Cluster Analysis as Selection and Dereplication Tool for the Identification of New Natural Compounds from Large Sample Sets

by **Katalin Böröczky**[a])[1]), **Hartmut Laatsch**[b]), **Irene Wagner-Döbler**[c]), **Katja Stritzke**[a]), and **Stefan Schulz***[a])

[a]) Institute of Organic Chemistry, Technical University of Braunschweig, Hagenring 30, D-38106 Braunschweig (fax: (+49) 531-3915272; e-mail: stefan.schulz@tu-bs.de)

[b]) Institute of Organic and Biomolecular Chemistry, University of Göttingen, Tammannstr. 2, D-37077 Göttingen

[c]) Helmholtz Institute of Infection Biology, Division of Cell Biology, Mascheroder Weg 1, D-38124 Braunschweig

Cluster analysis of gas-chromatographic (GC) data of *ca.* 500 bacterial isolates was used as an aid in detection and identification of new natural compounds. This approach reduces the number of GC/MS analysis (dereplication) and concomitantly improves the selection of samples with high probability to contain unknown natural products. Lipophilic bacterial extracts were derivatized and analyzed by GC under standardized conditions. A program was developed to convert chromatographic data into a two-dimensional matrix. Based on the results of hierarchical cluster analysis samples were selected for further investigation by GC/MS and NMR. This approach avoided unnecessary analysis of similar samples. By this method, the unusual oligoprenylsesquiterpenes **1** and **2** as well as new aromatic amides **7** and **8** were identified.

Introduction. – Marine bacteria, among other microorganisms, have been found to be a rich source of new natural products with interesting activities [1]. Isolation of new compounds by chemical screening is conventionally performed by fermentation of the strains, followed by solvent extraction of the broth. In typical workup procedures, a 'defatting' step, *e.g.*, partition between MeOH and cyclohexane, is often included to remove lipophilic material. However, the investigation of the cyclohexane phases can also yield novel compounds with interesting properties [1][2]. Unfortunately, these lipidic extracts not only contain complex mixtures of similar compounds, like fatty acids and hydrocarbons, but also contaminants. Phthalates, for example, are often introduced by the large amounts of solvents typically used during the isolation procedure. The cyclohexane extracts can be conveniently analyzed by GC/MS. Nevertheless, characterization of the components often requires several derivatization procedures including transesterification for the analysis of bound fatty acids, methylation for analysis of free fatty acids and phenols, and/or trimethylsilylation for volatilization of more-polar compounds carrying one or more carboxy, hydroxy, or amino groups. Double-bond localization in long-chain compounds requires derivatization with dimethyl disulfide. For each of the derivatizations, an additional GC/MS analysis has to be performed.

[1]) Present address: 117 Chemical Ecology Laboratory, Department of Entomology, Pennsylvania State University, University Park, PA 16802, USA

In a recent research program on bacterial isolates from the North Sea, we had to analyze more than 500 lipidic extracts of bacteria, resulting in more than 1,500 projected GC/MS analyses. The number of these analyses by far exceeded our GC/MS capabilities. Obviously, a selection and dereplication tool was needed, leading to extracts which had the highest possibility to contain unknown components. Owing to the similarity of the lipidic extracts, we developed a simple GC-based method using cluster analysis as a data-mining tool to select samples of interest for further analysis (selection) and to prevent unnecessary investigation of similar ones (dereplication). GC-Experimental time is generally less costly than GC/MS time, and unlimited access to GC was available to us.

Data-mining methods such as hierarchical cluster analysis (HCA) and principal component analysis (PCA) reveal the basic structure of a data set, if such a structure exists [3]. Objects are hierarchically arranged in clusters by HCA, and the result is presented usually in a dendrogram, where the branch length represents the distance between objects. PCA is a commonly used method to reduce dimensionality. The original variables of a data set are converted to uncorrelated components. The first few components called the principal components should represent more than 50% of the variance of the data. This allows visualization of data structure through presentation of samples as points in a two- or three-dimensional coordinate system with principal components as axes. Both techniques are routinely used to perform classification on chromatographic data sets [4].

We now report the application of HCA to reveal the relation between GC patterns of more than 500 lipophilic bacterial extracts. Exploration of the data set with our method allowed insight into the data structure and facilitated the discovery of novel lipidic natural compounds. We describe the isolation and structure elucidation of the new oligoprenylsesquiterpenes **1** and **2** (see below), as well as the identification of new *N*-(2-arylethyl) amides **7** and **8**.

Results and Discussion. – Generally, bacterial isolates were cultured under different conditions and extracted with AcOEt. After evaporation, these extracts were partitioned between MeOH and cyclohexane. The latter, lipophilic fraction was the object of our investigations. Although primarily lipophilic, the extracts also contained polar compounds, which are difficult to analyze by GC. To volatilize as many compounds as possible and to improve chromatographic peak shape, each sample was treated with 2,2,2-trifluoro-*N*-methyl-*N*-(trimethylsilyl)acetamide (MSTFA) to convert protic functional groups to the corresponding trimethylsilyl (TMS) derivative. Derivatized samples were analyzed once by GC on an apolar *BPX-5* capillary column under standardized conditions, resulting in *ca.* 500 gas chromatograms.

The data of the chromatograms were analyzed using the statistical software packages SPSS. To allow a meaningful analysis, the time axis of each chromatogram was divided into segments of equal length (typically 0.5 min corresponding to the width of broader peaks), and all peak areas in each segment were summarized. Decreasing the segment size led to too many variables, whereas larger segments caused a loss of chemical information. Variation of retention times was accounted for by running hydrocarbon reference samples after every 10 injections. If needed, computational peak alignment was performed after data acquisition using a polynominal fit procedure.

The segment values were normalized for each sample to the sum of peak area, and empty segments as well as segments containing identified impurities (see later) were eliminated. The final number of variables for data analysis varied between 100 and 200, which provided a good ratio to the total number of samples (>500) [5]. In general, the number of variables used in HCA or PCA should be significantly lower than the number of samples [5].

