doi:10.2533/chimia.2007.339 CHIMIA **2007**, *61*, No. 6

Chimia 61 (2007) 339–345

ISSN 0009–4293

Burning the Hay to Find the Needle – Data Mining Strategies in Natural Product Dereplication

D. Wolf* and K. Siems

Abstract: The acquisition and use of data from the LC/MS-ELSD analysis of extracts is described. The methodology requires MS spectra to be recorded in the positive/negative ESI mode, as well as the determination of retention time and peak area from ELSD. Subsequent calculation of molecular weight, referenced retention time, and normalized peak area, results in the creation of a peak library, which can be used for different data mining strategies: i) the dereplication of previously isolated natural products; ii) clustering/ranking of extracts for the creation of highly diverse natural product libraries; iii) a selection tool for the focused isolation of bioactive natural products and iv) to search for alternative sources of a target natural product.

Keywords: Chemodiversity profiling · Dereplication · Extract · Mass spectrometry · Natural product

1. Introduction

Natural products are still an important source for the discovery of new lead compounds in the pharmaceutical and agrochemical industry. Furthermore, from the field of metabolomics there is a growing interest in identifying metabolites from samples to obtain a deeper insight into the metabolism of different organisms.[1] Currently, leading groups in the metabolomics field are working on the META-PHOR project (financed by the European Union EU)[2] on the implementation of standard methods for data acquisition, data storage, and data mining.

Both topics, the discovery of lead compounds and metabolomics, have a strong need for the identification of compounds in complex extracts.

1.1. *Bioassay-Guided Fractionation and Pure Compound Screening*

There are two main approaches to the discovery of lead compounds from extracts: bioassay-guided fractionation and pure compound screening.[3] The first approach, bioassay-guided fractionation, typically starts with the screening of extracts, followed by the repeated fractionation of biologically active extracts and fractions until the successful isolation of a natural product which is then passed on to structure elucidation. Extracts can be excluded from further work, if they contain compounds or compound classes already known for their activity in the specific assay. The process to identify already known purified natural products and the avoidance of repetitive work is called dereplication. The second approach, pure compound screening, starts with an automated process of isolation and structure elucidation of the majority of the secondary metabolites contained in the crude extract, followed by screening of the purified and structurally elucidated natural products. In this approach, it is necessary to select extracts containing compounds that are not already present in the library of pure compounds. Both approaches must identify as many natural products (known or unknown biological activity) in extracts or fractions as possible at the earliest stage to avoid redundant work.

Beside the identification of already isolated compounds in extracts, it is advantageous to also allocate unknown compounds in different extracts. This information could be used for

i) extract ranking in bioassay-guided fractionation,
ii) identification of the active principle by comparing active and inactive extracts and
iii) pure compound screening as discussed later.

1.2. *Methods for Dereplication of Natural Products*

The most common method used for the dereplication of natural products consists of the separation of single metabolites and their identification. The separation step is dominated by chromatography; the identification step is dominated by spectroscopy. Mainly of historical importance is the use of staining reagents in DC chromatography for the identification of specific natural product classes or functional groups.

Today, methods of choice are GC/MS for volatile compounds and HPLC/MS, DAD and/or NMR for all other compounds. In LC/MS electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI) is used.[4] When using high-resolution Q-TOF-MS/MS,[5] FT-ICR-MS[6] instruments or MS/MS and MS/MS/MS techniques,[7] it is possible to determine the molecular mass with exceptional accuracy resulting in a list of molecular formula(s) per peak, which could be very helpful for dereplication. Although some natural products only show poor fragmentation in ESI MS, the fragmentation pattern can be very helpful in dereplication, mainly when dealing with glycosylated compounds. MS/MS

*Correspondence: Dr. D. Wolf
AnalytiCon Discovery GmbH
Hermannswerder Haus 17
14473 Potsdam, Germany
Tel.: +49 331 2300 309
Fax: +49 331 2300 333
E-Mail: d.wolf@ac-discovery.com

experiments are helpful to determine precursor ions and to exclude misinterpretation in the case of co-elution of two or more compounds, but it requires careful selection of MS/MS parameters during the analysis of extracts. Using information-dependent acquisition (IDA[8]), the molecular weight for each peak can be detected on-line and the parameters for the MS/MS experiment (daughter scan) are adjusted automatically for the next scan. Thus, it is no longer necessary to run a second LC/MS to record the daughter spectra, which was time consuming, required manual interpretation of MS, and clearly did not fulfill the requirements of a high throughput data acquisition and interpretation process.

Even when using LC/MS/MS there is still the risk of co-elution of isomers. Theoretically, coupling LC/MS with NMR would allow unambiguous identification of compounds, but lack of sensitivity is often a problem and automation of interpretation is difficult. The necessary sensitivity for on-line NMR (>>1 µg per compound of interest) can be achieved by HPLC with deuterated solvents or by on-line SPE-coupling (if necessary trapping the same peak from more than one HPLC run before elution of the pure compound with deuterated solvents into the NMR spectrometer).[9,10] Another problem is solvent suppression when using H_2O in the HPLC gradient. Technically it is possible to suppress the proton signal of water very efficiently, but especially the anomeric signals of sugars, which are quite important in interpretation of glycosylated compounds (*e.g.* saponins or antibiotics), typically with resonances between 4.5 and 5.5 ppm, are sometimes also suppressed or show at least reduced intensity. Nevertheless, fully integrated dereplication systems, including necessary hardware and suitable software packages (*e.g.* the Metabolic Profiler from Bruker), are commercially available. An overview of the different hyphenated methods is given in Table 1.

