ppm). Second, the computational complexity is determined by the number of computational operations for comparing two spectra of N data points, which is in the order of N^2 for peak lists, and N and $N\,logN$ for numerical vectors and trees, respectively. Theoretically, for a similar number of data points, numerical vectors are the fastest to compare, followed by trees and peak lists.

Searching databases

Three database search paradigms are useful in dereplication: (1) identity, (2) ranking and (3) interpretative; each search paradigm produces a different output format [127, 132]. We explain each paradigm below.

Identity search: Identity search returns a single compound with a spectrum that is equivalent to the query spectrum [111, 127]. Identity search requires no manual investigation and so it can be very useful in automating dereplication. However, identity search has two limitations: (1) Searching small-coverage databases may return empty results, if the exact spectrum is not in the database. (2) Setting strict equivalence criteria may miss spectra that are affected by variations in experimental conditions or inadequate preprocessing.

Ranking search: Ranking search returns a ranked list of compounds with spectra closest to that of the query by computing similarity scores the query spectrum and all spectra in the database [125]. By investigating common substructures of highly similar compounds in the list, we can deduce chemical class or functional groups of the query compound. Similarity scores can also be computed using a subset of the query spectrum, allowing users to focus on distinctive peaks. One limitation for ranking search is that deducing chemical classes and functional groups still requires manual investigation, which hampers automatic dereplication.

Interpretative search: Interpretative search returns a list of matching fragments by assigning peaks from the query spectrum to connected fragments of reference compounds [115, 133]. The output fragments, which belong to different reference compounds, can then be combined to deduce the query compound structure and so interpretative search can identify novel compounds that are not included in the reference database. Currently, interpretative search is not applicable to ^{1}H spectra because of the sensitivity of chemical shifts to spatial interactions [115] and because peak overlap prevents spectral peaks to be assigned to corresponding atoms.

Software summary for compound identification

For data reduction, we focused on three spectral representations: i) peak lists obtained by peak picking, ii) numerical vectors obtained by binning and feature extraction and iii) trees. Table 5 shows that peak picking is the most implemented method, available in 13 out of 16 software tools, followed by binning and then feature extraction, available in six and two tools, respectively.