Cluster analysis of converted GC data was then performed, and similarities between samples were presented in a dendrogram (*Fig. 1,a*). We preferred HCA over PCA, since the number of principal components representing more than 50% of the variance was much too high in PCA [6]. Similar approaches including peak alignment have been described in metabolomics research, but, to the best of our knowledge, have not been used for natural product identification so far. Mostly PCA is used in metabolomics research, because, contrary to our case, large data sets of similar chromatograms are analyzed [7].

The resulting dendrogram was used as a tool for the selection of representative samples for in-depth analysis. Samples were randomly chosen from different clusters and analyzed for new or unusual compounds by GC/MS. Thus, only few GC/MS runs were needed until an unknown compound is detected. The GC/MS analysis of a sample from a small cluster (*Fig. 1,a*, sample *258*, strain Bio017) revealed the presence of two late eluting compounds **A** and **B** with a molecular mass of 474 (**A**, $C_{35}H_{54}$, determined by GC/HR-MS: found 474.420, calc. 474.423) and 476 (**B**, $C_{35}H_{56}$, found 476.439, calc. 476.438) with unknown mass spectra containing characteristic ions at *m/z* 119 (*Fig. 2*). These compounds were detected in several strains in the cluster, but were also found later to be present in other strains on different locations on the dendrogram. This distribution is expected, since by-products and concentrations varied between strains, so that the influence of the target compounds on the position of a sample in the dendrogram varies. Therefore, it was necessary to restrict the HCA analysis to the relevant retention time interval, 84–89 min in this instance. As a result, *ca.* 30 strains clustered together with few falsely positive samples (*Fig. 1,b*). Selected *Roseobacter* strains that produced both compounds reliably were refermented on a larger scale, which made isolation of the substances possible.

Hydrogenation of extracts containing both compounds lead to a compound with a molecular mass of 484 and a very strong ion peak at *m/z* 119 (*Fig. 2,c*). This result pointed to five or four double bond equivalents in **A** and **B**, respectively. The mass spectra resembled those of phenylalkanes, which show a base peak at *m/z* 91, with two additional Me groups on the ring or the C(α)-atom. Less than 1 mg of each compound was then isolated by normal-phase HPLC and subjected to extensive NMR studies (*Tables 1* and *2*). Compound **A** possessed a 4-methyl-(1-methylalkyl)-substituted aromatic system connected to an oligoprenyl side chain. The aromatic head group can easily form the observed ion at *m/z* 119 by cleavage next to the α-Me group. According to the ^{13}C-NMR data, the prenyl groups were (all-*E*)-configured. Compound **B** was significantly less stable than **A** and slowly oxidized upon standing yielding the latter compound. The NMR data suggested a 4-methyl-1-(methylalkyl)cyclohexa-1,4-diene headgroup for **B**. Both MS and NMR data of the two compounds were strikingly similar to those of β-curcumene and *ar*-curcumene [8]. Therefore, compounds **A** and **B** were identified to be tetraprenyl-*ar*-curcumene (**1**) and tetraprenyl-β-curcumene (**2**).

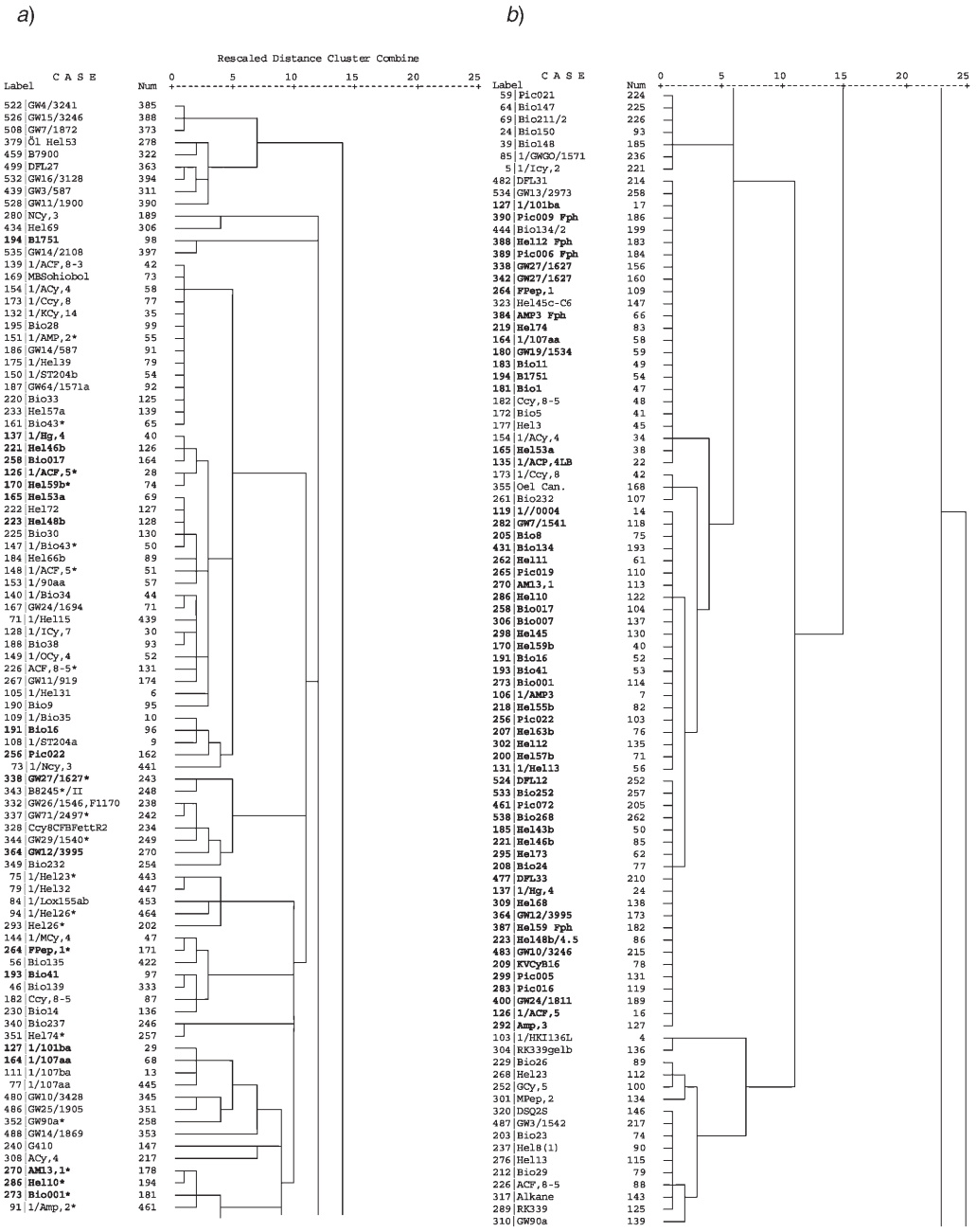


Fig. 1. a) *Part of the dendrogram obtained by HCA* (Squared Euclidean Distance, Average Linkage) *of* ca. *500 samples.* Strains containing **1** and **2** are marked in bold. b) *Dendrogram obtained by using data from the retention-time window between 84 and 89 min only.*