Knowing the taxonomy could be a useful tool for efficient dereplication. For plants, taxonomy based on morphological criteria is usually sufficient and for microorganisms, methods based on molecular biology (*e.g.* ribotyping[26]) or investigation of peptide fingerprints by MALDI-TOF[27] have been successfully used.

1.3. *Quantitation of Natural Products in Extracts*

In addition to identification, the quantitation of secondary metabolites in extracts is also important. In metabolomics, data from relative quantitation of small metabolites (*e.g.* amino and other organic acids, sugars, secondary metabolites such as alkaloids, flavonoids or terpenoids) are the basis for comparative analysis of different genotypes. In lead compound discovery, an absolute quantitation is needed to get an impression of how much biomaterial (*e.g.* fermentation volume or dried plant material) has to be produced and extracted to obtain sufficient amounts of the target compound for structure elucidation and screening. Due to the lack of standards for most of the natural products, quantitation is a challenging task, especially for unknown natural products. The optimal quantitation method should be independent of the chemical structure. In lead compound discovery, HPLC/UV-detection recorded at a non-specific wavelength (*e.g.* 210 nm), charged aerosol detector (CAD), evaporative light scattering detection (ELSD) or NMR are used, but the moderate limit of detection (LOD) for NMR and ELSD leads to a restrictive use in quantitation. Sensitivity enhancement is possible by use of solid-phase extraction (SPE) in the offline or highly automated online[28] mode to enrich compounds of interest. Depending on the requirements, the enhancement of sensitivity can also be achieved by preparative HPLC fractionation and analysis of the concentrated fractions.

Information about the concentration of individual known and unknown natural products may be useful for bioassay-guided fractionation; however, the abundant data acquired from the extracts are worth more than the sum of the parts.

Table 1. Hyphenated techniques used in dereplication of natural products

Separation	Detection	Comment	References
TLC	Staining reagents	Chemical screening	Grabley[11]
GC	MS	Essential oils	Roman[12]
		Tropane alkaloids	Witte[13]
		Metabolomics for small polar compounds	Kopka[14]
		Dereplication database of mycotoxins	Nielsen[15]
		Derivatized polar compounds from bacteria	Böröczky[16]
HPLC	DAD	Physical chemical screening	Fiedler[17]
		X-hitting algorithm	Larsen[18]
HPLC	DAD/MS/NMR	Plant extracts (saponins, iridoids, xanthones)	Hostettmann[19]
HPLC	ELSD/MS	Natural product libraries	Zeng[20]
HPLC	ELSD/ESI-MS	Chemodiversity profiling of extract libaries	Jakupovic[21]
HPLC	NMR/MS^n	Coupled with MSPD, NMR quantitation	Preiss[22]
HILIC	ESI-MS	Extremely polar natural products	Strege[23]
SFC	APCI-MS	Dereplication of artemisinin	Dost[24]
CE	CE/MS	Indole alkaloids in cell cultures	Stöckigt[25]

2. LC/MS-ELSD-Based Identification and Quantitation of Natural Products

Our group is using several parameters for identification and quantitation of natural products in extracts derived from the LC/MS-ELSD data per peak:

i) referenced retention time from ELSD;
ii) molecular weight calculated from (±)-ESI-MS;
iii) other information like UV spectra and taxonomic information and
iv) concentration by referenced ELSD peak area.

2.1. *Highly Reproducible Retention Times in HPLC*

The accuracy of retention times in HPLC and GC is important for the unambiguous identification of natural products in mixtures. Slight changes in the chromatographic conditions, *e.g.* pH of the solvent or RP material, might change the retention time dramatically. In GC, Kovats established a method to improve the accuracy of retention times more than 40 years ago.[29] Frisvad and Thrane have introduced a similar system in the taxonomical identification of fungi by characterization of their mycotoxin patterns by HPLC/DAD chromatography[30] using bracketed alkylphenone retention indices.[31] Our group uses an external standard mixture of twelve natural products from different polarities and compound classes (*e.g.* alkaloids, phenols, and saponins). It is injected every 10th LC/MS run and relative retention times are calculated by interpolation between standard compounds for each peak in the extracts.[21] Using this method, the variance of retention times of the same compound originating from different extracts (different concentrations in different matrices) recorded on different instruments (*e.g.* using high pressure and low pressure gradient mixing) and with different column batches could be reduced to less than 0.5% (about ±10s in a 30 min gradient).