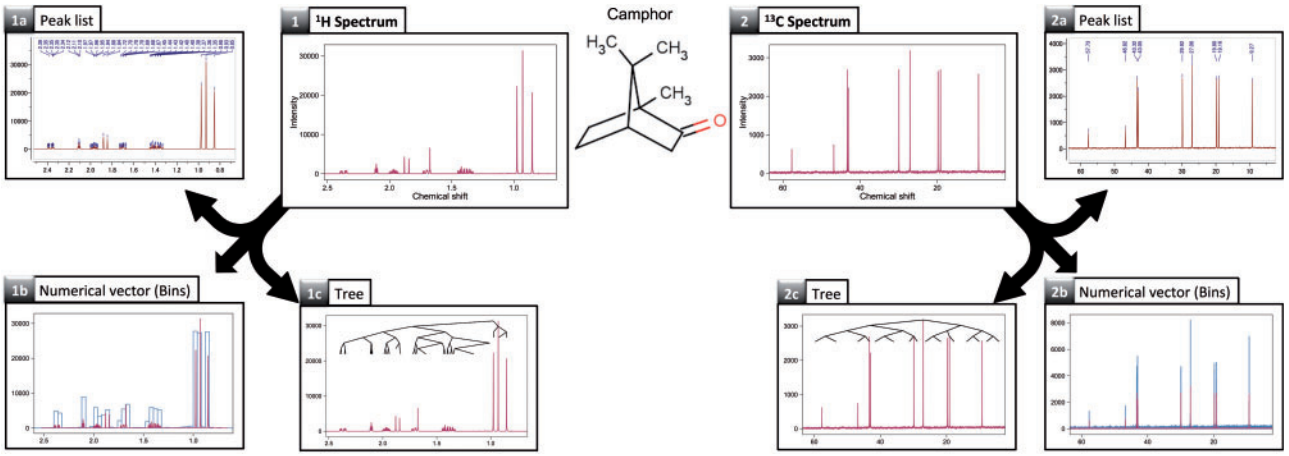

Figure 3. Data reduction of spectra. ^{1}H and ^{13}C NMR spectra of camphor, a natural compound, demonstrate the effect of each data reduction method on different types of spectra. Peak picking reduces the ^{13}C spectrum to a few peaks (2A), but fails with the ^{1}H spectrum (1A), as resonance coupling generates numerous overlapping multiplet peaks. Binning produces in a large vector (1532 bin) in the ^{13}C spectrum and a small one in the ^{1}H spectrum (47 bins). Both spectra are reduced to relatively few nodes when represented as trees. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

No software tools implement tree representation of spectra; however, the pseudocode is available [109]. Automics [58] and MVAPACK [62] are the only tools implementing the three data reduction methods. rNMR [67], NMRPipe [65] and PERCH have both peak picking and binning functionalities.

Spectral comparison methods, such as inner product and partial correlation, are available in statistical software frameworks, such as R and Matlab.

Finally, among database search paradigms, ranking search is implemented in spectral databases, such as NMRShiftDB [39] and CSEARCH [38], because chemical class or functional groups can be deduced by investigating the ranked compound list.

V- Future perspectives

Despite the abundance of computational resources that are useful for dereplication, we need to overcome several challenges to realize the aspired automation. We discuss four proposed solutions to existing challenges that can enhance the speed and quality of natural products dereplication results.

Enriching databases using automated machine leaning methods

The deficiency of necessary data, namely, measured spectra and source organisms, presents a challenge to the development of a dereplication database, being summarized into two points: (1) The scarcity of measured spectra prevents spectral searchability from producing reliable results. (2) The absence of source organisms data prevents their use to limit dereplication candidates. Two machine learning-derived approaches will provide a fast and automated way to add data to databases and complete missing data: spectral prediction and literature text mining. First, compound spectra can be predicted from existing spectra on the basis of compound structural similarity [134]. Several machine learning algorithms have been proposed to predict NMR spectra [135–137], of which prediction accuracy increases with training data size [138]. Similar algorithms have also been developed for other types of