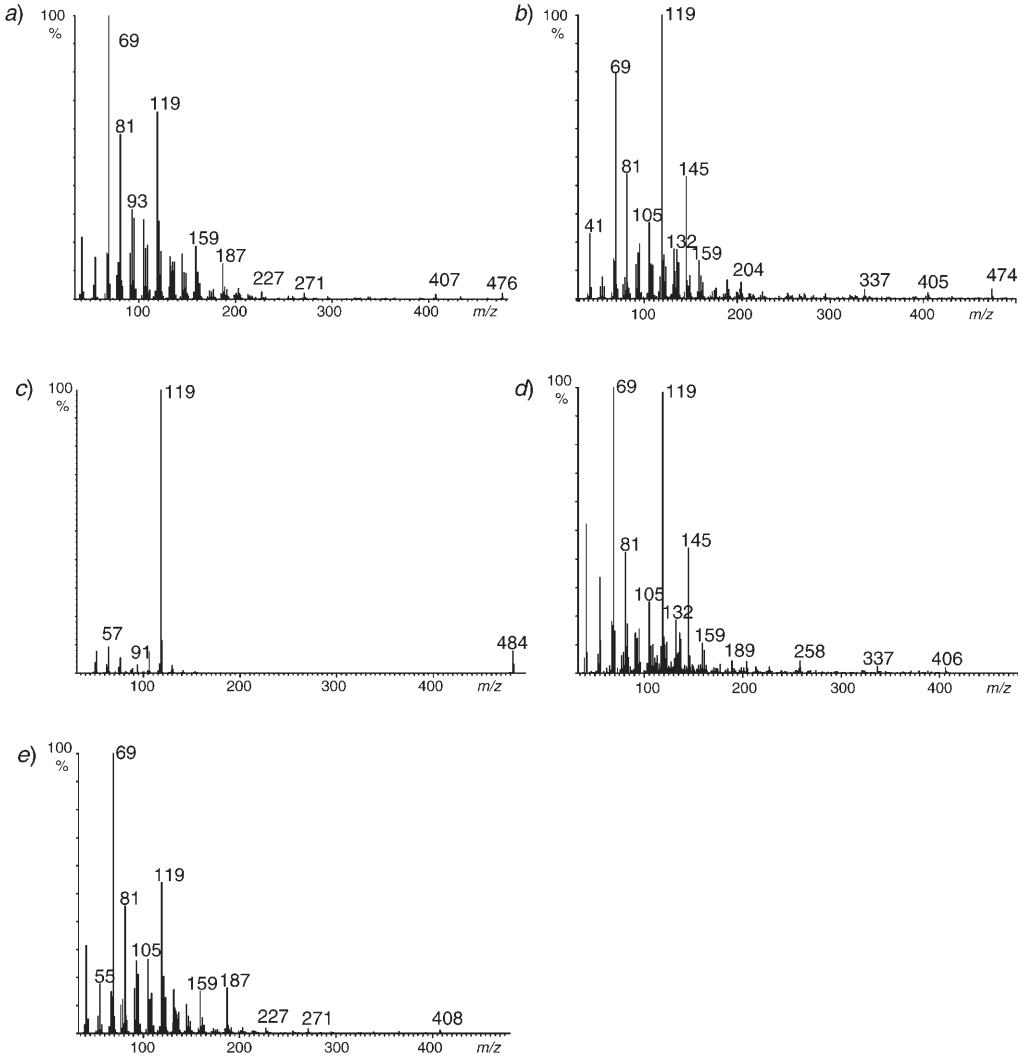
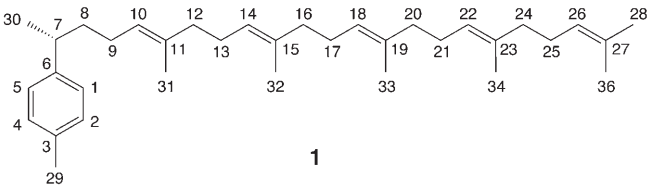


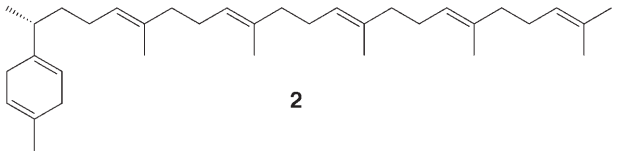
Fig. 2. *Mass spectra of* **A** (**1**; *a*), **B** (**2**; *b*), *hydrogenated* **A** *and* **B** (*c*), **3** (*d*), *and* **4** (*e*)

Compounds **1** and **2** were often accompanied by two minor constituents lacking one prenyl group. These compounds were identified as triprenyl-*ar*-curcumene (**3**) and triprenyl-β-curcumene (**4**), based on their mass spectra (*Fig. 2*).