It is also possible to calculate retention times for natural products known in the lit-

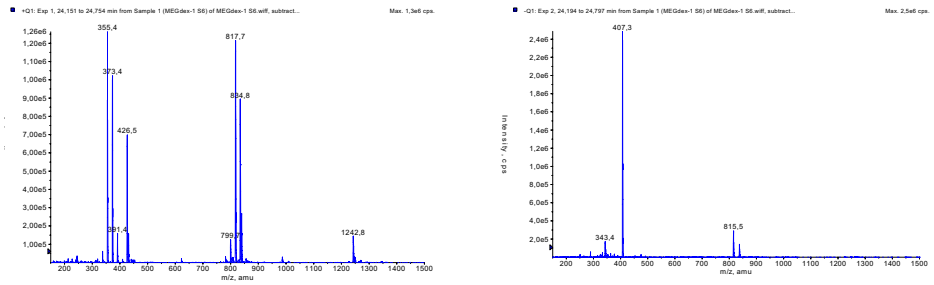
erature but unavailable by using referenced retention times and structures of analyzed compounds for self-learning programs (*e.g.* based on neural networks). Although this principle is well documented in the literature,[32] it has not yet been applied for prediction of HPLC retention times for dereplication purposes.

2.2. *Automated Calculation of Molecular Weight from (±)-ESI-MS*

Searching MS spectra in spectral libraries is well established for GC/MS spectra recorded under EI conditions, but in LC/MS with ESI or APCI ionization algorithms for searching EI spectra often fail, because the intensity of adducts of Na^+ and solvent additives used in HPLC (*e.g.* $[NH_4]^+$ or $[COOH]^-$) depend on specific instrument conditions and therefore they are essentially irreproducible. Sometimes different adducts even lead to different fragmentations.[19] Thus, in LC/MS it is not the comparison of experimental MS spectra, but rather the interpretation of spectra that is the recommended approach. The calculation of molecular weight by use of adducts and ions occurring under different ionization conditions ((+)- and (–)-ESI) has the highest importance. Our group uses the following algorithm:

i) Export the ten most intensive m/z values from (+)- and (–)-ESI MS for each peak in the chromatogram.
ii) Calculate for each m/z value all possible corresponding molecular weights. When using ammonium formiate buffer as mobile phase in (+)-ESI, a m/z value could be caused by $[M+H]^+$, $[M+NH_4]^+$, or $[M+Na]^+$, respectively. In (–)-ESI, a m/z value could be caused by $[M–H]^-$, or $[M+HCOO]^-$, respectively.
iii) Sum the relative intensities of the corresponding m/z values.
iv) Sort the molecular weights descending by the sum of the relative intensities.
v) Check each of the potential molecular weights, if it could be interpreted as a dimer of one of the other molecular weights. If so, add the sum of the relative intensities to the sum of the potential monomers.
vi) Check each of the potential molecular weights, whether it could be interpreted as a fragment of one of the other molecular weights. The interpretation of the fragmentation process depends on a list of fragments commonly observed in the analysis of natural products by ESI-MS, *e.g.* 18 (loss of water), 132 (loss of pentose), 146 (loss of deoxyhexose or cumaric acid), 162 (loss of hexose or caffeic acid).

An example for a compound with the molecular weight of 408 g/mol is given in Fig. 1.



Left: (+)-ESI-MS, Right: (–)-ESI-MS

i) Export the ten most intensive m/z values per ionisation method and

ii) Calculation of potential molecular weights by use of corrected m/z values

(+)-ESI					(–)-ESI			
m/z	rel. Intensity	M+1	M+18	M+23	M/z	rel. Intensity	M–1	M+45
817.7	21%	816	799	794	407.3	75%	408	362
355.4	21%	354	337	332	815.5	9%	816	770
373.4	17%	372	355	350	343.4	5%	344	298
834.8	15%	833	816	811	837.5	4%	838	792
426.5	12%	425	408	403	289.2	2%	290	244
839.7	5%	838	821	816	333.3	2%	334	288
391.4	3%	390	373	368	363.5	1%	364	318
431.5	3%	430	413	408	377.4	1%	378	332
1242.8	2%	1241	1224	1219	353.2	1%	354	308
799.7	2%	798	781	776	251.3	1%	252	206

iii) Calculation of the sum of the relative intensity (divided by two for each ionisation method), and

iv) Check for dimers and recalculation (not all data shown)

MW	Sum of rel. intensity	Comment	Final sum[a]
408	0.44		0.69
362	0.37		0.37
816	0.25	Dimer of 408	–
332	0.11		0.11
354	0.11		0.11

[a] The result is a value between 0 and 1 for each potential MW which could be interpreted as a probability value of the molecular weight.

The molecular weight is 408 g/mol.

Fig. 1. Example for MW calculation

2.3. *Identification of Natural Products in Crude Extracts*

Each peak in the chromatogram is characterized by its relative retention time, molecular weight and, if present, important fragments. In some cases, a certain peak exhibits more than one plausible molecular weight calculated from the MS data due to co-eluting compounds or non-implemented fragmentation patterns. The database query includes referenced retention time and molecular weight as the first filter. The resulting hit list of potential structures could be further reduced by using fragment information (*e.g.* the hit compound has an O-bonded glucose, but the ESI-MS of the query peak did not show the [M–162] fragment, which usually means that the compound could be excluded from the hit list) and UV data (*e.g.* the hit compound is a flavonoid, but the UV spectrum of the peak did not show the typical UV spectrum of flavonoids).

