Journal of Microbiological Methods 88 (2012) 83–97

Contents lists available at SciVerse ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

MALDI-typing of infectious algae of the genus *Prototheca* using SOM portraits

Henry Wirth [a,b,c,*], Martin von Bergen [b,d], Jayaseelan Murugaiyan [b,e], Uwe Rösler [e], Tomasz Stokowy [f,g], Hans Binder [a,c,*]

[a] Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107, Härtelstr. 16-18, Leipzig, Germany
[b] Helmholtz-Centre for Environmental Research; Department for Proteomics, D-04318 Leipzig, Permoserstr. 15, Germany
[c] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE); Universität Leipzig, D-4103 Leipzig, Philipp-Rosenthalstr. 27, Germany
[d] Helmholtz-Centre for Environmental Research, Department for Metabolomics, D-04318 Leipzig, Permoserstr. 15, Germany
[e] Free University of Berlin, Institute for Animal and Environmental Hygiene, D-10117 Berlin, Luisenstraße 56, Germany
[f] Silesian University of Technology, Department of Automatic Control, ul.Akademicka 16, 44-100 Gliwice, Poland
[g] Cancer Center and Institute of Oncology Gliwice branch, Nuclear Medicine and Endocrine Oncology Department, ul. Wybrzeże Armii Krajowej 15 44-101 Gliwice, Poland

ARTICLE INFO

Article history:
Received 8 August 2011
Received in revised form 17 October 2011
Accepted 20 October 2011
Available online 26 October 2011

Keywords:
MALDI-typing
Mass spectrometry
Peaklist
Prototheca
SOM
Classification

ABSTRACT

Background: MALDI-typing has become a frequently used approach for the identification of microorganisms and recently also of invertebrates. Similarity-comparisons are usually based on single-spectral data. We apply self-organizing maps (SOM) to portray the MS-spectral data with individual resolution and to improve the typing of Prototheca algae by using meta-spectra representing prototypes of groups of similar-behaving single spectra.
Results: The MALDI-TOF peaklists of more than 300 algae extracts referring to five Prototheca species were transformed into colored mosaic images serving as molecular portraits of the individual samples. The portraits visualize the algae-specific distribution of high- and low-amplitude peaks in two dimensions. Species-specific pattern of MS intensities were readily discernable in terms of unique single spots of high amplitude MS-peaks which collect characteristic fingerprint spectra. The spot patterns allow the visual identification of groups of samples referring to different species, genotypes or isolates. The use of meta-peaks instead of single-peaks reduces the dimension of the data and leads to an increased discriminating power in downstream analysis.
Conclusions: We expect that our SOM portray method improves MS-based classifications and feature selection in upcoming applications of MALDI-typing based species identifications especially of closely related species.

1. Introduction

Identification of microorganisms was for a long time performed by applying specifically designed differential culture conditions but has over the last 20 years become the domain of molecular based methods, like PCR (Erlich et al., 1991; Irenge et al., 2010), microarray hybridization (Pozhitkov et al., 2007; Pozhitkov et al., 2011), DNA-sequencing (Tautz et al., 2002), FTIR spectroscopy (Helm et al., 1991) or mass spectrometry (Demirev et al., 1999). The former methods analyze DNA- and RNA-nucleotide sequences with typically rich phylogenetic information content. Instead, the latter two methods usually apply to metabolites and proteins. Whereas metabolites are limited in their phylogenetic informational content diversity of proteins is comparable with the richness of the genetic or transcriptomic information.

The method of MALDI-typing was developed for the rapid identification of bacterial samples (Bright et al., 2002). It is based on most simple extraction procedures like using acidic solution that can afterwards directly be mixed with MALDI-matrices. This allows a rapid preparation that can be performed in an automated way. Besides the extraction also the ionisation process used in MALDI-MS has a crucial impact on the selection of detected proteins and peptides. The rules that have been found include a general preference for proteins in the range of 2–20 kDa with an inverse correlation between the mass and the intensity of detection and a sometimes very strong dependency on the used matrix substance.

The phyla, from which species identification by mass spectrometry of proteins is reported, range from microorganisms as bacteria (Jehmlich et al., 2009), yeast (van Veen et al., 2010) and algae towards small invertebrates (Campbell, 2005; Feltens et al., 2010) and even vertebrates (Mazzeo et al., 2008). MALDI-typing poses a cost efficient alternative to PCR based approaches due to the suitability of simple and fast extraction and measurement procedures.

In parallel with the further improvement of the experimental protocols of cultivation, extraction and sample preparation and of the

* Corresponding authors at: Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Härtelstr., 16-18, Germany.
E-mail addresses: wirth@izbi.uni-leipzig.de (H. Wirth), binder@izbi.uni-leipzig.de (H. Binder).

0167-7012/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2011.10.013

MS-techniques applied, data processing and analysis is a crucial step in typing tasks. Besides primary spectral cleaning and peak extraction it requires appropriate classification algorithms allowing to distinguish the spectra of different species with maximum resolution. Also visualization of the high-dimensional data given by the amplitudes of a few thousands MS-peaks in hundreds of samples is an important requirement for comprehensive analytics.

Machine learning using self-organizing maps (SOM) allows portraying the molecular landscape with individual resolution. The basics of the method were developed by Kohonen about 30 years ago (Kohonen, 1982). It projects data from the original high dimensional space to reference vectors of lower dimension. SOM analysis was successfully applied to high-dimensional microarray gene expression data using either a gene-centered perspective to cluster genes (Tamayo et al., 1999) or a sample-centered mode to classify samples into diagnostic groups (Golub et al., 1999; Covell et al., 2003; Buckhaults et al., 2003). SOM machine learning was also applied to genetic SNP-data of human populations (Binder et al., 2011) and to NMR and mass spectrometry data in the context of metabolic and proteomic profiling. These latter applications in the first instance address methodical issues of machine learning such as the automated characterization of subclass-related metabolic interactions (Suna et al., 2007), multi-factorial classification of metabolites (Wongravee et al., 2010) and metabolite profiling using one-dimensional (Meinicke et al., 2008) or fuzzy-labeled (Villmann et al., 2008) SOMs.

The SOM method can be configured also in such a way that it combines both, the sample- and feature-centered perspectives (Nikkilä et al., 2002; Wang et al., 2002; Eichler et al., 2003). This specific approach decodes the pattern of thousands of single features per sample into a two-dimensional mosaic pattern which allows the sample-to-sample comparison by direct visual inspection. It has been demonstrated that such SOM portraits are featured by several important benefits (Nikkilä et al., 2002; Wang et al., 2002; Eichler et al., 2003): (i) they provide an individual visual identity for each sample; (ii) they reduce the dimension of the original data; (iii) they preserve the information richness of the original data allowing the detailed, multivariate explorative comparisons between samples, (iv) they are highly intuitive not-requiring specific knowledge of the underlying algorithmic kernel of the method, and (v) they can be treated as new, complex objects for next level analysis in terms of visual recognition.

In this publication we apply this portraying-approach to MALDI-TOF typing of *Prototheca* species using previously published data (von Bergen et al., 2009). We will demonstrate that SOM-portraying not only improves the comparability of the peaklists of different species in an intuitive fashion but also extracts meta-spectra representing the fundamental set of peaks inherent in the data. We show that classification based on these metaspectra clearly outperforms classification based on single spectra.

The application of our method to colorless algae of the genus *Prototheca* within the Chlorellaceae family is motivated by the fact that these algae are the only known plants that cause infections in humans and animals. The taxonomic status of *Prototheca* has been evolving in recent decades. Five species are currently assigned to this genus: *P. zopfii*, *P. wickerhamii*, *P. blaschkeae*, *P. stagnora* and *P. ulmea* (Pore, 1985; Arnold and Ahearn, 1972; Roesler et al., 2006). Cases of human protothecosis are predominantly caused by *P. zopfii GT 2* and *P. wickerhamii* and occur as local (predominantly cutaneous) and systemic infections mainly in immune-compromised patients, e.g. patients infected with HIV or treated with glucocorticoids (Matsuda and Matsumoto, 1992; Bianchi et al., 2000; Lass-Flörl et al., 2004; Lass-Flörl and Mayr, 2007). Bovine prototheca mastitis is worldwide mainly caused by *P. zopfii GT2*, more seldom by *P. blaschkeae*. *P. blaschkeae* were further isolated from some cases of onychomycosis (Roesler et al., 2006). Canine protothecosis is caused by *P. wickerhamii* and *P. zopfii GT2*, and is characterized by similar clinical symptoms as in humans (Stenner et al., 2007).