spectra, such as fragmentation pattern in MS spectra [139–141], ultraviolet spectra (UV) [413] and chromatographic retention index [143–145]. Comparison and accuracy assessment of NMR prediction algorithms are reviewed in [146–148]. Second, text mining of chemical information [149, 150] can automatically extract compound associated data such as NMR assignments and source organisms from the literature.

Developing software suite from building blocks

The wide use and integration of dereplication to current experimental design is hampered by the unavailability of open-source software to process NMR spectra, to link and to summarize information across all submitted spectra. While all steps for dereplication are implemented in software packages (Table 5), the dereplication process requires the use of different tools and familiarity of programming languages. To accelerate dereplication, a software suite combining available software packages through a unified graphical interface that can be used intuitively by experimental researchers on natural products is needed.

Integrating different spectral types

Relying on NMR data only for compound identification becomes insufficient as molecular complexity [151, 152] increases, as exemplified by fatty acids and peptides. The integration of different spectral data into dereplication can resolve structural ambiguities in these chemical classes. Several studies in dereplication showed promising results by integrating MS fragmentation with UV spectra [153, 154], and in combination with NMR spectra [155, 156]. However, current studies have two limitations: (1) Other spectral types, such as chromatographic retention times, can differentiate between compounds that are otherwise similar. While these spectra utilized in metabolomics [61, 128, 157], they are not yet incorporated in dereplication. (2) Similarity scores between query and database compounds are calculated based on only one spectral type, and candidate structures are then filtered using the other spectra. Calculating similarity scores based on all available spectra is still lacking.

Sorting databases for efficient search

Calculating similarity scores between a query spectrum and a database containing hundreds of thousands of spectra can be computationally intensive. Classifying database compounds using molecular characteristics such as complexity [152, 158],

common substructures [17, 159] have proved useful in efficient compound identification [17, 18] and mining of chemical databases [160]. Applying similar strategies to spectral databases presents promising possibilities.