Finally, chemotaxonomic information, *e.g.* plant genus and species or type of organism (fungi or bacteria) might be additional filtering criteria by use of positive and negative hit lists. For example in the case of micro-organisms, flavonoids are quite unusual except when extracted from the cultivation media, *e.g.* genistein. In case of plants, there are many compound classes quite specific for certain families or *vice versa* never occurring in particular plant families, *e.g.* iridoids, which are common in several plant families (*e.g.* Apocynaceae, Rubiaceae, Scrophulariaceae, Lamiaceae), but have not yet been detected in the well-investigated Asteraceae family. The identification of the antifungal alkaloid balanol (CAS 63590-19-2) is given in Fig. 2 as an example.

2.4. *Quantitation via ELSD*

There are two main indications for quantitation *via* ELSD: i) purity determination of compound libraries[34] and ii) quantitation of secondary metabolites contained in a crude extract.[3] In general, the ELSD can be considered a quasi-universal and mass-dependent detector of the quantity of non-volatile secondary metabolites in crude extracts, although the response factors of the detector are influenced by the nature of the solvents and analytes.[35,36] After analysis of crude extracts by standardized LC/MS-ELSD, the ELSD data of each of the peaks detected in the medium polar range were used to predict the concentration in the crude extract by correction, considering the sigmoidal response curve with the concentration and the response variation with the mobile phase composition. We have analyzed 326 isolated, pure natural products in concentrations from 0.05 mg/ml to 2 mg/ml by LC/MS-ELSD and used the corresponding dependence for the general calibration of acquired ELSD peak areas. The application of the corrected ELSD data per peak allows the quantitation of known and unknown secondary metabolites regardless of their chemical structure, molecular weight, concentration, and polarity. Although many groups are using ELSD for the quantitation of special natural product classes by use of reference compounds, *e.g.* active ingredients in traditional Chinese medicine,[37,38] to our knowledge, there is no application in the literature using ELSD as a quantitation tool even for unknown secondary metabolites without using identical natural products as reference material.

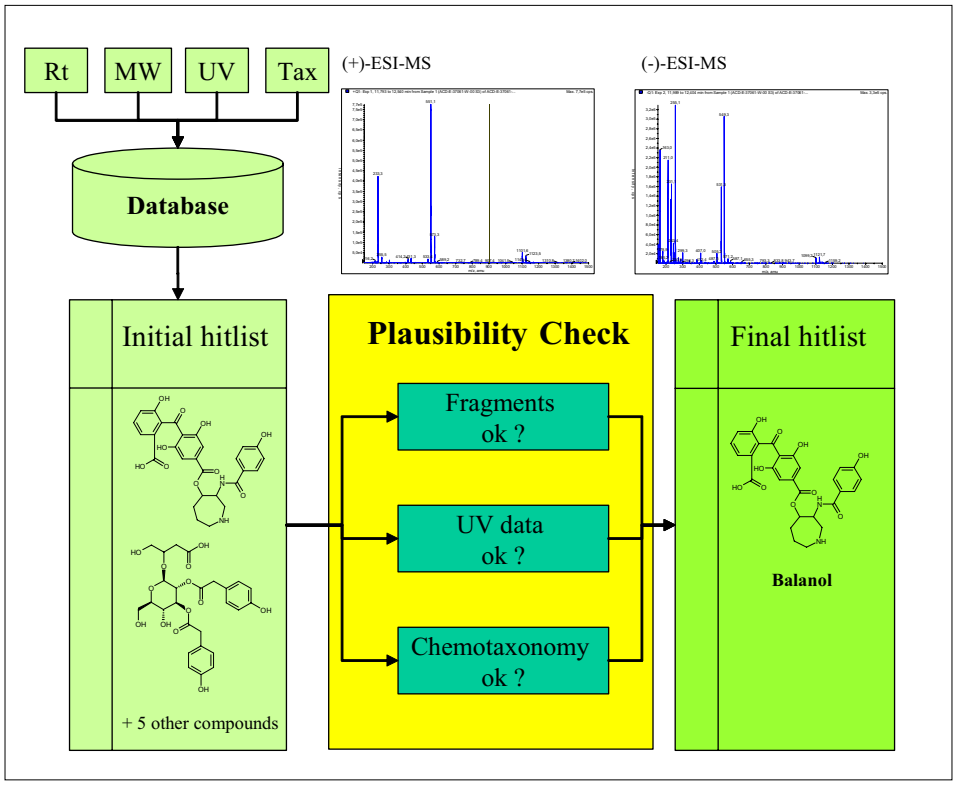


Fig. 2. Database search. Input data for the query: Megdex 1173, MW 550, fragment 233 (+)-ESI), UV active at 254 nm, taxonomy: produced by a fungi. The initial hitlist is the result from the query for MW and Megdex (± 30 units) only. The final hitlist is the result of a refined search using fragmentation, UV and taxonomy data.