The discrimination between harmless and pathogenic variants remains difficult and is typically performed by sequence analysis of the 18S rDNA, or by diagnostic PCR or RFLP (Roesler et al., 2006). Although this method is the most accurate available up to now, this test is not capable of differentiating *Prototheca* species in a single run. In our previous paper we demonstrated that this diagnostic gap can be overcome by using MALDI-TOF MS spectra for identification. We here present an improved analysis of these data using spectral SOM portraits.

The paper is organized as follows: In the methodical part we describe the functioning of SOM-machine learning if applied to MS peak lists. The gallery of SOM images of *Prototheca* samples is discussed in the second part. Finally we compare the classification power between single and meta spectra-based discriminant analysis.

2. Materials and methods

2.1. Sorting MS-peak pattern using self-organizing maps

The MS-spectra of different species provide characteristic peak pattern differing mainly in their peak amplitudes. Fig. 1 illustrates the basic idea of self-organizing map (SOM) clustering how to compress the information content of the respective peak lists by removing redundant pattern and to visualize the remaining meta peak profiles in terms of sample-related mosaic maps: In the first step one extracts single peak profiles from the peak lists. Each profile contains the amplitudes of one of the peaks in all samples considered. The second step analyzes all obtained single peak profiles and extracts so-called meta peak profiles which serve as proxies of all non-redundant pattern inherent in the original data. This step reduces the number of profiles considered because single peak profiles of similar shape merge into one meta profile. In the last step all extracted meta profiles are collected into sample-related meta peak lists ('meta spectra'). They are visualized using a series of mosaic images where each image shows the amplitudes of all meta peaks in one of the samples. The whole procedure thus removes redundant information from the data and represents the non-redundant part in terms of mosaics using a suited color or grey scale.

To solve the peak re-sorting task we applied SOM-machine learning to MS-peak lists (see Fig. 2 for illustration). The SOM method applies a neural network algorithm to project high dimensional data onto a two-dimensional visualization space (Bishop et al., 1998; Kohonen, 1982). Our algorithm initializes a sufficient number K of meta (peak) profiles and arranges them into a two-dimensional rectangular grid of size $K = K_x \cdot K_y$. These meta profiles represent vectors of dimensionality M given by the number of spectra included in the study. Then each peak from the peak lists of all spectra is transformed into a single peak profile as illustrated in Figs. 1 and 2. Each single peak profile is associated with the meta profile of closest similarity using Euclidian distance metrics. Then, each meta profile is adjusted so that it more closely resembles the profiles of the associated single peaks features. An iterative procedure progressively optimizes the similarity between all meta- and the associated single peak profiles where also the meta profiles of adjacent tiles in the mosaic are adjusted using a distance dependent weight. The resulting final grid consists of regions of similar meta profiles. Each meta profile represents a minicluster of several single peaks with similar profiles. The meta profiles can be understood as a sort of prototypes characterizing the multitude of non-redundant single profiles inherent in the data.

Sample-specific mosaic images are generated by color coding the amplitude of each meta profile in the respective sample. These maps can be understood as two-dimensional meta spectra where the underlying basal grid defines the peak-prototypes and the color (i.e. the z-dimension) their amplitudes in the respective sample.

Our SOM was trained using $N = 1406$ MS-amplitudes identified in $M = 324$ samples. The data matrix thus comprises M columns

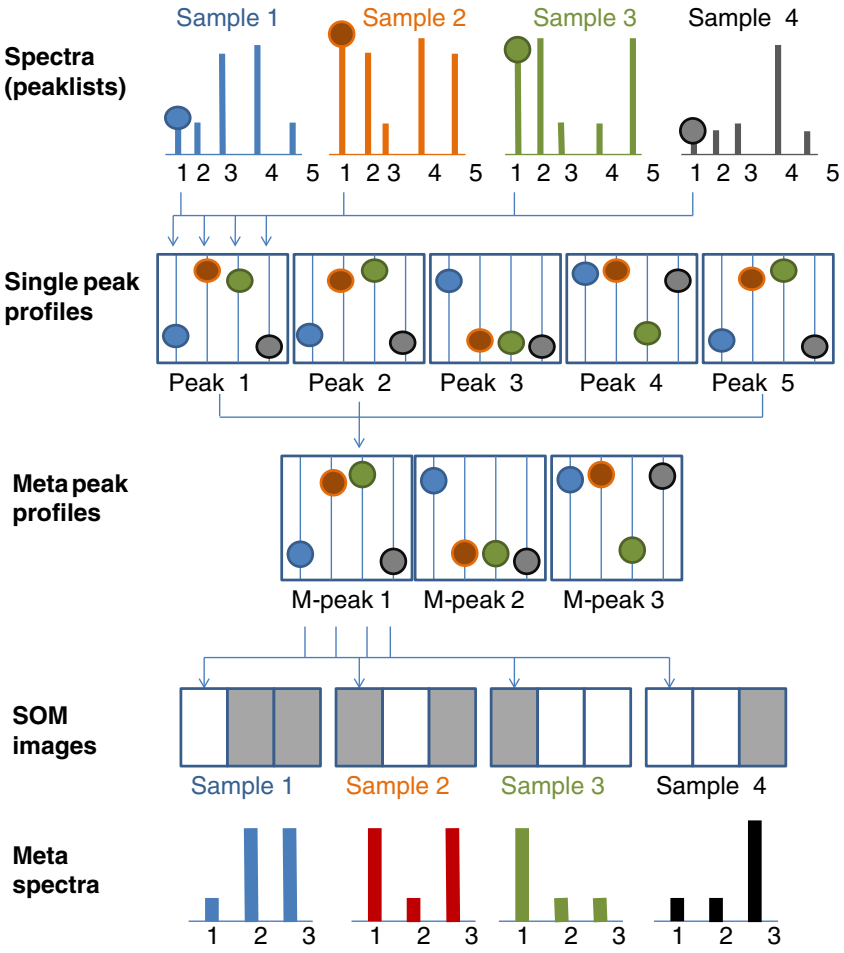

Fig. 1. Re-sorting MS-peaks by SOM machine learning: The five MS-peaks observed in the four samples reduce into only three 'meta peaks' because peaks 1, 2 and 5 contain almost redundant information in terms of their common single peak profiles. Each sample possesses a unique peak pattern; however only peak 3 is uniquely expressed in sample 1. All the other peaks show high amplitudes in at least two samples. The obtained meta spectra are visualized as grey-scaled tiles. The thin arrows illustrate the flow of information for selected peaks. For example, the values of meta peak profile no. 1 are transform into the grey scale of the first tile of the SOM images of the samples. The unique single peak 3 transforms into meta peak 2 which is uniquely high expressed in the SOM image of sample 1. Hence unique peaks are visualized by unique features in the SOM images.

representing the samples and N rows representing the intensity values extracted from the MS-spectra. These *N* amplitudes are sampled using the same set of supporting points along the *m/z*-axis in each MS-spectrum considered. This means that missing peaks are assigned to zero amplitudes in the respective samples. The vector of *M* intensity values at one particular *m/z*-value in all samples represents the peak profile and the *N* intensity values of one sample are called peaklist. The peaklists were quantile-normalized (Bolstad et al., 2003) before training the SOM. We use a number of $K = 20 \times 20 = 400$ tiles to 'pixelate' the mosaic portraits. SOM training requires about 100000 iterations which are processed in about 3 min on a standard personal computer. The number of pixels is considerably smaller than the number of MS-peaks resulting in an occupancy of 3.5 single profiles per meta profile cluster, on the average. We applied a home-made R-program which uses the CRAN package 'som'.