Key Points

- Dereplication is the process of rapid identification of previously isolated natural products compounds, which increases the chance of isolating new compound, and accelerates drug discovery therefrom.
- Automating dereplication requires the utilization and integration of diverse computational resources.
- We review the currently available computational resources that are useful in dereplication by categorizing them into databases, methods and software.
- We conclude by discussing current computational challenges to automating dereplication with proposed solutions.

Funding

This work is partially supported by JSPS KAKENHI 24300054. Ahmed Mohamed would like to thank the Rotary Yoneyama Memorial Foundation, Inc. for the scholarship awarded to him.

References

1. Li JW, Vederas JC. Drug discovery and natural products: end of an era or an endless frontier? *Science* 2009;**325**:161–5.
2. Beutler JA. Natural products as a foundation for drug discovery. *Curr Protoc Pharmacol* 2009;**Chapter 9**:Unit 9.11.
3. Koehn FE, Carter GT. The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* 2005;**4**:206–20.
4. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 2012;**75**:311–35.
5. Berdy J. Bioactive microbial metabolites. *J Antibiot* 2005;**58**:1–26.
6. Clemons PA, Bodycombe NE, Carrinski HA, *et al*. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci USA* 2010;**107**:18787–92.
7. Over B, Wetzel S, Grutter C, *et al*. Natural-product-derived fragments for fragment-based ligand discovery. *Nat Chem* 2013;**5**:21–8.
8. Buckingham J. *Dictionary of Natural Products*. CRC Press, Boca Raton, FL, USA, 1993.
9. Blunt JW, Munro MHG. 22 is there an ideal database for natural products research? In: Natural Products: Discourse, Diversity, and Design. John Wiley & Sons, Inc., Hoboken, NJ 2014:413.
10. Wolfender J-L, Marti G, Ferreira Queiroz E. Advances in techniques for profiling crude extracts and for the rapid identificationof natural products: Dereplication, quality control and metabolomics. *Curr Org Chem* 2010;**14**:1808–32.
11. Lang G, Mayhudin NA, Mitova MI, *et al*. Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. *J Nat Prod* 2008;**71**:1595–9.
12. Gerwick WH, Moore BS. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* 2012;**19**:85–98.
13. Rosenblum ML, Gerosa MA, Wilson CB, *et al*. Stem cell studies of human malignant brain tumors. Part 1: development of the stem cell assay and its potential. *J Neurosurg* 1983;**58**:170–6.
14. Molinski TF. Microscale methodology for structure elucidation of natural products. *Curr Opin Biotechnol* 2010;**21**:819–26.
15. Halabalaki M, Vougogiannopoulou K, Mikros E, *et al*. Recent advances and new strategies in the NMR-based identification of natural products. *Curr Opin Biotechnol* 2014;**25**:1–7.
16. Liu Y, Green MD, Marques R, *et al*. Using pure shift HSQC to characterize microgram samples of drug metabolites. *Tetrahedron Lett* 2014;**55**:5450–3.
17. Watrous J, Roach P, Alexandrov T, *et al*. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA* 2012;**109**:E1743–52.
18. Yang JY, Sanchez LM, Rath CM, *et al*. Molecular networking as a dereplication strategy. *J Nat Prod* 2013;**76**:1686–99.
19. Elyashberg ME. Identification and structure elucidation by NMR spectroscopy. *TrAC Trends Anal Chem* 2015;**69**:88–97.
20. Robinette SL, Bruüschweiler R, Schroeder FC, *et al*. NMR in metabolomics and natural products research: two sides of the same coin. *Acc Chem Res* 2011;**45**:288–97.
21. Wishart DS, Jewison T, Guo AC, *et al*. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res* 2013;**41**:D801–7.
22. Blinov KA, Carlson D, Elyashberg ME, *et al*. Computer-assisted structure elucidation of natural products with limited 2D NMR data: application of the StrucEluc system. *Magn Reson Chem* 2003;**41**:359–72.
23. Breton RC, Reynolds WF. Using NMR to identify and characterize natural products. *Nat Prod Rep* 2013;**30**:501–24.
24. Smolinska A, Blanchet L, Buydens LM, *et al*. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal Chim Acta* 2012;**750**:82–97.
25. Ji HF, Li XJ, Zhang HY. Natural products and drug discovery. *EMBO Rep* 2009;**10**:194–200.
26. Dandapani S, Marcaurelle LA. Grand challenge commentary: accessing new chemical space for'undruggable'targets. *Nat Chem Biol* 2010;**6**:861–3.
27. Wang B, Fang A, Heim J, *et al*. DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Anal Chem* 2010;**82**:5069–81.
28. Wishart DS. Quantitative metabolomics using NMR. *TrAC Trends Anal Chem* 2008;**27**:228–37.
29. Beckonert O, Keun HC, Ebbels TMD, *et al*. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protocols* 2007;**2**:2692–703.
30. Füllbeck M, Michalsky E, Dunkel M, *et al*. Natural products: sources and databases. *Nat Prod Rep* 2006;**23**:347–56.
31. Blunt J, Munro M, Upjohn M. The role of databases in marine natural products research. In: *Handbook of Marine Natural Products*. Springer, 2012, 389–421.
32. Lagunin AA, Goel RK, Gawande DY, *et al*. Chemo-and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Nat Prod Rep* 2014;**31**:1585–611.

33. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.
34. Hastings J, de Matos P, Dekker A, *et al*. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 2013;**41**: D456–63.
35. Seiler KP, George GA, Happ MP, *et al*. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 2008;**36**:D351–9.
36. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
37. ChemSpider. http://www.chemspider.com/(1/6/2014 2014, date last accessed).
38. Kalchhauser H, Robien W. CSEARCH: a computer program for identification of organic compounds and fully automated assignment of carbon-13 nuclear magnetic resonance spectra. *J Chem Inf Comput Sci* 1985;**25**:103–8.
39. Steinbeck C, Kuhn S. NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* 2004;**65**: 2711–7.
40. Li Q, Cheng T, Wang Y, *et al*. PubChem as a public resource for drug discovery. *Drug Discov Today* 2010;**15**:1052–7.
41. Goodman J. Computer software review: reaxys. *J Chem Inf Model* 2009;**49**:2897–98.
42. Barth A. SpecInfo: an integrated spectroscopic information system. *J Chem Inf Comput Sci* 1993;**33**:52–8.
43. Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;**45**:177–82.
44. Laatsch H. AntiBase, a Database for rapid dereplication and structure determination of microbial natural products. In: *Book AntiBase, A Database for Rapid Dereplication and Structure Determination of Microbial Natural Products*. John Wiley & Sons, Inc., Hoboken, NJ, 2010.
45. Hammami R, Zouhir A, Le Lay C, *et al*. BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol* 2010;**10**:22.
46. Ntie-Kang F, Mbah JA, Mbaze LM, *et al*. CamMedNP: building the Cameroonian 3D structural natural products database for virtual screening. *BMC Complement Alternat Med* 2013;**13**: 88.
47. Ntie-Kang F, Onguéné PA, Scharfe M, *et al*. ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Adv* 2014;**4**:409–19.
48. Lopez-Perez JL, Theron R, del Olmo E, *et al*. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics* 2007;**23**: 3256–7.
49. Mangal M, Sagar P, Singh H, *et al*. NPACT: naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Res* 2013;**41**:D1124–9.
50. Valli M, dos Santos RN, Figueira LD, *et al*. Development of a natural products database from the biodiversity of Brazil. *J Nat Prod* 2013;**76**:439–44.
51. Hammami R, Ben Hamida J, Vergoten G, *et al*. PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res* 2009;**37**:D963–8.
52. Dunkel M, Fullbeck M, Neumann S, *et al*. SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res* 2006;**34**:D678–83.
53. Banerjee P, Erehman J, Gohlke B-O, *et al*. Super Natural II—a database of natural products. *Nucleic Acids Res* 2014;**34**: D678–83.
54. Chen CY. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 2011;**6**:e15939.
55. Gu J, Gui Y, Chen L, *et al*. Use of natural products as chemical library for drug discovery and network pharmacology. *PloS One* 2013;**8**:e62839.
56. Roldán C, de la Torre A, Mota S, *et al*. Identification of active compounds in vegetal extracts based on correlation between activity and HPLC–MS data. *Food Chem* 2013;**136**: 392–9.
57. Vu TN, Laukens K. Getting your peaks in line: a review of alignment methods for NMR spectral data. *Metabolites* 2013; **3**:259–76.
58. Wang T, Shao K, Chu Q, *et al*. Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* 2009;**10**:83.
59. Hao J, Astle W, De Iorio M, *et al*. BATMAN–an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 2012;**28**:2088–90.
60. Hanson BA. ChemoSpec: An R Package for Chemometric Analysis of Spectroscopic Data and Chromatograms (Package Version 1.61-3) 2013.
61. Kim S, Fang A, Wang B, *et al*. An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure. *Bioinformatics* 2011;**27**:1660–6.
62. Worley B, Powers R. MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chem Biol* 2014;**9**:1138–44.
63. Wist J, Patiny L. Structural analysis from classroom to laboratory. *J Chem Educ* 2012;**89**:1083–83.
64. Helmus JJ, Jaroniec CP. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *J Biomol NMR* 2013;**55**:355–67.
65. Delaglio F, Grzesiek S, Vuister GW, *et al*. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 1995;**6**:277–93.
66. Izquierdo JL. NMRS: NMR Spectra Preprocessing (Package Version 1.0) 2009.
67. Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem* 2009;**47** (Suppl 1):S123–6.
68. Vu TN, Valkenborg D, Smets K, *et al*. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* 2011; **12**:405.
69. Olboyle NM, Banck M, James CA, *et al*. Open Babel: An open chemical toolbox. *J Cheminf* 2011;**3**:33.
70. Cao Y, Charisi A, Cheng LC, *et al*. ChemmineR: a compound mining framework for R. *Bioinformatics* 2008;**24**: 1733–4.
71. Guha R. Chemical informatics functionality in R. *J Stat Softw* 2007;**18**:1–16.
72. Cao D-S, Xiao N, Xu Q-S, *et al*. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions. *Bioinformatics* 2014:**31**:279–81.
73. Lancashire RJ. The JSpecView Project: an Open Source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chem Cent J* 2007;**1**:31.