3. Applications

The next step is the application of the acquired LC/MS-ELSD data from crude extracts to the dereplication of previously isolated and fully structurally elucidated natural products. We are using an internal database (calculated molecular weight, referenced retention time, UV data, source organism) of 20000 isolated and structurally elucidated pure natural products (10000 from microbial sources and 10000 from plants) with a novelty rate[39] of 60% and 40%, respectively. In the case of microbial extracts, additional data from the analysis of blind media extracts by LC/MS-ELSD can be easily incorporated in order to allocate unwanted compounds extracted from the fermentation medium, *e.g.* soy saponins or any other, even unknown peaks. Depending on the strategy used in natural product drug discovery, a threshold for concentration per detected peak might be used as a filter for data evaluation and to narrow down the number of natural products for isolation and identification.

3.1. *Burning the Hay to Find All Needles: Selection Tool for the Creation of Highly Diverse Pure Natural Product Libraries*

Due to the time-consuming bioassay-guided fractionation and disappointing experiences with natural product extracts in modern screening formats,[40] many pharmaceutical companies either terminated or reduced their activities in natural product drug discovery or switched to pure natural compound libraries in a ready-to-screen format. For the pure natural compound strategy, the creation of a maximal diverse natural product library using a broad variety of source organisms is essential. Small-scale extracts, either prepared from the fermentation of micro-organisms (*e.g.* 10 ml cultivation under different conditions) or from plant material, will be analyzed by LC/MS-ELSD and a database ('peak library') comprising the calculated data for all peaks from all extracts will be compiled. The peak library is the starting point for the exploration of the diversity ('ranking') of

Table 2. Possible criteria for compiling a peak sublibrary

Criteria	Data Source	Example Filter
Molecular Weight	(±)-ESI-MS	MW between 300–1000 g/mol
Polarity	Referenced Retention Time or ClogP of pure natural products as reference	Medium-polar natural products with ClogP between 1 and 3
Abundance	Frequency of the dataset MW and Rt in the complete peak library	Unique peaks (single abundance)
Concentration	Quantitation by ELSD	≥0.1 mg/l (strains) or ≥1 mg in 500 g (plants)
Novelty I	Dereplication of previously isolated natural products	New (unpublished) natural products
Novelty II	Dereplication of previously isolated natural product	Natural products previously not isolated by our group
Novelty III	Dereplication of previously isolated natural product	Commercially available natural products provided by our group

all extracts considering criteria for the creation of tailor-made pure compound libraries. Setting tailor-made filters for all peaks in the peak library, a second, reduced peak library ('peak sublibrary') is created.

Table 2 summarizes the possible filters that can be applied in any combination. For each extract, the level of overall uniqueness ('ranking points'), which is a function of the uniqueness of its peak sublibrary, is calculated. The result is a ranking list, which contains the greatest number of most unique peaks at the top position, or in other words, the extracts could be ordered by chemodiversity ('chemodiversity profiling').

Besides the quantitation of known and unknown natural products in complex extracts, the limit of detection of the ELSD might be crucial, since the target compounds and presumably biologically active natural products are present in less than 1% by weight of the crude extract. An initial search in the literature regarding the concentration of natural products from micro-organisms obtained by classical bioassay-guided fractionation reveals a wide range from 1–600 mg/l (isolated amount of biologically active compound to large-scale cultivation volume). Applying the generally accepted LOD of 50 ng of the ELSD[41] to the chemical profiling by LC/MS-ELSD (extract concentration: 10 mg/ml, injection of 50 µl), compounds can be detected down to 0.01% in the crude extract. For a typical extract with an average extract weight of 25 g, prepared from a 10 l large-scale fermentation of a micro-organism, all compounds in calculated amounts of 2.5 mg can be detected by ELSD.

Thus, a library of thousands of small-scale extracts can be classified and the organisms exhibiting the highest chemical diversity can be selected for large-scale extract preparation, isolation and structure confirmation and/or identification. Although other methods, such as HPLC-ES-MS,[42] HPLC-UV,[43] and direct-infusion ES-MS,[44] are known for the exploration of the diversity of extract libraries, the present data mining strategy represents the most universal and flexible method to qualify and quantify extracts.

3.2. *Burning the Hay to Find the Golden Needle: Selection Tool for the Focused Isolation of Potentially Bioactive Natural Products*

The second strategy is based on the combination of bioassay-guided fractionation and the dereplication technology described before. The main task of the classical bioassay-guided fractionation is the biological screening of extracts either from a randomized collection or selected by chemodiversity profiling. One question arises from the results of the biological screening of extract libraries: What is the biologically active principle in a specific extract? Besides answering the question by classical bioassay-guided fractionation, an alternative strategy might be to remove all natural products with predicted biological inactivity and to focus on a reduced number of isolations and unambiguous identifications of natural products with high probability to exhibit biological activity. This strategy, developed by our group, consists of chemical analysis, biological screening and data mining:

i) analysis of thousands of extracts by LC/MS-ELSD;
ii) automated calculation of molecular weights and referenced retention time (ELSD or TIC retention times);
iii) creation of a peak library consisting of millions of datasets without any restrictions, *e.g.* threshold in the ELSD chromatogram;
iv) biological screening of the complete extract library or parts thereof;
v) dividing the huge peak library into two sublibraries: sublibrary A contains all datasets from all biologically inactive extracts;
vi) sublibrary B contains all datasets from all biologically active extracts;
vii) matching sublibrary A with sublibrary B.