The so-called second level SOM analysis was performed to supplement the single and meta profile-centered views with a sample-centered one (Guo et al., 2006; Wirth et al., 2011). Here, similarity relations between the samples are visualized in a two-dimensional grid obtained by SOM training of the 'first level SOM'. The tiles of the 2nd level SOM then characterize the mosaic portraits of prototypic meta-samples. Their number in the 2nd level SOM exceeds the number of real samples to ensure sufficient resolution of the map. A considerable fraction of tiles are thus empty with no sample assigned, leaving enough space for the metasamples to unfold the complete sample-space. In addition to 2nd level SOM, we generate maximum (weight) spanning trees (MST) based on the matrix of Pearson correlation coefficients between the meta profiles for all pairwise combinations of samples using the R-package 'igraph'. Phylogenetic trees were constructed based on single and meta peak profiles using a neighbor-joining algorithm (Saitou and Nei, 1987) implemented in R-package 'ape'.

2.2. MS spectrometry of algae

The diagnosis of *Prototheca* in the case of systemic infection of human is often hampered by their visual similarity with yeast. In the field of infection of animals the actual question is how to discriminate between harmless environmental samples and the harmful ones. Since they cannot be distinguished by visual inspection, the molecular species detection offers a straight forward approach.

We used recently published MALDI-TOF data (von Bergen et al., 2009). The spectra refer to a total of 19 strains, representing 7 algae species (5 *Prototheca* spec. and 2 *Chlorella* spec.) and two genotypes of *Prototheca zopfii* were used in this study (see ref. (von Bergen et al., 2009) for details). All strains were either type, reference, or other well-characterized isolates that have been previously identified by sequence and biochemical analysis. These strains stem in parts from environmental samples but also several clinical isolates were included. Especially, most of the investigated *P. zopfii GT2*, *P. blaschkeae* and *P. wickerhamii* strains were isolated from bovine prototheca

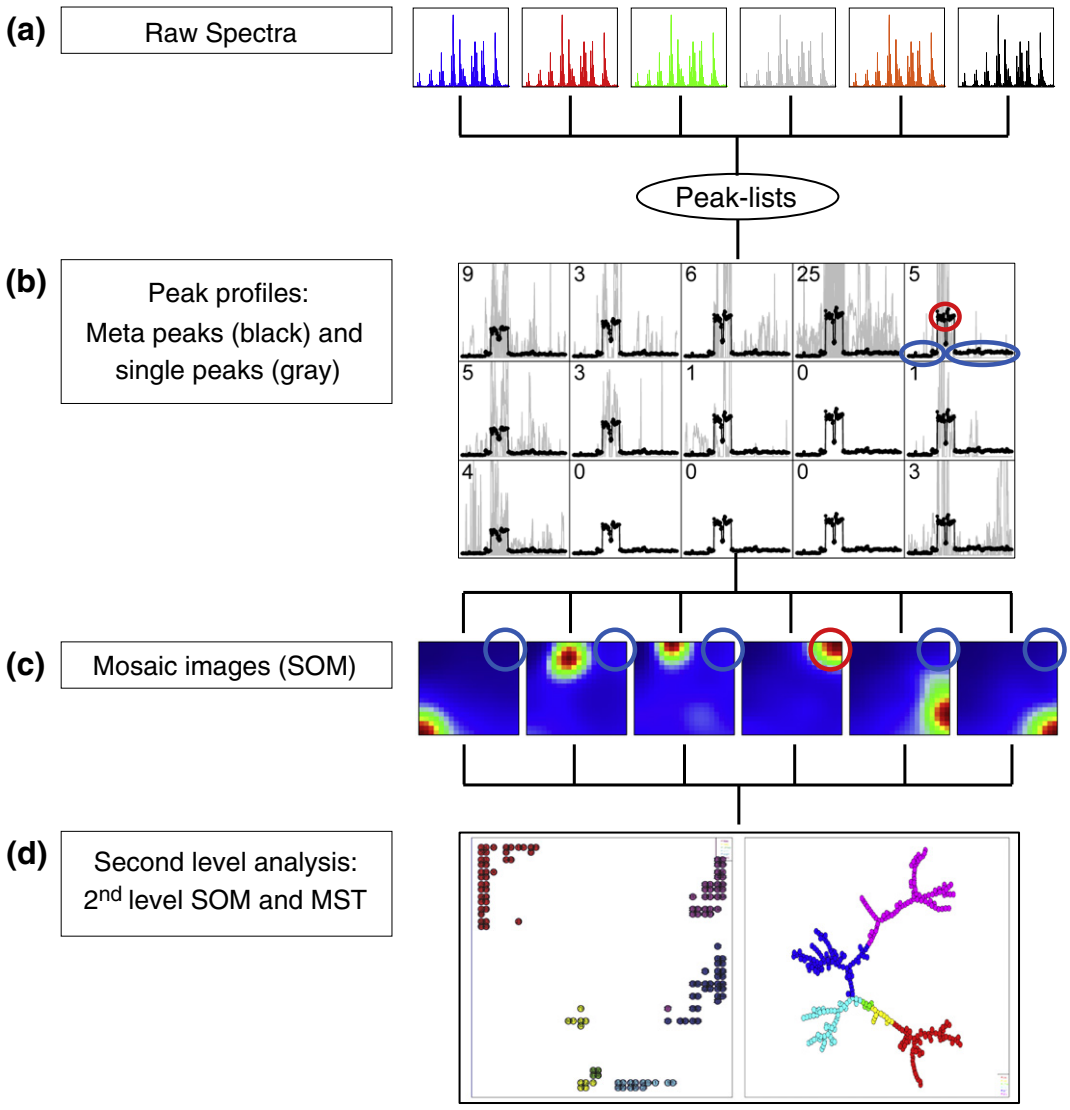

Fig. 2. MS-spectral profiling using self organizing maps (SOMs): Peaklists derived from MS-spectra referring to different samples are subjected to SOM-machine learning which distributes the meta profiles and the associated single peak profiles in a two dimensional grid where each tile refers to one meta profile and a variable number of associated single profiles (see the number in each of the tiles). The meta profiles are transformed into colored mosaics where red areas refer to large and blue areas to small amplitudes. The red and blue circles assign samples with high and low amplitudes of the respective meta peaks. Second level analysis visualizes similarity relations between the meta profiles of the samples.

mastitis, canine protothecosis, and human protothecosis. All type strains of the five *Prototheca* species were also included in the study.

Samples were prepared using a modified standard protocol, in which colonies were washed prior to applying them to a MALDI-target (MTP ground steel, Bruker Daltonik, Bremen, Germany) (Maier and Kostrzewa, 2007). Mass spectra were recorded in the range ~(4000–17000) Da with a resolution of 150–200 ppm and subsequently binned into intervals of width ~3.2 Da (i.e. ~750 ppm at 4000 Da) to account for small mass shifts between different spectra. The resulting penalty in spectral resolution is counterbalanced by a gain in stability of the selected features with impact for the applied distance metrics used to train the SOM (vide supra).

Peaks were detected from the raw mass spectra after baseline subtraction using the centroid algorithm with S/N > 6 implemented in the FlexAnalysis 2.4 program (Bruker Daltonics, Bremen, Germany) where the highest 100 peaks were labeled to normalize the spectra. The generated peak lists were exported and processed using the MS-Screener program (Version 1.0.1) (Schmidt et al., 2003). Spectral alignment of all sample spectra extracts discrete supporting points along the *m/z*-axis which meet the condition of non-zero signal amplitude at minimum one sample spectrum of the series. The final peaklist contains 1406 intensity amplitudes sampled at common *m/z*-values in all samples in the range (4135–16954) Da. Note that these intensities represent assigned *m/z* values rather than real peaks the number of which is effectively smaller. We will use the term 'peaklist' as short name of the data vector extracted from the MS-spectrum of each sample despite the fact that some of the 'real' peaks remain unresolved and extend over two-to-three subsequent *m/z*-bins. Our SOM method aims at sorting the data matrix obtained from the vectors of all samples into unique (showing non-zero amplitude only in one class of samples) as well as ubiquitous (showing non-zero amplitudes in more than one class) high and medium amplitude peaks, but also into low amplitude noisy features as shown below. The consequences of the partially imperfect spectral resolution of single peaks will be discussed below.