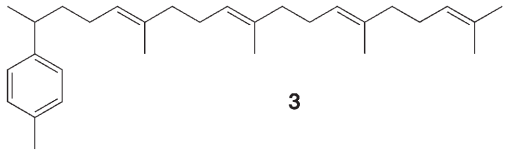
The absolute configuration of **1** was determined by oxidative degradation. Ruthenium tetroxide [9] cleaved the aromatic ring as well as the C=C bond to form Dimethyl 2-methylpentanedioate (**6**), thus preserving the stereogenic center at C(7), although in low yield (*Scheme*). Degradation of (*S*)- and (*R*)-citronellene (**5**) under similar conditions gave access to (*S*)-**6** and (*R*)-**6**, respectively, as reference compounds. After methylation, GC/MS analysis using a chiral cyclodextrin phase revealed the



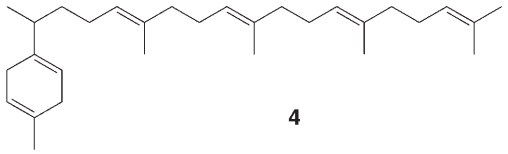
1



2



3



4

Table 1. *^{1}H- and ^{13}C-NMR Data of* **1** *and* ar-*Curcumene* (δ in ppm)

Position[a])	**1**[b])		*ar*-Curcumene[c])	
	δ(H)	δ(C)	δ(H)	δ(C)
30	1.20 (*d*)	23.1	1.19 (*d*)	22.9 (*q*)
31	1.52 (*s*)	16.1	1.50 (*s*)	17.7 (*q*)
32–35	1.59 (*s*)	16.2 (3 C), 17.8	–	–
28	1.67 (*s*)	25.8	1.65 (*s*)	25.8 (*q*)
8, 12, 16, 20, 24	1.82–2.00 (*m*)	38.9, 40.3, 40.2 (3 C)	1.51–1.65 (*m*)	38.8 (*t*)
9, 13, 17, 21, 25	2.02–2.10 (*m*)	27.2 (5 C)	1.81–1.89 (*m*)	26.6 (*t*)
29	2.29 (*s*)	21.1	2.30 (*s*)	21.0 (*q*)
7	2.60–2.70 (*m*)	39.5	2.60–2.67 (*m*)	39.4 (*d*)
10, 14, 18, 22, 26	5.07–5.14 (*m*)	125.2, 124.7–124.9 (3 C), 125.0	5.04–5.10 (*m*)	125.1 (*d*)
1, 5	7.07 (*dd*)	n.a.	7.04 (*s*)	127.3 (*d*)
2 ,4		129.4 (2 C)		129.3 (*d*)

[a]) For numbering, see structural formula of **1**. [b]) ^{1}H-NMR was recorded in CD_2Cl_2 at 400 MHz. The ^{13}C-NMR chemical-shift values are calculated from the HMBC experiments recorded in C_6D_6. No correlation was observed between H–C(7) and C(1/5). [c]) ^{1}H-NMR in $CDCl_3$, ^{13}C-NMR in C_6D_6; see [8].

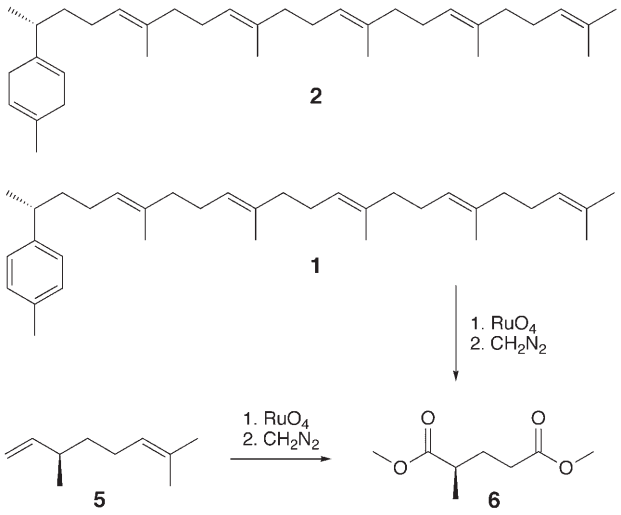
absolute configuration to be (*R*). We assume that the terpene **2** has the same configuration, because this compound was slowly oxidized to **1** during the isolation procedure. Terpene cyclases in plants are known to produce a hexa-1,4-diene ring,

Table 2. *^{1}H-NMR Data of* **2** *and β-Curcumene* (δ in ppm)

Position[a])	**2**[b])	β-Curcumene[c])
	δ(H)	δ(H)
30	0.98 (*d*)	0.99 (*d*)
31–35	1.58 (s), 1.60 (*s*)	1.20–1.60 (*m*)
8, 12, 16, 20, 24	1.87–2.02 (*m*)	1.20–1.60 (*m*)
28, 29	1.67 (*s*), 1.65 (*s*)	1.67 (*s*)
7, 9, 13, 17, 21, 25	2.02–2.13 (*m*)	2.09 (*m*), 1.90 (*m*)
2, 5	2.56 (br. *s*)	2.58 (*m*)
10, 14, 18, 22, 26	5.06–5.15 (*m*)	5.09 (*m*)
1, 4	5.42 (*m*)	5.42 (*s*)

[a]) For atom numbering, see structural formula of **2**. [b]) Recorded in CD_2Cl_2 at 400 MHz. [c]) In $CDCl_3$; see [8].

Scheme. *Chemical Correlation of* **1** *and* **2** *for the Determination of Their Absolute Configuration*



which can be aromatized [10]. It is possible that the aromatic compound **1** is a nonenzymatic oxidization product of **2**, but enzymatic formation cannot be excluded. All terpenes **1**–**4**, showing an unusual arrangement of six or seven head-to-tail isoprene units, belong to the rare class of oligoprenylsesquiterpenes. Only few of these compounds are known so far from nature, *i.e.*, poduran from a springtail [11] and three furospinosulins from a sponge [12]. The widespread occurrence of **1** and **2** in unrelated bacterial strains (*Table 3*) and their easy oxidation may point to an important function in the physiology of the bacteria.