The remaining datasets from sublibrary B reduced by the datasets from sublibrary A are the potential, biologically active natural products, which need to be isolated and structurally identified (Fig. 3). Prior to the extraction of biomaterial followed by isolation and identification of natural products, further dereplication steps (known natural products biologically active in the assay of interest or frequent hitters) and/or clustering might be an option. The big advantages are obvious: once an extract library is analyzed by LC/MS-ELSD, the datasets can be used for each and every screening campaign.

3.3. *Finding the Golden Needle in Other Haystacks: Search for Alternative Sources of a Target Natural Product*

Another application strategy may be used at a later stage of the natural product drug discovery process after successfully identifying a biologically active natural product: the search for alternative sources for a target compound within a chemically profiled extract library. One way might be the identification using the dataset of the identified target compound, *e.g.* to find an organism or fermentation condition which provides the target compound in multiple amounts compared to the initially used organism. Another way is the identification using the dataset of derivative(s) of the identified target compound, *e.g.* the corresponding methyl ester in case of the acid for chemical or enzymatic transformation to the target compounds after identifying the most productive organism.

During a campaign for the creation of a diverse, pure natural product library (MEGx off-the-shelf) using the MEGAbolite® technology, the antifungal alkaloid balanol (CAS 63590-19-2) was isolated from *Fusarium sp*. Since balanol seems to be an interesting starting material for the creation of a semi-synthetic natural product library (NatDiverse™, Fig. 4), it was decided to produce balanol on a 50 g scale. The search for alternative sources in our extract library of more than 20000 extracts using the dataset (molecular weight, fragmentation pattern, referenced retention time, DAD data) showed that balanol was produced by more than ten different organisms, each of them showing a totally different compound pattern in ELSD chromatogram (Fig. 5). Consequently, it is not recommended to apply the overall chemical fingerprinting of a certain extract for the search for similar or alternative extracts, but to use the specific dataset of the target compound (retention time, MW and diagnostic fragments) for the search within the peak library created from all analyzed extracts.

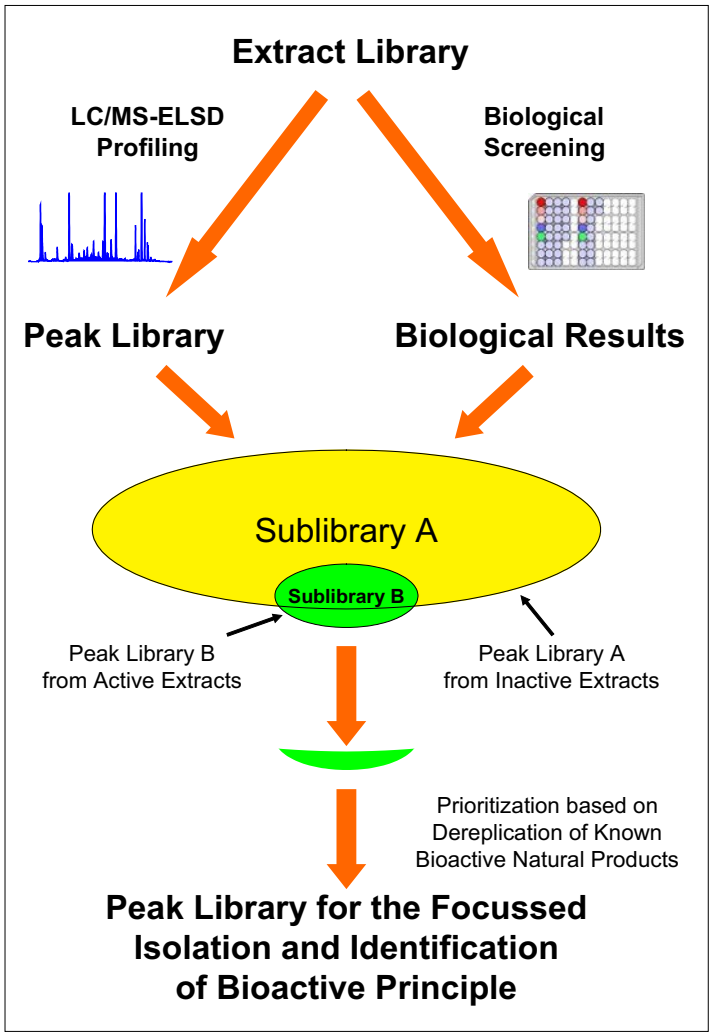


Fig. 3.