2.3. Random Forests classification

Predictive value of single- and meta-peak profiles is a useful measure of the proposed MALDI-TOF method to obtain stable, high accuracy

classifiers for the analyzed data sets. Random Forests (RF) analysis, a multiclass supervised classification method (Breiman, 2001) was performed to estimate prediction possibilities of single- and meta-peak profiles using the R/Bioconductor Random Forests package (Liaw and Wiener, 2002). The method generates one million decision trees that assure a stable list of differentiating features under consideration of all features studied (see, e.g., (Díaz-Uriarte and Alvarez de Andrés, 2006)). Obtained classification accuracies were presented as 'Out-Of-Bag' errors generated by the RF quality evaluation. Features differentially changed according to RF were visualized in principal component biplots together with the respective samples (Gabriel, 1971).

3. Results and discussion

3.1. MS-SOM atlas of Prototheca species

The SOM was trained using MS-peaklists of 324 samples referring to five *Prototheca* species where that of *Prototheca zopfii* splits into two genotypes. Each peaklist contains the amplitudes of 1406 peaks. Our method transformed the peak list of each sample into one mosaic image serving as molecular portrait of the underlying MS-pattern. Fig. 3 shows these SOM-images using a 20×20 mosaic grid. The color gradient of the map was chosen to visualize high and low peak amplitudes of the meta profiles in the particular samples: Maroon codes the highest intensity level; red, yellow and green indicate intermediate levels and blue corresponds to the lowest level of peak amplitude. Each mosaic exhibits characteristic spatial and color patterns serving as MS-fingerprint of the *Prototheca* samples studied.

Comparison of the individual SOM portraits within each species reveals very similar pattern for *P. blaschkeae*, *P. ulmea* and the two genotypes of *P. zopfii*. The SOM portraits of *P. stagnora* and *P. wickerhamii* show isolate-specific differences. Different species show consistent differences between their mosaic patterns. Hence, comparison of the SOM-textures allows the straightforward classification of the samples according to their taxonomic membership.

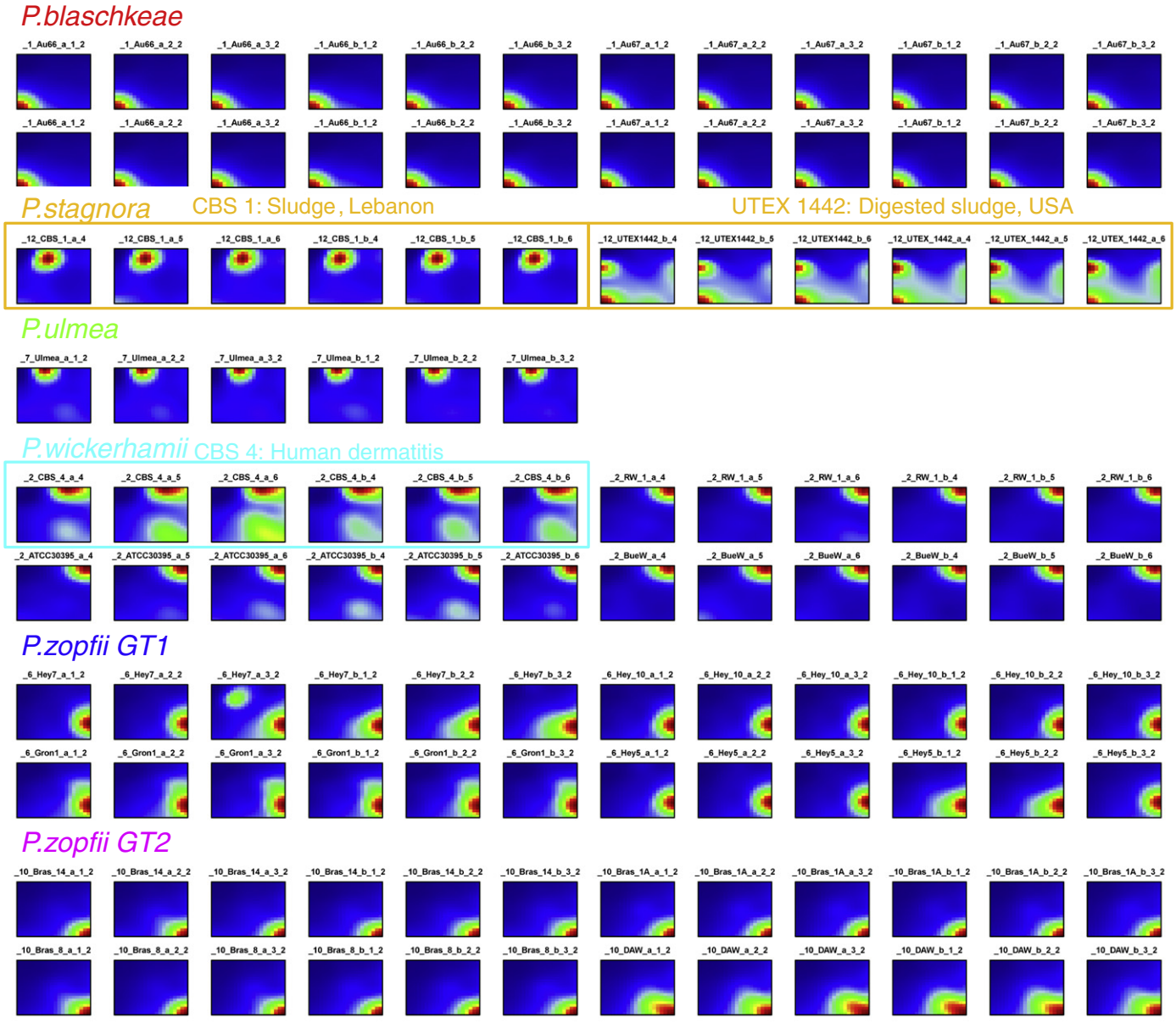

Fig. 3. SOM profiles of 114 selected *Prototheca* samples. The images are sorted according to their taxonomic categories. The color of the heading of each category is used also in the other figures throughout the paper for assignment. Selected batches of samples of different origin are marked by colored frames.

Most of the individual SOM portraits show only one high-intensity spot. Its position however varies in a species- and, partly, isolate-specific fashion. This property means that each species is characterized by a set of virtually unique peaks showing high amplitudes only for this particular species but small amplitudes for all other ones.

The observed spot patterns of the images potentially depend on the particular realization of SOM training and on the particular set of samples included in the analysis. Technical details of the method such as the number of tiles per mosaic image, the initialization method (e.g. linear or random), the consideration of adjacent tiles (e.g., Gaussian or 'bubble' neighborhood) and also the lattice-type (e.g., rectangular or hexagonal grids) can affect the number of spots and their position in the map. In a recent study we demonstrated however that the number of spots and their mutual arrangement is virtually independent of the chosen lattice type, initialization, neighborhood and the pixelation of the image if the number of tiles exceeds the number of relevant clusters of strongly-correlated feature profiles roughly by two-orders of magnitude (Wirth et al., 2011). Each *Prototheca* image shows typically not more than two of such spot-clusters. Hence, the chosen number of 400 pixels fairly meets the condition of convergent spot patterns. This result is further confirmed by the fact that the zoom-in step applied to the *P. zopfii* samples does not affect the number and relative arrangement of the detected spots (see below). The observed spot patterns consequently reflect intrinsic properties of the associated mass spectra in the ensemble of samples considered presuming that training of the map reached equilibrium. Each SOM portrait then can serve as a visual identity of the underlying mass spectrum in the context of the set of samples which are trained together. The spot patterns can change if one alters the ensemble of sample-spectra included in the study. We previously used this option in a so-called zoom-in step which considers only a subensemble of the initial sample series in order to 'amplify' the landscape of feature-values for these selected cases (Wirth et al., 2011). Below we apply the zoom-in view to the two *P. zopfii* genotypes to illustrate the potency of this option for MALDI-typing. There is however basically no need to further increase the resolution of the SOM applied to the *Prototheca* species because the relevant spots in the obtained images are well resolved allowing to identify the different species with high specificity (see below).