74. Bienfait B, Ertl P. JSME: a free molecule editor in JavaScript. *J Cheminform* 2013;**5**:24.
75. Csizmadia F. JChem: Java applets and modules supporting chemical database handling from web browsers. *J Chem Inf Comput Sci* 2000;**40**:323–4.
76. Davies AN, Lampen P. Jcamp-Dx for NMR. *Appl Spectrosc* 1993;**47**:1093–9.
77. Goddard TD, Kneller DG. *Sparky—NMR Assignment and Integration Software*. University of California, San Francisco, 2006.
78. Zhang F, Brüschweiler R. Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angew Chem Int Ed* 2007;**46**:2639–42.
79. Robinette SL, Zhang F, Bruschweiler-Li L, *et al*. Web server based complex mixture analysis by NMR. *Anal Chem* 2008;**80**: 3606–11.
80. Rubtsov DV, Jenkins H, Ludwig C, *et al*. Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics* 2007;**3**:223–9.
81. Downing J, Murray-Rust P, Tonge AP, *et al*. SPECTRa: the deposition and validation of primary chemistry research data in digital repositories. *J Chem Inf Model* 2008;**48**: 1571–81.
82. Vranken WF, Boucher W, Stevens TJ, *et al*. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* 2005;**59**:687–96.
83. Chignola F, Mari S, Stevens TJ, *et al*. The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR. *Bioinformatics* 2011;**27**:885–6.
84. Hall SR. The STAR file: A new format for electronic data transfer and archiving. *J Chem Inf Comput Sci* 1991;**31**: 326–33.
85. Hall SR, Spadaccini N. The STAR file: Detailed specifications. *J Chem Inf Comput Sci* 1994;**34**:505–08.
86. Spadaccini N, Hall SR. Extensions to the STAR File syntax. *J Chem Inf Model* 2012;**52**:1901–6.
87. Dietrich W, Rüdel CH, Neumann M. Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra. *J Magn Reson* 1991;**91**:1–11.
88. Cobas JC, Bernstein MA, Martin-Pastor M, *et al*. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *J Magn Reson* 2006;**183**:145–51.
89. Bao Q, Feng J, Chen F, *et al*. A new automatic baseline correction method based on iterative method. *J Magn Reson* 2012; **218**:35–43.
90. Shao X, Ma C. A general approach to derivative calculation using wavelet transform. *Chemometr Intell Lab Syst* 2003;**69**: 157–65.
91. Shao X, Cai W, Pan Z. Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis. *Chemometr Intell Lab Syst* 1999;**45**:249–56.
92. Brown DE. Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials. *J Magn Reson Ser A* 1995;**114**:268–70.
93. Gan F, Ruan G, Mo J. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometr Intell Lab Syst* 2006;**82**:59–65.
94. Xi Y, Rocke DM. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC bioinformatics* 2008;**9**:324.
95. Boelens HFM, Dijkstra RJ, Eilers PHC, *et al*. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *J Chromatogr A* 2004;**1057**: 21–30.
96. Ruckstuhl AF, Jacobson MP, Field RW, *et al*. Baseline subtraction using robust local regression estimation. *J Quant Spectrosc Radiat Transf* 2001;**68**:179–93.
97. Komsta Ł. Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia* 2011;**73**:721–31.
98. Liu X, Zhang Z, Sousa PFM, *et al*. Selective iteratively reweighted quantile regression for baseline correction. *Anal Bioanal Chem* 2014;**406**:1985–98.
99. Zhang Z-M, Chen S, Liang Y-Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* 2010;**135**:1138–46.
100. Tawfike AF, Viegelmann C, Edrada-Ebel R. Metabolomics and dereplication strategies in natural products. In: *Metabolomics Tools for Natural Product Discovery*. Humana Press, Totowa, NJ, 2013, 227–44.
101. Koradi R, Billeter M, Engeli M, *et al*. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* 1998;**135**:288–97.
102. Brodsky L, Moussaieff A, Shahaf N, *et al*. Evaluation of peak picking quality in LC-MS metabolomics data. *Anal Chem* 2010;**82**:9177–87.
103. Yang C, He Z, Yu W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics* 2009;**10**:4.
104. Davis RA, Charlton AJ, Godward J, *et al*. Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometr Intell Lab Syst* 2007;**85**:144–54.
105. De Meyer T, Sinnaeve D, Van Gasse B, *et al*. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal Chem* 2008; **80**:3783–90.
106. Anderson PE, Mahle DA, Doom TE, *et al*. Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics* 2011;**7**:179–90.
107. Hinneburg A, Porzel A, Wolfram K. *An Evaluation of Text Retrieval Methods for Similarity Search of Multi-dimensional NMR-spectra*. In: *Bioinformatics Research and Development*. Springer, Berlin, Heidelberg, 2007, pp. 424–438.
108. Luts J, Poullet JB, Garcia–Gomez JM, *et al*. Effect of feature extraction for brain tumor classification based on short echo time 1H MR spectra. *Magn Reson Med* 2008;**60**: 288–98.
109. Castillo AM, Uribe L, Patiny L, *et al*. Fast and shift-insensitive similarity comparisons of NMR using a tree-representation of spectra. *Chemometr Intell Lab Syst* 2013;**127**:1–6.
110. Castillo AM, Bernal A, Patiny L, *et al*. A new method for the comparison of 1H NMR predictors based on tree-similarity of spectra. *J Cheminform* 2014;**6**:1–6.
111. Singh AP, Halloran J, Bilmes JA, *et al*. Spectrum identification using a dynamic Bayesian network model of tandem mass spectra. *arXiv* 2012;**1210**.4904.
112. Jeong J, Shi X, Zhang X, *et al*. Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry. *BMC Bioinformatics* 2012;**13**:27.
113. Green DE. Quantitation of cannabinoids in biological specimens using probability based matching GC/MS. *NIDA Res Monogr* 1976:70–87.