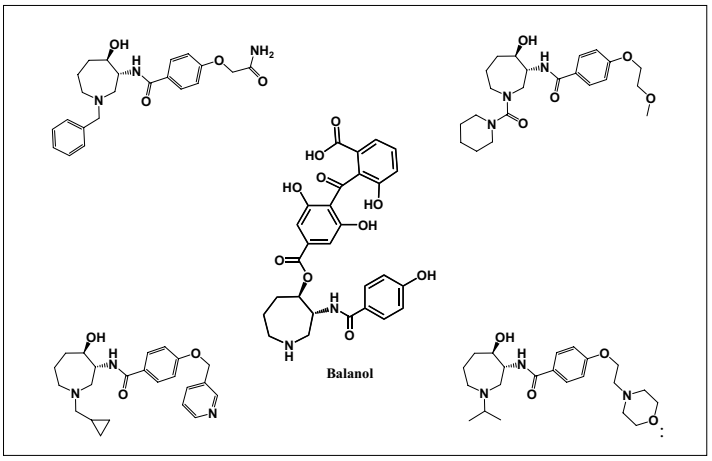


Fig. 4. Examples for Natdiverse™ compounds based on balanol. Starting with 50 g of the natural product balanol a library of 250 compounds (50–100 mg each) was synthesized.

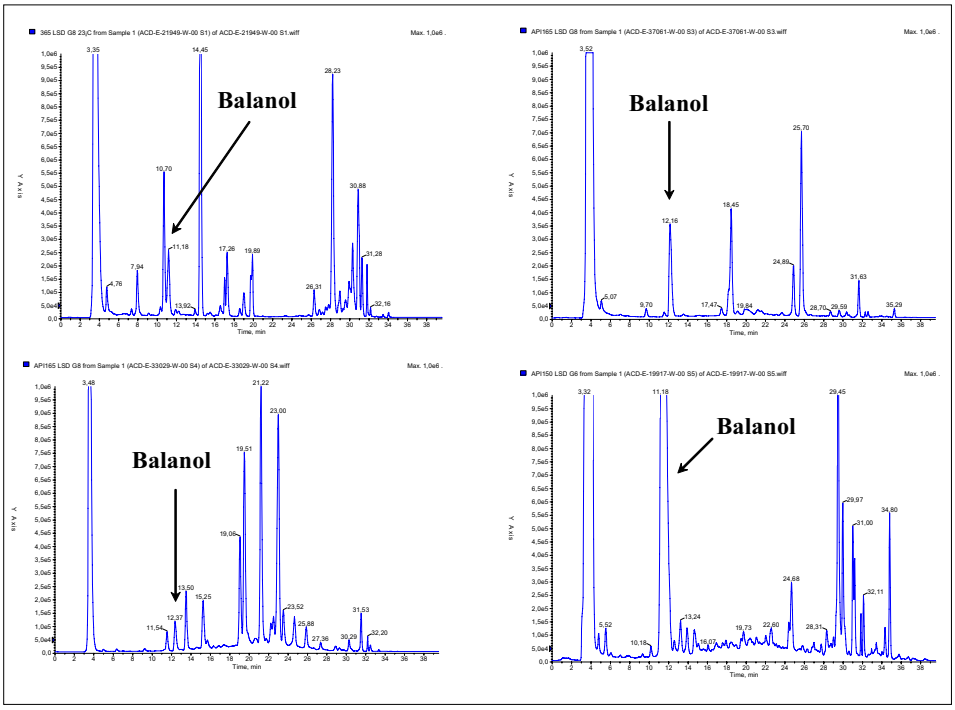


Fig. 5. ELSD Chromatograms of crude extracts containing balanol. HPLC conditions: RP-Select B 250×5, solvent A: ammonium formiate buffer, solvent B: methanol-acetonitril 1:1; gradient 15% B to 100% B in 30 min followed by 15 min 100% B, 1 ml/min flowrate, evaporative light scattering detection (ELSD); all samples have a concentration of 10 mg/ml raw extract in DMSO. Balanol is produced by different strains with completely different by-product patterns.

4. Conclusion

The described methodology can be used as a tool for identifying compounds, *e.g.* secondary metabolites in complex mixtures, for selecting the most productive organisms for further processing, and for speeding-up the lead compound identification in the early stages of the drug discovery process.

Acknowledgment

The authors thank Frank E. Koehn from Wyeth Research (Pearl River, USA) for valuable and helpful discussions. We also wish to thank our colleague Brantley Haigh for critically reviewing the manuscript.

Received: March 30, 2007

[1] R. D. Hall, *New Phytologist* **2006**, *169*, 453.

[2] META-PHOR, EC Project Code FOOD-CT-2006-036220. For an overview on the aims of the project see R. D. Hall, *Agro FOOD industry hi-tech* **2007**, *18*, 14 and *www.meta-phor.eu*.

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

[4] J.-L. Wolfender, S. Rodriguez, K. Hostettmann, *J. Mass. Spectrom. Rap. Comm. Mass. Spectr.* **1995**, 35.

[5] J.-L. Wolfender, P. Waridel, K. Ndjoko, K. R. Hobby, H. J. Major, K. Hostettmann, *Analusis* **2000**, *28*, 895.

[6] L. A. McDonald, L. R. Barbieri, G. T. Carter, G. Kruppa, X. Feng, J. A. Lotvin, M. M. Siegel, *Anal. Chem.* **2003**, *75*, 2730.

[7] Y. Konishi, T. Kiyota, C. Draghici, J.-M. Gao, F. Yeboah, S. Acoca, S. Jarussophon, E. Purisima, *Anal. Chem.* **2007**, *79*, 1187.