Each tile of the mosaic images refers to one of 400 meta peak profiles characterizing the peak-landscape of the samples. Each meta peak serves as representative for a minicluster of correlated single peak profiles the number of which varies between the meta peaks. The meta profiling map illustrates the variation of the meta profiles throughout the map using a coarse grained resolution of the mosaic (Fig. 4). The position of the profile-graphs in the grid agrees with their position in the higher-resolved SOM mosaics to enable direct comparison. The profiles refer to the samples of each species in consecutive order as indicated by the color bars above the graphs shown in each of the 10×10 tiles.

The profiles of the meta peaks reveal high amplitudes for distinct algae species in the distinct regions of the map. For example, the meta profiles in the bottom left corner show strong intensities for *P. blaschkeae* (red bar) and weak intensities for essentially all other species as mentioned above. Contrarily, the top right corner is occupied by meta profiles possessing high amplitudes for *P. wickerhamii* (cyan bar) and bottom right corner by meta profiles possessing high amplitudes for *P. zopfii* (blue and magenta bars). Please note also that the shape of the meta profiles changes gradually along the borders of the map. Similar profiles are located closely together whereas their shape diverges with increasing distance in the map. The strong similarity of the profiles in adjacent tiles gives rise to the smooth color texture of the high resolution images expressing pronounced spots in distinct areas of the map (Fig. 3).

The meta peak profiling map nicely illustrates the systematic character of alterations of the meta profiles within the SOM. It also shows that the meta profiles largely differ in the overall variance of their profiles: In the center and in direction towards the top left corner one finds meta profiles of low variance whereas the profiles located alongside the top and low borders more strongly vary. Please note, that the number of single peak profiles per meta profile strongly alters as indicated by the numbers given in each tile of the mosaic in Fig. 4.

The profiling map in Fig. 4 also reveals that the top left corner is occupied by virtually invariant and low-amplitude meta peak profiles carrying virtually no species-specific information. In consequence, this region lacks characteristic high-amplitude spots and therefore it is consistently colored in blue in all sample portraits. Please note that more than 500 out of the 1406 single peak profiles considered cluster together in this region of the map (the numbers given in the tiles in Fig. 4 indicate the number of single peaks in the respective meta peak cluster). This result illustrates that more than one third of the features of the peak list used are virtually non-informative with respect to different forms of *Prototheca*. The SOM machine learning method automatically and effectively separates these features from the more informative ones solely based on the criterion of minimum Euclidian distance to cluster similar profile.

Note also that the regions of the high-amplitude spots in the remaining three corners of the map typically show local population maxima of 25–50 single peak profiles per spot which are typically surrounded by regions of sparsely populated tiles. This result indicates that the selected peaks exhibit partly binary 'present–absent' characteristics with respect to their appearance in different species. Their amplitudes are either relatively intense or relatively weak. Intermediate amplitude values are relatively rare giving rise to sparsely populated regions around the high amplitude spots. Two of the meta profiles taken from the top and bottom right corners of the profiling map are shown with enlarged resolution in the right part of Fig. 4. These graphs also depict the respective single peak profiles (grey curves) expressing high amplitudes in the spectra of *P. wickerhamii* (see the cyan bar above the curves) or in the spectra of the two *P. zopfii* genotypes (blue and magenta bars), respectively. We mark these peaks in representative mass spectra of all species studied (Fig. 5). This representation clearly shows that the selected peaks indeed form a characteristic set which protrudes with larger amplitudes uniquely only in the spectra of one of the species. The closely related *P. zopfii* genotypes are characterized by specific peaks, labeled in blue and magenta in Fig. 5, respectively. Hence, SOM machine learning clusters the MS-peaks into species-specific spots.

The SOM images also reveal isolate effects of algae samples of the same species which are marked in Fig. 3 by colored frames. For example, the samples of *P. stagnora* split into two different isolates with clear differences of their spot patterns. Fig. 6a shows representative spectra of the two *P. stagnora* isolates *UTEX1442* and *CBS1* and of *P. blaschkeae*. MS peaks of high amplitude cluster into three specific spots which allow clear discrimination between the different sample types. The spot in the left bottom corner is commonly found in the portraits of *P. blaschkeae* and *P. stagnora UTEX1442* which however show a second isolate-specific spot referring to a unique set of MS-peaks differing from that observed in the second isolate *P. stagnora CBS1*. Fig. 6b depicts the characteristic SOM spots and the associated MS-peaks for the two isolates of *P. wickerhamii* (*CBS4* and *ATCC30395)*. The spots are in adjacent positions and partly overlap. This pattern refers to a superposition of two isolate-specific sets of MS-peaks marked by grey and cyan dots, respectively.

Note also that high signal amplitudes at subsequent m/z-values referring to not resolved peaks typically cluster together in the same spot (see, e.g. the pairs of dots marking the same peak in part of the spectra shown in Figs. 5 and 6). Hence, the imperfect resolution of single peaks in the original peaklists used to train the SOM is not crucial for the typing task. Intensity values referring to the same unresolved peak although treated as independent features in SOM