114. McLafferty FW, Hertel RH, Villwock RD. Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures. *Organic Mass Spectrometry* 1974;**9**: 690–702.
115. Koichi S, Arisaka M, Koshino H, *et al*. Chemical structure elucidation from 13C NMR chemical shifts: Efficient data processing using bipartite matching and maximal clique algorithms. *J Chem Inf Model* 2014;**54**:1027–35.
116. Levandowsky M, Winter D. Distance between sets. *Nature* 1971;**234**:34–5.
117. Egert B, Neumann S, Hinneburg A. Fast approximate duplicate detection for 2D-NMR spectra. In: *Data Integration in the Life Sciences*. Springer, Science & Business Media, Berlin, 2007, 139–55.
118. Hinneburg A, Egert B, Porzel A. Duplicate detection of 2d-nmr spectra. *J Integr Bioinform* 2007;**4**:53.
119. Beer I, Barnea E, Ziv T, *et al*. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004;**4**:950–60.
120. Atwater BL, Stauffer DB, McLafferty FW, *et al*. Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Anal Chem* 1985;**57**:899–903.
121. Tabb DL, MacCoss MJ, Wu CC, *et al*. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 2003;**75**:2470–7.
122. Li J, Hibbert DB, Fuller S, *et al*. Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Anal Chem* 2005;**77**:639–44.
123. Linusson A, Wold S, Nordén B. Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry, *Chemometr Intell Lab Syst* 1998;**44**:213–27.
124. Julian RK, Higgs RE, Gygi JD, *et al*. A method for quantitatively differentiating crude natural extracts using high-performance liquid chromatography-electrospray mass spectrometry. *Anal Chem* 1998;**70**:3249–54.
125. Tsipouras A, Ondeyka J, Dufresne C, *et al*. Using similarity searches over databases of estimated 13C NMR spectra for structure identification of natural product compounds. *Anal Chim Acta* 1995;**316**:161–71.
126. Rasmussen GT, Isenhour TL. The evaluation of mass spectral search algorithms. *J Chem Inf Comput Sci* 1979;**19**: 179–86.
127. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom* 1994;**5**:859–66.
128. Kim S, Koo I, Jeong J, *et al*. Compound identification using partial and semipartial correlations for gas chromatography–mass spectrometry data. *Anal Chem* 2012;**84**: 6477–87.
129. Koo I, Zhang X, Kim S. Wavelet-and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Anal Chem* 2011; **83**:5631–8.
130. Horai H, Arita M, Nishioka T. Comparison of ESI-MS spectra in MassBank database. *2008 International Conference on BioMedical Engineering and Informatics*, 2008, 853–7.
131. Koo I, Kim S, Zhang X. Comparative analysis of mass spectral matching-based compound identification in gas chromatography–mass spectrometry. *J Chromatogr A* 2013;**1298**: 132–8.
132. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 2004;**1**: 195–202.
133. Penchev PN, Schulz K-P, Munk ME. INFERCNMR: A 13C NMR Interpretive Library Search System. *J Chem Inf Model* 2012;**52**: 1513–28.
134. Katritzky AR, Kuanar M, Slavov S, *et al*. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem Rev* 2010;**110**: 5714–89.
135. Blinov KA, Smurnyy YD, Churanova TS, *et al*. Development of a fast and accurate method of 13C NMR chemical shift prediction. *Chemometr Intell Lab Syst* 2009;**97**:91–7.
136. Binev Y, Aires-de-Sousa J. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *J Chem Inf Comput Sci* 2004;**44**:940–5.
137. Aires-de-Sousa J, Hemmer MC, Gasteiger J. Prediction of 1H NMR chemical shifts using neural networks. *Anal Chem* 2002;**74**:80–90.
138. Binev Y, Corvo M, Aires-de-Sousa J. The impact of available experimental data on the prediction of 1H NMR chemical shifts by neural networks. *J Chem Inf Comput Sci* 2004;**44**: 946–49.
139. Heinonen M, Rantanen A, Mielikäinen T, *et al*. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom* 2008;**22**:3043–52.
140. Wolf S, Schmidt S, Müller-Hannemann M, *et al*. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 2010;**11**:148.
141. Allen F, Pon A, Wilson M, *et al*. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* 2014:**42**: W94–9.
142. Fitch WL, McGregor M, Katritzky AR, *et al*. Prediction of ultraviolet spectral absorbance using quantitative structure-property relationships. *J Chem Inf Comput Sci* 2002;**42**: 830–40.
143. Peng CT. Prediction of retention indices: V. Influence of electronic effects and column polarity on retention index. *J Chromatogr A* 2000;**903**:117–43.
144. Liu SS, Liu Y, Yin DQ, *et al*. Prediction of chromatographic relative retention time of polychlorinated biphenyls from the molecular electronegativity distance vector. *J Separ Sci* 2006;**29**:296–301.
145. Liao L, Mei H, Li J, *et al*. Estimation and prediction on retention times of components from essential oil of *Paulownia tomentosa* flowers by molecular electronegativity-distance vector (MEDV). *J Mol Struct THEOCHEM* 2008;**850**:1–8.
146. Lodewyk MW, Siebert MR, Tantillo DJ. Computational prediction of 1H and 13C chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem Rev* 2011;**112**:1839–62.
147. Kuhn S, Egert B, Neumann S, *et al*. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics* 2008;**9**: 400.
148. Elyashberg M, Blinov K, Smurnyy Y, *et al*. Empirical and DFT GIAO quantum-mechanical methods of 13C chemical shifts prediction: competitors or collaborators? *Magn Reson Chem* 2010;**48**:219–29.