In the dendrogram of the entire data set, clusters are not all well-defined. Nevertheless, grouping of similar samples was observed due to characteristic GC peaks of compounds present in higher concentrations. As an example, *N*-acyl derivatives of 2-

Table 3. *Phylogenetic Affiliations of Strains from the North Sea Containing Both Compounds* **1** *and* **2**[a])

Sample	*Species*
Bio8	*Cytophaga* sp.
Hel11	*Vibrio* sp.
AM13,1	*Cytophaga* sp. (AM13,1 = Pic84)
Hel10	*Jannaschia helgolandensis*
Bio017	*Brevundimonas vesicularis*
Hel45	*Oceanibulbus indolifex*
Bio007	*Sulfitobacter pontiacus*
Hel12	*Microbacterium* sp.
1/AMP3	*Cytophaga* sp. (AMP,3 = Pic86 = Pic400 = Pic302)
Hel59b	*Marinomonas* sp.
Bio41	*Clavibacter* sp.
Bio001	*Cytophaga* sp.
Hel43b	*Jannaschia helgolandensis*
Hel46b	*Sulfitobacter pontiacus*
1/Hg,4	*Flavobacterium* sp. (Pic265)
Hel68	*Halomonas* sp.
Hel48b/4.5	*Pseudoalteromonas* sp.
Pic016	*Cyclobacterium marinum*
Hel73	*Brevibacter linens*

[a]) The assignments were made based on similarity of 16sRNA. Strains Pic019, Hel63b, Hel55b, Pic022, Bio16, Bio24, Pic005, Hel57b, 1/ACF,5, GW7/1541, GW14/1750, and KVCyB16 were not sequenced.

phenylethylamine and other biogenic amines were identified in numerous samples from *Cytophaga* and *Frigoribacter* strains. Higher concentrations of *N*-acetyl-2-phenylethylamine, found earlier by us in extracts of marine *Streptomyces* strains [13], caused the grouping of samples containing these amides. In the samples FPep1 (*Cytophaga*/*Flexibacterium*) and Pic006 (*Frigoribacter*) previously unknown 2-phenylethyl amides of typical bacterial fatty acids (C_{14}, iso-C_{15}, anteiso-C_{15}, C_{15}, iso-C_{16}, C_{16}, and iso-C_{17}; **7b**–**7h**) and of senecioic acid (**7a**) were identified. The *Vibrio* strain Hel11 contained ten acylated tyramines **8**, of which the *N*-propanoyl (**8b**), *N*-butanoyl (**8d**), *N*-(3-methylbutanoyl) (**8f**), *N*-(3-methylbut-2-enoyl) (**8g**), *N*-(4-methylpentanoyl) (**8h**), *N*-[3-(methylsulfanyl)propanoyl] (**8i**), and *N*-phenylacetyl (**8j**) amides have not been reported before. These amides can be readily identified by GC/MS and use of a GC retention-index system which allows determination of Me-group positions in long alkyl chains [14]. The mass spectra of 2-phenylethyl amides are characterized by a base ion peak at *m*/*z* 104, a *McLafferty* rearrangement ion peak at *m*/*z* 163, an acylium ion, and the loss of 91 amu from the molecular ion (*Fig. 3*). The position of the Me group in the branch can be easily deduced using the increment system introduced by us earlier [14]. The related acylated tyramines possess a small molecular ion and a base ion peak at *m*/*z* 120. The fatty acid part can be deduced from the ion pair acylium ion/acylium ion +18. Characteristic cleavages near substituents on the chain indicate their position.

Additional distinguishing components were (*E*)- and (*Z*)-octadec-9-enamide, produced by several bacteria. On the other hand, the extracts contained several components unspecific for bacteria: industrial plasticizers, sugars, aromatic impurities,