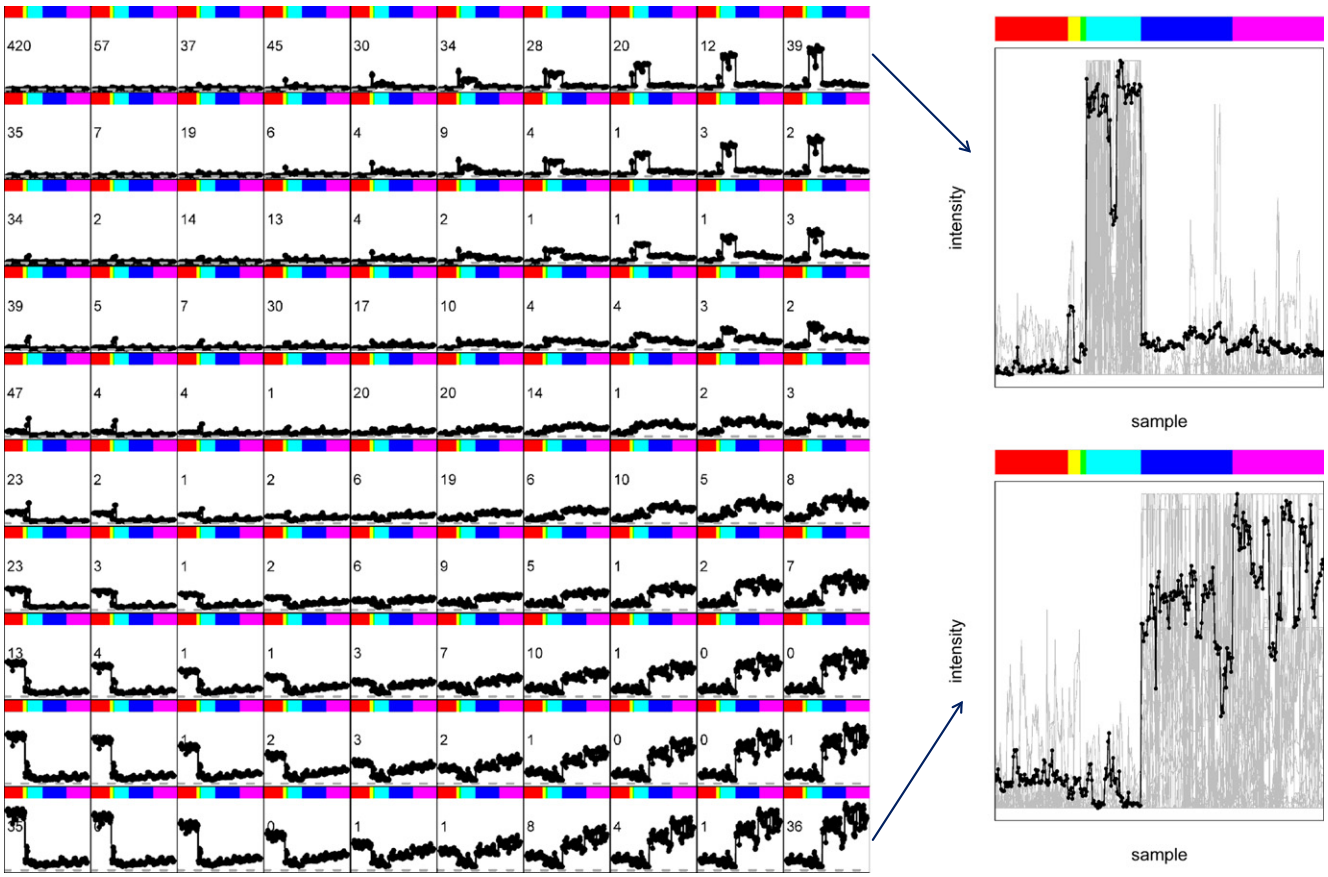

Fig. 4. Meta peak profiling map of *Prototheca* species: The graphs in the left part show the meta profiles using a more coarse granulation of the SOM mosaic images of 10 × 10 tiles instead of the 20 × 20 pixels in the original portraits shown in Fig. 3. The color bar on top of each of the tiles assigns the *Prototheca* species (compare with the headings in Fig. 3). The numbers in each tile provide the population of the respective minicluster with single peaks. Two graphs are amplified in the right part: The meta profiles are shown by thick curves whereas the grey ones show the profiles of the associated single peaks. One sees that peaks showing large amplitudes in *P. zopfii GT1* (blue bar) are represented by the meta profile in the top right corner of the map whereas high-intensity peaks of *P. zopfii GT2* (magenta bar) cluster together in the down right corner.

training are usually assigned to the same meta peak profile owing to their correlated amplitude profiles.

In summary, SOM serves as an effective 'sorting machine' which distributes the individual amplitude profiles over a series of meta peak profiles representing intrinsic modes in the data space provided by the experimental data. The spot patterns of the obtained SOM images 'portray' each of the *Prototheca* species and part of the isolates which allows their identification and assignment by visual inspection of the individual portraits.

3.2. Spot analysis

The texture of the SOM images clusters and visualizes high amplitudes of the meta peak profiles. The obtained spot like clusters can serve as specific markers of different algae species and isolates. For an overview about all observed spots we create the so-called high-amplitude summary map shown in Fig. 7a. It represents an integral 'master' map collecting all high amplitude spots observed in the individual SOM portraits (Fig. 3). In total, eight spots (labeled A….H) are identified by selecting 2% of the meta peaks with largest amplitude in each sample. Most of the spots can be uniquely associated with just one species (see the right part of Fig. 7a). The typical spots of *P. wickerhamii* appear in two variants located at adjacent positions in the map (spots E and F). *P. stagnora* shows a slightly more complex pattern of the spots A–C as described above.

The so-called spot-amplitude heatmap is shown in part b of Fig. 7. It visualizes the profiles of the mean amplitude averaged over all meta peaks of each of the spots in all samples studied. Importantly, it scales the amplitudes between the samples in absolute units whereas each of the SOM portraits in Fig. 3 are scaled individually between the maximum and minimum peak values in each sample. The spot-amplitude heatmap consequently allows comparison of the spot amplitudes between the samples and also between the different spots. For example, the amplitude of spot A markedly exceeds that of spot E. Most of the spots possess high amplitudes in only one of the species. Especially spot A shows consistently high amplitudes in *P. blaschkeae* whereas spots B–D are linked to the different isolates of *P. stagnora* and to *P. ulmea* as described above. Spots G and H collect the high intensity peaks of both *P. zopfii* genotypes however with alternating maximum values. These peaks appear also in *P. blaschkeae* however with smaller amplitudes. Hence, part of the spots detected is more ubiquitous appearing in different species. Nevertheless, also these spots can serve as specific markers due to their species- and/or isolate-dependent intensity levels.

The color bars on top of the heatmap in Fig. 7b assign different strains and isolates according to reference (von Bergen et al., 2009) (see Table 1 therein). Partly, these characteristics correlate with modulations of the spot amplitudes. Unfortunately only about 5% of the samples are supplemented by strain information disabling a more detailed and systematic study of these effects.

In summary, SOM analysis enables identification of species specific peaks in the mass spectra and thus further accurate identification of discriminating markers. We suggest tryptic digestion of the intact protein extracts and subsequent shotgun mass mapping (Jehmlich et al., 2009) to match proteins to characteristic spectra peaks.

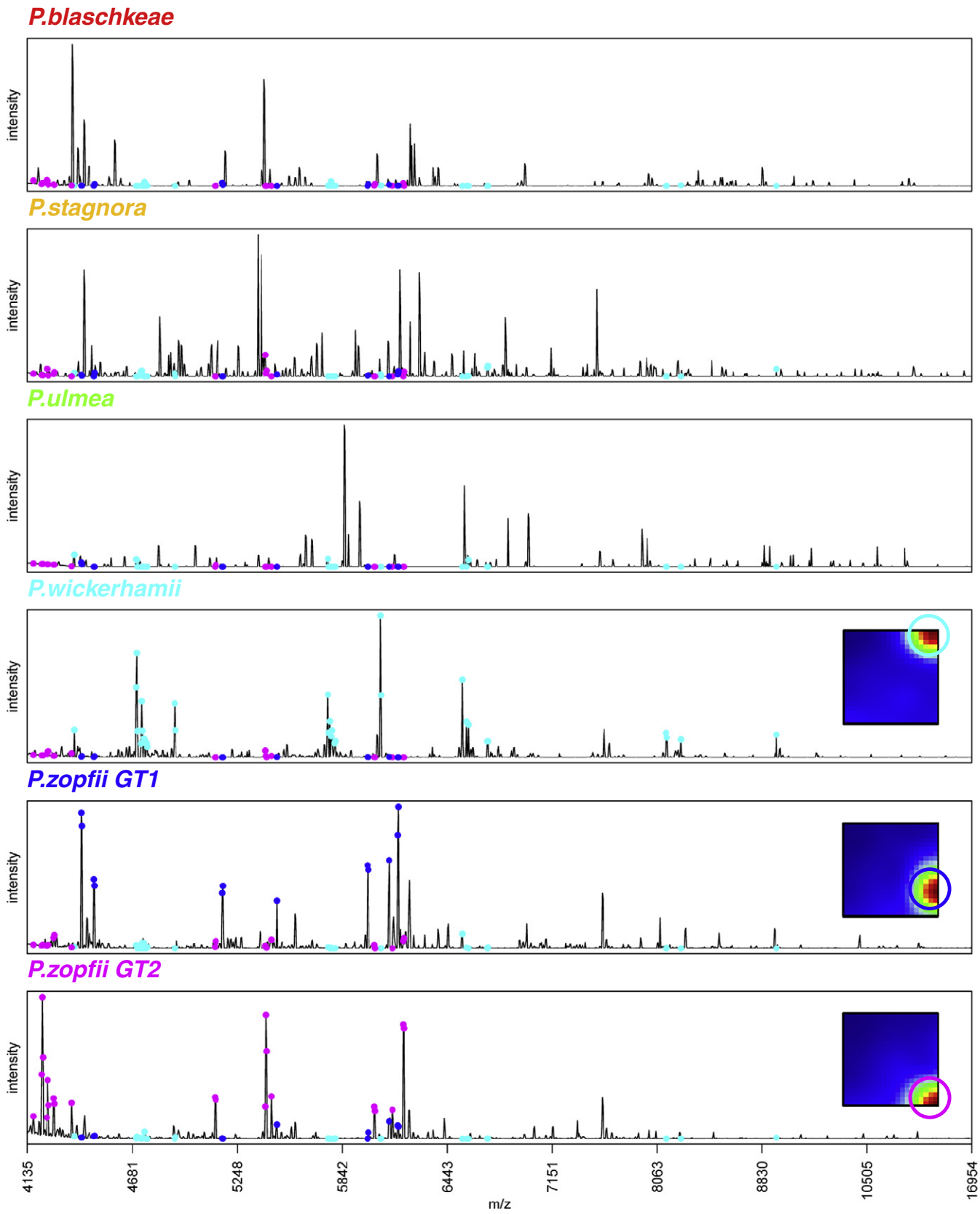

Fig. 5. Typical MALDI-TOF spectra of different Algae species. Peaks indicated by cyan, blue or magenta dots are taken from the respective red 'high-amplitude' spots in SOM images of *P. wickerhamii*, *P. zopfii GT1* or *P. zopfii GT2*, respectively. The respective spots are indicated by colored circles in the images shown in the right part of the figure. Note that some of the peaks are marked by two dots of the same color due their imperfect resolution in the peaklists used to train the SOM.