[8] Several technical notes about IDA are available at *www.appliedbiosystems.com* under the keyword ‘information depend acquisition’.

[9] C. Clarkson, D. Stark, S. H. Hanson, P. J. Smith, J. W. Jaroszewski, *J. Nat. Prod.* **2006**, *69*, 1280.
[10] M. Lambert, D. Stark, S. H. Hanson, M. Sairafianpour, J. W. Jaroszewski, *J. Nat. Prod*. **2005**, *68*, 1500.
[11] S. Grabley, R. Thiericke, A. Zeeck, in 'Drug Discovery from Nature', Eds. S. Grabley, A. Thiericke, Springer, **1999**.
[12] R. Oprean, M. Tamas, R. Sandulescu, L. Roman, *J. Pharm. Biomed. Analysis* **1998**, *18*, 651.
[13] K. Doerk-Schmitz, L. Witte, A. W. Alfermann, *Phytochem*. **1994**, *35*, 107.
[14] 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, J. Kopka, *FEBS Lett*. **2005**, *579*, 1332.
[15] K. F. Nielsen, J. Smedsgaard, *J. Chromatogr. A* **2003**, *1002*, 111.
[16] K. Böröczky, H. Laatsch, I. Wagner-Döbler, K. Stritzke, S. Schulz, *Chem. Biodiv.* **2006**, *3*, 622.
[17] H.-P. Fiedler, *J. Chromatogr. A* **1984**, *316*, 487.
[18] T. O. Larsen, B. O. Petersen, J. Ø. Duus, D. Sørensen, J. C. Frisvad, M. E. Hansen, *J. Nat. Prod.* **2005**, *68*, 871.
[19] K. Hostettmann, J.-L. Wolfender, S. Rodriguez, *Planta Medica* **1997**, *63*, 2.
[20] P. A. Cremin, L. Zeng, *Analyt. Chem*. **2002**, *74*, 5492.
[21] J. Jakupovic, H. Binkele, D. Wolf, K. Siems, WO Patent No. 002003006152, **2003**.
[22] M. Sandvoss, A. Weltring, A. Preiss, K. Levsen, G. Wuensch, *J. Chromatogr. A* **2001**, *917*, 75.
[23] M. A. Strege, *Anal. Chem.* **1998**, *70*, 2439.
[24] K. Dost, G. Davidson, *Analyst* **2003**, *128*, 1037.
[25] J. Stöckigt, Y. Sheludko, I. Gerasimenko, M. Unger, H. Warzecha, D. Stöckigt, D, *J. Chromatogr. A* **2002**, *967*, 85.
[26] F. V. Ritacco, B. Haltli, J. E. Janso, M. Greenstein, V. S. Bernan, *J. Ind. Microbiol. Biotechnol.* **2003**, *30*, 472.
[27] M. Erhard, H. von Döhren, P. R. Jungblut, *Nature Biotechnology* **1997**, *15*, 906.
[28] Spark Holland Symbiosis^{TM}, *www.spark.nl*.
[29] E. Kovats, *Helv. Chim. Acta* **1958**, *41*, 1915.
[30] J. C. Frisvaad, U. Thrane, *J. Chromatogr.* **1987**, *404*, 195.
[31] D. W. Hill, T. R. Kelley, K. J. Lagner, K. W. Miller, *Anal. Chem.* **1984**, *56*, 2576.
[32] R. Kaliszan, T. Baczek, A. Bucinski, B. Buszewski, M. Sztupecka, *J. Sep. Sci.* **2003**, *26*, 271.
[33] A. Fredenhagen, C. Derrien, E. Gassmann, *J. Nat. Prod.* **2005**, *68*, 385.
[34] B. Yan, L. Fang, M. Irving, S. Zhang, A. M. Boldi, F. Woolard, C. R. Johnson, T. Kshirsagar, G. M. Figliozzi, C. A. Krueger, N. Collins, *J. Comb. Chem*. **2003**, *5*, 547.
[35] G. Guichon, A. Moysan, C. Holley, *J. Liq. Chrom.* **1988**, *11*, 2547.
[36] C. E. Kibbey, *Mol. Divers.* **1999**, *1*, 247.
[37] Y. Lu, H. B. Qu, Y. Y. Cheng, *Chromatographia* **2007**, *65*, 19.
[38] W. Li, J. F. Fitzloff, *J. Liq. Chrom. & Rel. Technol.* **2002**, *25*, 29.
[39] Novelty means not published in 'Dictionary of Natural Products' on CD ROM, Copyright © **1982–2007**, Chapman & Hall/CRC.
[40] W. R. Strohl, *Drug Discov. Today* **2000**, *5*, 39.
[41] M. Ganzera, H. Stuppner, *Curr. Pharm. Anal.* **2005**, *1*, 135.
[42] R. K. Julian, R. E. Higgs, J. D. Gygi, M. D. Hilton, *Anal. Chem*. **1998**, *70*, 3249.
[43] J. B. García, J. R. Tormo, *J. Biomol. Screening* **2003**, *8*, 305.
[44] R. E. Higgs, J. A. Zahn, J. D. Gygi, M. D. Hilton, *Appl. Environ. Microbiol*. **2001**, *67*, 371.