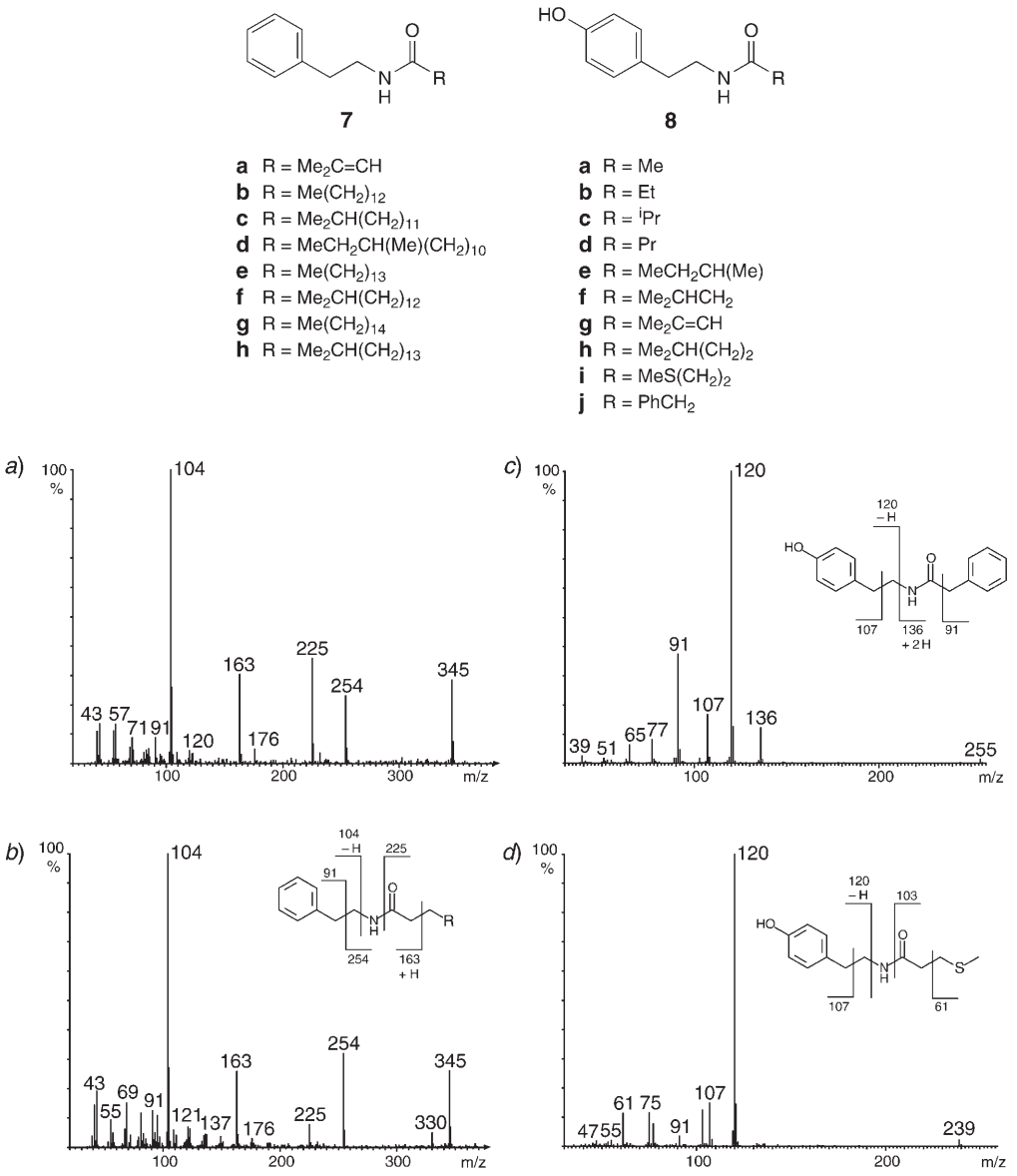


Fig. 3. *Mass spectra and fragmentation pattern of different amides identified in strain F/Pep1 and Pic006. N*-(2-Phenylethyl)pentadecanamide (**7e**; *a*), 13-Methyl-*N*-(2-phenylethyl)tetradecanamide (**7c**; *b*), *N*-[2-(4-hydroxyphenyl)ethyl]-2-phenylethanamide (**8j**; *c*), and *N*-[2-(4-hydroxyphenyl)ethyl]-3-(methylsulfanyl)propanamide (**8i**, *d*).

and some piperazine-diones. Proline-derived piperazine-diones, often occurring as mixtures of diastereoisomers, were present in many extracts as well as in the fermentation medium, indicating their non-natural formation. These substances are

distributed throughout the whole time axis, thus suppressing the influence of actual natural products for the cluster analysis. To avoid this effect, segments containing phthalates and the most common piperazine-diones were eliminated before data analysis. This procedure provided better results and cleaner clusters (see *Fig. 1,b*).

Conclusions. – In summary, we developed a new explorative data-analysis method based on a standardized GC screening combined with data conversion and HCA to evaluate large sets of bacterial extracts. The following general method is the most appropriate for dereplication to avoid superfluous analyses and for fast selection of strains containing new natural products. First, a cluster is arbitrarily selected in the original dendrogram, and a sample from this cluster is analyzed by GC/MS. When an unknown component is detected, a refined dendrogram is produced based on GC data of the retention-time interval around this component. Finding all the samples containing the compound of interest, this way saves further exploration of the original data set. Finally, the structure of the compound is elucidated by appropriate methods. Additional new samples are easily integrated into the data set and allow evaluation whether further analysis seems to be promising for chemical screening. As an example, the later analyzed sample Pic004 clustered closely to the sample Pic006, which was already analyzed by GC/MS. Because of its position in the dendrogram, no further analysis had to be performed, because no new compounds could be expected.

The analysis of the dendrograms revealed interesting facts on the distribution of metabolites in this large sample set. For example, **1** and **2** are present in more than 6% of the strains of mostly unrelated bacteria. Nevertheless, a phylogenetic classification of the strains based on the dendrograms is not justified because samples were obtained from different fermentation media, incubation times, *etc.* Cluster analysis, applied usually to classify chromatographic data, was used here as a visualization tool to reveal the degree of similarity between GC patterns in a very large data set. Similar approaches have been used in metabolomics research [7]. Here, in contrast to our approach, PCA is preferred because often closely related data sets are investigated. Although it is still a challenge to minimize the effect of the fermentation medium on data analysis, the examples presented demonstrate the effectiveness of this tool in discovering novel natural compounds.

This work was funded by a grant from the *Volkswagenstiftung* within the *Lower Saxony Cooperative Research Project* on Marine Biotechnology.

Experimental Part

General. Solvents used for extraction of the broth and for synthetic work were of chemical grade (*ACROS*), hexane was of HPLC grade (*ACROS*), CH_2Cl_2 and pentane were of GC trace analysis grade (*Suprasolv*, *Merck*). Chemicals used were 2,2,2-trifluoro-*N*-methyl-*N*-(trimethylsilyl)acetamide (MSTFA; *CE Chromatography Service*), Pd/C 10% (*ACROS*), $RuCl_3 \cdot (H_2O)_n$ (*Aldrich*), and H_5IO_6 (*Merck*). Column chromatography (CC) was performed with silica gel (230–400 mesh).

Isolation and Fermentation of Bacteria. Most strains investigated were isolated from diverse habitats of the North Sea, including dinoflagellate cultures (strain designation DFL), picoplancton (Pic), water column samples (Hel), biofilm developed on glass plates (Bio), *Laminaria* surfaces (LM) and diatoms (DT). Details of the isolation procedure can be found in [15]. Strains were identified by sequencing the

16S rRNA gene as described in [15]. Seawater-based solns. like *Marine Broth 2216* (*Difco*) or *LBSS* (26 g of *Luria Bertani* broth (*Sigma*), 17.08 g of sea salts (*Sigma*) per 1000 ml of dist. water were used as fermentation medium. The broth was extracted with AcOEt, filtered, and the solvent was removed. The residue was partioned between MeOH and cyclohexane. The cyclohexane phase was used for further experiments.

Derivatization. Extracts were dissolved in CH_2Cl_2 in an approximate concentration of 1 mg/100 µl. This soln. (50 µl) was mixed with MSTFA (100 µl) in a 2-ml vial with micro-inset. The vial was closed tightly and maintained at 50° for 1 h. The remaining MSTFA and the solvent were evaporated at 50° under a gentle stream of N_2. The samples were redissolved in CH_2Cl_2 (100 µl) prior to GC analysis.

GC Screening. Samples were analyzed with a *GCTop 8000* gas chromatograph connected to an *AS800* autosampler (*ThermoQuest*). H_2 was used as carrier gas at a flow rate of 1 ml/min at 100°. Components were separated on a *BPX-5* (25 m × 0.25 mm × 0.25 µm) cap. (*SGE Inc.*) under identical conditions (injector temp. 250°, splitless injection with valve time of 0.75 min, FID temp. 280°, oven temp. program: 2 min at 50° (2 min), then with 3°/min to 320°. Data acquisition was performed by the software *ChromCard 1.19* (*ThermoQuest*). Retention times were regularly checked with hydrocarbon test solns. and adjusted by post-analysis peak alignment.

GC/MS Analysis. A *HP6890* gas chromatograph coupled with a *HP5973 MSD* and a *HPG1513A* autosampler (*Hewlett Packard*) were used. Compounds were separated on a *BPX-5* (25 m × 0.25 mm × 0.25 µm) cap. (*SGE Inc.*) with He as carrier gas at a constant flow rate of 1 ml/min under the same conditions as described for the GC analysis. Data were recorded by the software ChemStation v. A.03.00 (*Hewlett Packard*).