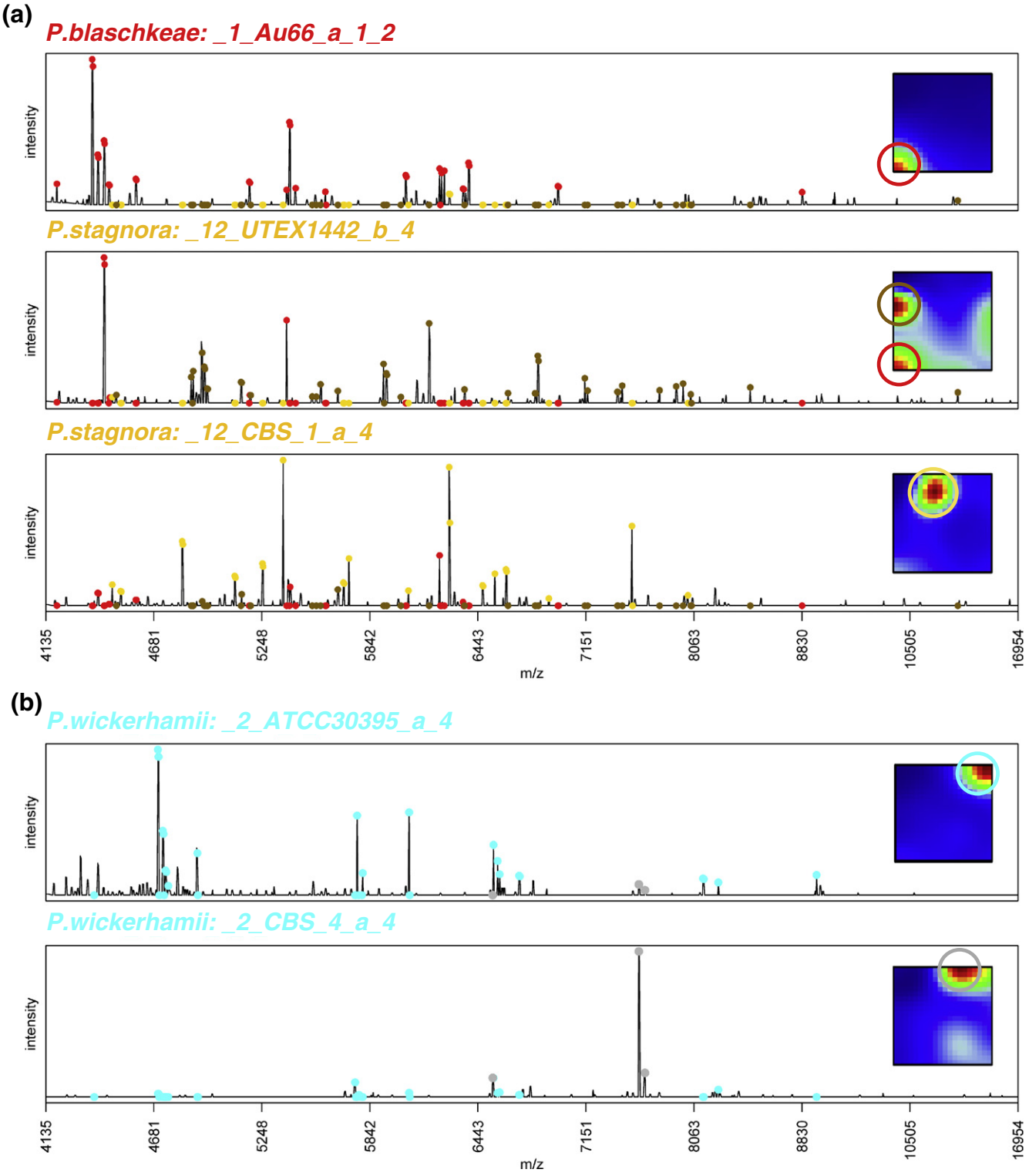

Fig. 6. Selected MALDI-TOF spectra of different isolates indicated in Fig. 3: *P. stagnora UTEX1442* and *CBS1* and *P. blaschkeae AU66* (panel a) and *P. wickerhamii CBS4* and *ATCC30395* (panel b). The respective MS-peaks are marked by colored dots taken from the high-amplitude spots in the SOM images as indicated in the right part of the figure.

3.3. Similarity analysis

Guo et al. proposed a second level SOM analysis step (Guo et al., 2006). It maps all samples together into one two-dimensional mosaic pattern to visualize the degree of similarity between the 'first-level' SOM portraits. The second level SOM algorithm uses the meta profiles of each sample as input. After training, each tile of the mosaic is characterized by the profile of one 'metasample' which serves as the condensation nucleus of the associated minicluster of real samples possessing similar SOM pattern. The mutual distances between the samples in the map are related to the degree of similarity of their SOM expression pattern. Our second level SOM uses a resolution of 20×20 nodes, which exceeds the number of samples: On average only 0.8 samples are assigned to each node. In consequence a