149. Tharatipyakul A, Numnark S, Wichadakul D, *et al*. ChemEx: information extraction system for chemical data curation. *BMC Bioinformatics* 2012;**13** (Suppl 17):S9.
150. Vazquez M, Krallinger M, Leitner F, *et al*. Text mining for drugs and chemical compounds: methods, tools and applications. *Mol Inform* 2011;**30**:506–19.
151. Bertz SH. On the complexity of graphs and molecules. *Bull Math Biol* 1983;**45**:849–55.
152. Nikolic S, Trinajstic N, Tolic IM. Complexity of molecules. *J Chem Inf Comput Sci* 2000;**40**:920–6.
153. El-Elimat T, Figueroa M, Ehrmann BM, *et al*. High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. *J Nat Prod* 2013;**76**:1709–16.
154. Nielsen KF, Mansson M, Rank C, *et al*. Dereplication of microbial natural products by LC-DAD-TOFMS. *J Nat Prod* 2011;**74**: 2338–48.
155. Staerk D, Kesting JR, Sairafianpour M, *et al*. Accelerated dereplication of crude extracts using HPLC-PDA-MS-SPE-NMR: quinolinone alkaloids of Haplophyllum acutifolium. *Phytochemistry* 2009;**70**:1055–61.
156. Motti CA, Freckelton ML, Tapiolas DM, *et al*. FTICR-MS and LC-UV/MS-SPE-NMR applications for the rapid dereplication of a crude extract from the sponge Ianthella flabelliformis. *J Nat Prod* 2009;**72**:290–4.
157. Menikarachchi LC, Cawley S, Hill DW, *et al*. MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. *Anal Chem* 2012;**84**:9388–94.
158. Bywater RP. Membrane-spanning peptides and the origin of life. *J Theor Biol* 2009;**261**:407–13.
159. Koch MA, Schuffenhauer A, Scheck M, *et al*. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 2005;**102**: 17272–7.
160. Batista J, Bajorath J. Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J Chem Inf Model* 2007;**47**:59–68.