GC/HR-MS Analysis. High-resolution (HR) MS data were acquired on a *MAT95XLT* (*ThermoQuest Finnigan*) apparatus equipped with a *BPX-5* (25 m × 0.25 mm × 0.25 µm) cap. at 1 s/decade in a magnetic scan range of 35–550 amu with a resolution of 7,000–9,000 (10% valley).

NMR Analysis. NMR Spectra were recorded with a *DRX-400* (^{1}H: 400 MHz, ^{13}C: 100 MHz) instrument (*Bruker*). The internal standard was Me_4Si (TMS).

Data Conversion. GC Report files generated by ChromCard (*CE Instruments*) were edited using the self-made program ChromConv written in *Perl*. The user is allowed to choose which section of the chromatogram should be involved in the process, which normalization method should be used, and which segments containing impurities should be deleted. The peak area was summed up within 0.5-min time segments, empty or unwanted segments were deleted, and values were normalized to the sum. As a result, a text file was created containing a 2D matrix with the samples in rows and the calculated segment values in columns.

Statistical Analysis. SPSS (*SPSS Inc.*) was used for the statistical analysis and generation of dendrograms. Euclidean distance measures were applied to calculate the distance between samples. Since the structure of the data set was unknown, average linkage was chosen for clustering the samples. Several other statistical methods including PCA were tried, but yielded inferior results.

Isolation and Structure Elucidation of the Tetraprenylsesquiterpenes **1** *and* **2**. Several lipophilic extracts from the different fermentations of the strains AMP,3 and Hel59 were separated by CC. Pentane was used for the elution of **1** and **2**. Squalene eluted together with both compounds. Thus, a further purification step by HPLC was necessary. HPLC isolation was performed on a Spectra System (*ThermoQuest*) equipped with a *P4000* pump, an *AS3000* autosampler, and a *SN4000* diode array detector. A *Superspher Si60* (250 × 4 mm, 4 µm) column (*Merck*) was used to separate the compounds with hexane as the mobile phase. Flow rate was set to 1 ml/min. Fractions were collected manually. Squalene, **1**, and **2** were successfully separated by this procedure. ChromQuest 2.1 (*ThermoQuest*) was used for data acquisition.

Hydrogenation. Terpene **2** was dissolved in 0.5 ml of isooctane in a 2-ml vial at r.t. and stirred continuously. A tiny portion of Pd/C (10%) was added, and the mixture was stirred for 10 min. H_2 was introduced into the vial *via* stainless steel tubing and a needle at the end. A second needle provided gas purge through the septum. The pressure was maintained at 0.1 bar for 45 min before removing the extra cannula. The H_2 valve was closed and the mixture was stirred for another 15 min under H_2 atmosphere. The mixture was filtered, and the filtrate was concentrated under a gentle stream of N_2 at 50°. The residue was dissolved in CH_2Cl_2.

Oxidative Degradation. A mixture containing the isolated natural product **1** (*ca.* 0.5 mg, 1 μmol), H_5IO_6 (8 mg, 35 μmol, 35 equiv.), MeCN (20 μl), CCl_4 (20 μl), and dist. H_2O (30 μl) was stirred 5 min in a closed GC vial with a needle opening at r.t. A minute amount of $RuCl_3 \cdot (H_2O)_n$ was introduced, and the whole mixture was stirred for 5 h. Brine (200 μl) was added, and the mixture was extracted with CH_2Cl_2 (200 μl). The org. phase was concentrated in a gentle stream of N_2 at 30° and derivatized with a sat. CH_2N_2 soln. in CH_2Cl_2 (500 μl). The solvent was evaporated, and the residue was dissolved in CH_2Cl_2 (10 μl), resulting in a soln. containing **6**. (+)- and (−)-citronellene (**5**) were transformed into (*R*)-**6** and (*S*)-**6** in an identical manner.

Chiral GC Analysis. The enantiomers of **6** were separated on a *hydrodex-6-TBDMS* (15 m × 0.25 mm; *Macherey & Nagel*) at 60° with 67 cm/s H_2 gas flow. Retention times: (*S*)-**6** 36.5 min, (*R*)-**6** 37.2 min.

3-Methyl-N-*(2-phenylethyl)but-2-enamide* (**7a**). EI-MS: 203 (25, M^+), 104 (100), 91 (22) , 85 (4), 84 (5), 77 (4), 71 (4), 65 (7), 55 (13), 43 (19).

N-*(2-Phenylethyl)tetradecanamide* (**7b**). EI-MS: 331 (23, M^+), 240 (32), 211 (16), 163 (32), 104 (100), 91 (12), 71 (8), 69 (6), 57 (9), 55 (8), 43 (10), 41 (8).

13-Methyl-N-*(2-phenylethyl)tetradecanamide* (**7c**). EI-MS: 345 (26, M^+, 26), 330 (5), 254 (32), 225 (36), 163 (28), 104 (100), 91 (13), 71 (12), 69 (13), 55 (10), 43 (19), 41 (14).

12-Methyl-N-*(2-phenylethyl)tetradecanamide* (**7d**). EI-MS: 345 (25, M^+), 316 (3), 254 (37), 225 (36), 163 (28), 104 (100), 91 (10), 71 (5), 69 (5), 57 (6), 55 (7), 43 (12), 41 (7).

N-*(2-Phenylethyl)pentadecanamide* (**7e**). EI-MS: 345 (28, M^+), 254 (23), 225 (4), 163 (26), 104 (100), 91 (9), 71 (11), 69 (8), 57 (16), 55 (12), 43 (17), 41 (12).

14-Methyl-N-*(2-phenylethyl)pentadecanamide* (**7f**). EI-MS: 359 (22, M^+), 344 (5), 268 (31), 239 (7), 163 (29), 104 (100), 91 (10), 71 (7), 69 (7), 57 (12), 55 (11), 43 (19), 41 (10).

N-*(2-Phenylethyl)hexadecanamide* (**7g**). EI-MS: 359 (18, M^+), 268 (23), 239 (46), 163 (32), 104 (100), 91 (7), 71 (9), 69 (10), 57 (16), 55 (16), 43 (17), 41 (15).

15-Methyl-N-*(2-phenylethyl)hexadecanamide* (**7h**). EI-MS: 373 (32, M^+), 358 (6), 282 (47), 253 (9), 163 (32), 104 (100), 91 (9), 71 (6), 69 (6), 57 (12), 55 (9), 43 (15), 41 (13).

N-*[2-(4-Hydroxyphenyl)ethyl]ethanamide* (**8a**). EI-MS: 179 (2, M^+), 120 (100), 107 (34), 91 (4), 77 (9), 43 (12).

N-*[2-(4-Hydroxyphenyl)ethyl]propanamide* (**8b**). EI-MS: 193 (2, M^+), 120 (100), 107 (24), 91 (4), 77 (9), 74 (5), 57 (13).

N-*[2-(4-Hydroxyphenyl)ethyl]-2-methylpropanamide* (**8c**). EI-MS: 207 (2, M^+), 120 (100), 107 (18), 91 (4), 88 (10), 77 (8), 71 (11), 43 (12).

N-*[2-(4-Hydroxyphenyl)ethyl]butanamide* (**8d**). EI-MS: 207 (2, M^+), 120 (100), 107 (18), 91 (3), 88 (8), 77 (8), 71 (12), 43 (13).

N-*[2-(4-Hydroxyphenyl)ethyl]-2-methylbutanamide* (**8e**). EI-MS: 221 (2, M^+), 120 (100), 107 (14), 102 (17), 91 (4), 85 (9), 77 (7), 57 (25), 41 (6).

N-*[2-(4-Hydroxyphenyl)ethyl]-3-methylbutanamide* (**8f**). EI-MS: 221 (2, M^+), 120 (100), 107 (17), 102 (11), 91 (3), 85 (12), 77 (8), 57 (13), 43 (3), 41 (5).

N-*[2-(4-Hydroxyphenyl)ethyl]-3-methylbut-2-enamide* (**8g**). EI-MS: 219 (4, M^+), 120 (100), 107 (15), 100 (41), 98 (2), 83 (96), 77 (10), 55 (28), 39 (6).

N-*[2-(4-Hydroxyphenyl)ethyl]-4-methylpentanamide* (**8h**). EI-MS: 235 (1, M^+), 120 (100), 116 (11), 107 (14), 99 (7), 91 (3), 77 (6), 71 (6), 55 (3), 43(11).

N-*[2-(4-Hydroxyphenyl)ethyl]-3-(methylsulfanyl)propanamide* (**8i**). EI-MS: 239 (1, M^+), 120 (100), 107 (16), 103 (12), 91 (3), 77 (6), 75 (11), 61 (11), 47 (3).

N-*[2-(4-Hydroxyphenyl)ethyl]-2-phenylethanamide* (**8j**). EI-MS: 255 (2, M^+), 136 (12), 120 (100), 107 (17), 99 (7), 91 (37), 77 (8), 65 (7), 39 (2).

REFERENCES

[1] J. W. Blunt, B. R. Copp, M. H. G. Munro, P. T. Northcote, M. R. Prinsep, *Nat. Prod. Rep.* **2005**, *22*, 15, and earlier reviews in this series; R. H. Feling, G. O. Buchanan, T. J. Mincer, C. A. Kauffman, P. R. Jensen, W. Fenical, *Angew. Chem.* **2003**, *115*, 369.
[2] K. Stritzke, S. Schulz, H. Laatsch, E. Helmke, W. Beil, *J. Nat. Prod.* **2004**, *67*, 395.
[3] 'The Elements of Statistical Learning', Eds. T. Hastie, R. Tibshirani, J. Friedman, Springer, New York, USA, 2001.
[4] B. K. Lavine in 'Encyclopedia of Analytical Chemistry', Ed. R. A. Meyers, John Wiley & Sons Ltd, Chichester, UK, 2000, p. 9689; M. Otto, 'Chemometrics: Statistics and Computer Application in Analytical Chemistry', Wiley-VCH, 1999; Y. Tikunov, A. Lommen, C. H. Ric de Vos, H. A. Verhoeven, R. J. Bino, R. D. Hall, A. G. Bovy, *Plant Physiology* **2005**, *139*, 1125.
[5] K. Varmuza, *Anal. Chem., Symp. Ser.* **1983**, *15*, 19.
[6] 'Intelligent Data Analysis', Eds. M. Berthold, David J. Hand, Springer, 2003, p. 99.
[7] A. L. Duran, J. Yang, L. Wang, L. W. Sumner, *Bioinformatics* **2003**, *19*, 2283; M. Glinski, W. Weckwerth, *Mass Spectrom. Rev.* **2006**, *25*, 173.
[8] D. Joulain, W. A. König, 'The atlas of spectral data of sesquiterpene hydrocarbons', E. B.-Verlag, Hamburg, 1998.
[9] H. J. Carlsen, T. Katsuki, V. S. Martin, K. B. Sharpless, *J. Org. Chem.* **1981**, *46*, 3939; M. T. Nunez, V. S. Martin, *J. Org. Chem.* **1990**, *55*, 1928.
[10] M. L. Wise, R. Croteau, in 'Comprehensive Natural Products Chemistry, Vol. 2', Ed. D. E. Cane, Elsevier, New York, 1999, p. 97.
[11] S. Schulz, C. Messer, K. Dettner, *Tetrahedron Lett.* **1997**, *53*, 2077.
[12] G. Cimino, S. De Stefano, L. Minale, *Tetrahedron* **1972**, *28*, 1315.
[13] R. P. Maskey, P. N. Asolkar, E. Kapaun, I. Wagner-Döbler, H. Laatsch, *J. Antibiot.* **2002**, *55*, 643.
[14] S. Schulz, *Lipids* **2001**, *36*, 637; J. S. Dickschat, S. C. Wenzel, H. B. Bode, R. Müller, S. Schulz, *ChemBioChem* **2004**, *5*, 778.
[15] M. Allgaier, H. Uphoff, A. Felske, I. Wagner-Döbler, *Appl. Environ. Microbiol.* **2003**, *69*, 5051.

Received May 8, 2006