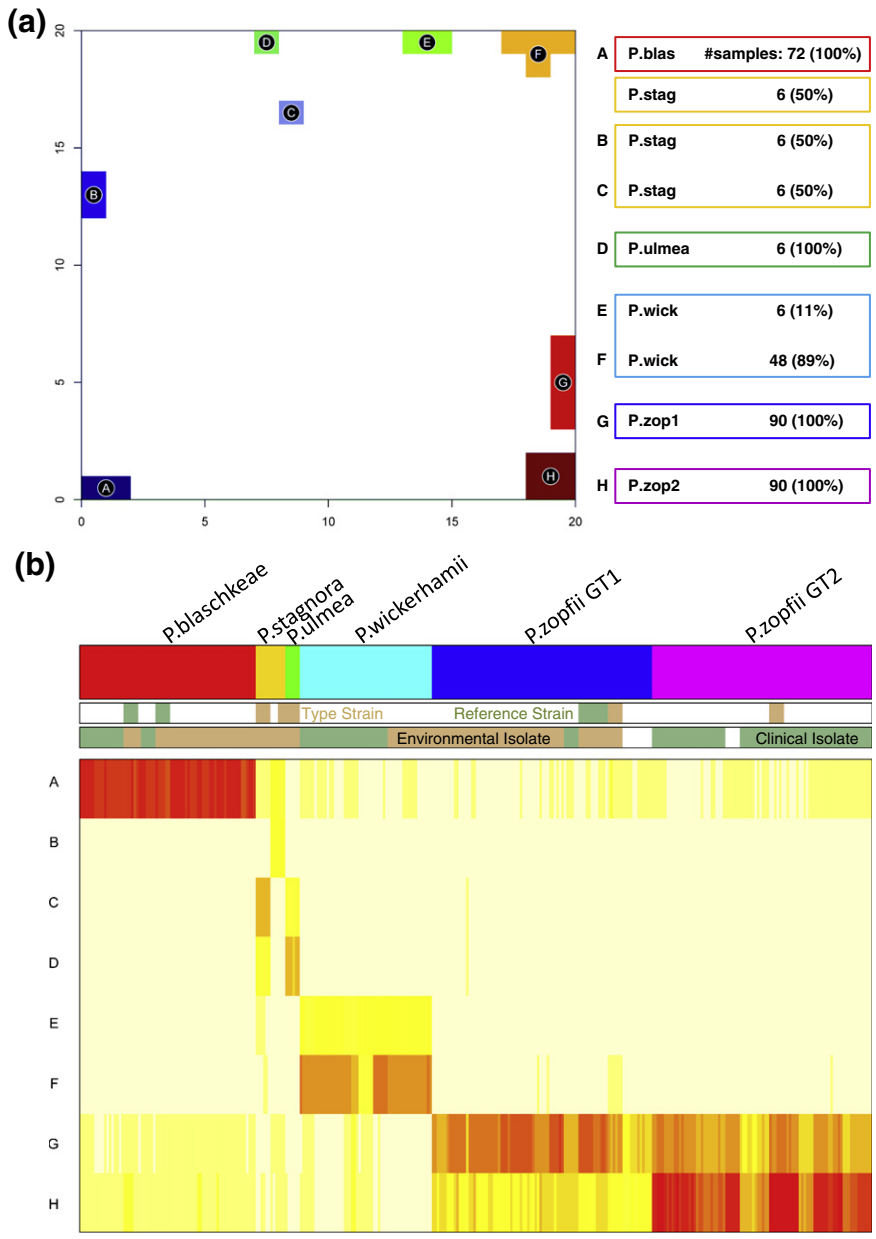

Fig. 7. High-amplitude summary map (part a) and spot amplitude heatmap (part b) provide an overview about all high-amplitude spots and their mean intensities observed in the series of samples studied. The spots are labeled using capital letters A…H. The spot amplitudes are color-coded in the heatmap with yellow-brown-yellow-white from high to low values. The percentages of samples of each species showing the respective spot are given in the right part of panel a.

considerable number of tiles remain empty. Our 1st level SOM-portraits use a denser population of tiles in the mosaic with, on the average, 3.5 single profiles per meta profile leaving only a few tiles unoccupied.

Fig. 8a shows the second level SOM of all 324 *Prototheca* samples studied: The position of each sample is represented by a small circle colored according to the respective species. Samples of each species mostly form compact and well separated clusters allowing the clear assignment of the different algae. The two isolates of *P. stagnora* (yellow circles) split into two separate clusters where the cluster of *P. stagnora* CBS1 is more closely located to the clusters of *P. ulmea* (green) and *P. zopfii* 1 (blue) than to that of the second isolate of *P. stagnora*. The observed relations thus reflect the spot patterns discussed above.

The maximum spanning tree shown in Fig. 8b provides an alternative option to visualize the similarity relations between the samples. It directly connects the images of highest mutual correlations between their metagenes. The algae clearly aggregate into species-specific branches except *P. stagnora* which splits into the two isolate-specific groups.

3.4. SOM-analysis improves MALDI-TOF typing

We previously built a dendrogram of *Prototheca* based on MALDI spectra which was in fairly good agreement with a dendrogram based on sequence information from 18S DNA (von Bergen et al., 2009). We recalculated the former dendrogram using a neighbor-joining algorithm (Saitou and Nei, 1987) (see Fig. 9a). For the sake of clarity we averaged the single-peaks over replicated measurements of each *Prototheca* sample and considered the obtained mean spectra in tree construction. The leaves of the tree refer consequently to the averaged samples and the lengths of the branches connecting

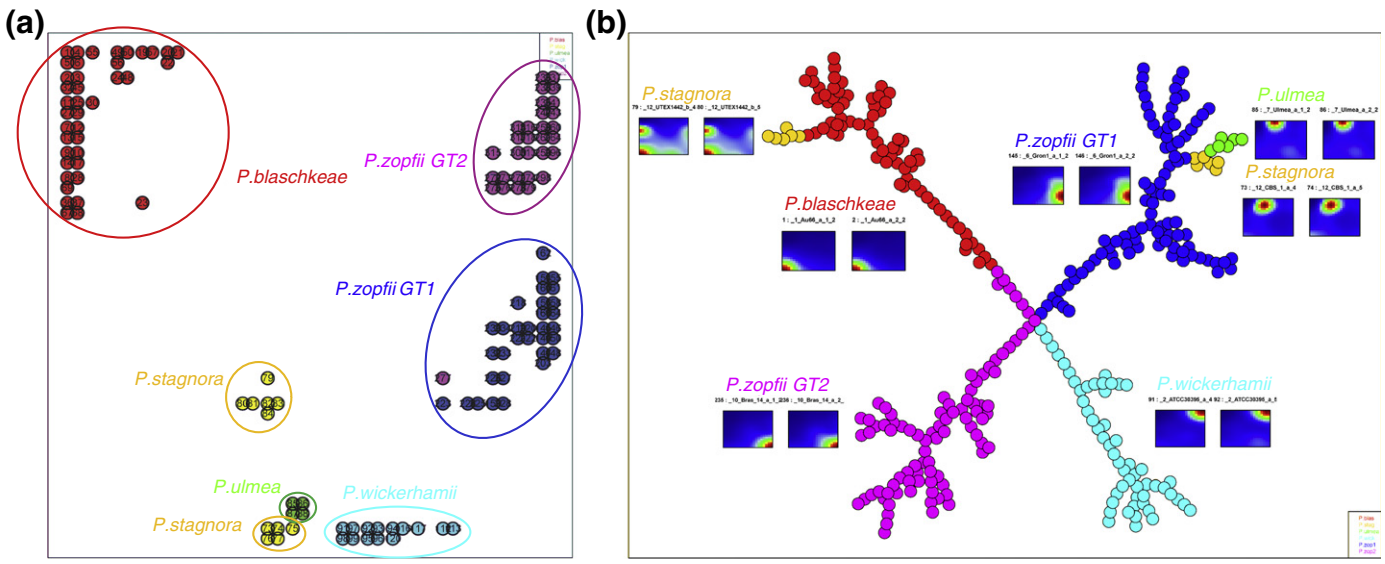

Fig. 8. Second level SOM (panel a) and maximum spanning tree (MST) of the 324 *Prototheca* samples studied: Each species is color coded by the circles according to its taxonomic category. The small mosaics show the relevant 1st level SOM pattern in panel b.

pairs of samples are directly related to the distances between them. For comparison we calculate a dendrogram using the profiles of the meta-peaks instead of those of the single peaks (Fig. 9b). Both dendrograms cluster the different *Prototheca* species into different branches of the tree in a similar fashion as discussed in the previous subsection. The single peak-based dendrogram in Fig. 9a forms a more compact polytomical, 'star-like' structure than the meta peak-based dendrogram in Fig. 9b which might be indicative for not fully resolved dichotomies in the phylogenetic tree. Detailed inspection reveals that the increased compactness of the former dendrogram indeed results from the smaller distances of the branches connecting different species. In other words, the ratio between the intra-species and inter-species distances is clearly smaller for the meta peak-based dendrogram. It consequently better separates the different groups of samples than the single peak-based dendrogram. This difference can be attributed to the improved signal-to-noise ratio of the meta peaks which are representatives of a large number of associated single peaks as illustrated in the right part of Fig. 4. Note in this context that the meta profiles are virtually mean profiles averaged over the single profiles in the respective microcluster of MS-peaks. This averaging step reduces the noise of the features and, in consequence, clearly improves the specificity of the dendrograms to distinguish between different algae species. The question whether also non-noisy and thus informative features of the single spectra get lost in this data compression step will be discussed below.

Note that the improved resolution of the meta peak-dendrograms reveals subtle substructures not clearly evident in the single peak-dendrogram: For example, the *P. zopfii GT1* sample SAG23610 is characterized by slight, but systematic differences between its SOM image and those of the other images of *P. zopfii GT1* (blue), a difference which is not evident in the tree based on single-peaks. On the other hand, sample POT2 protrudes as an outlier among the *P. zopfii GT2*

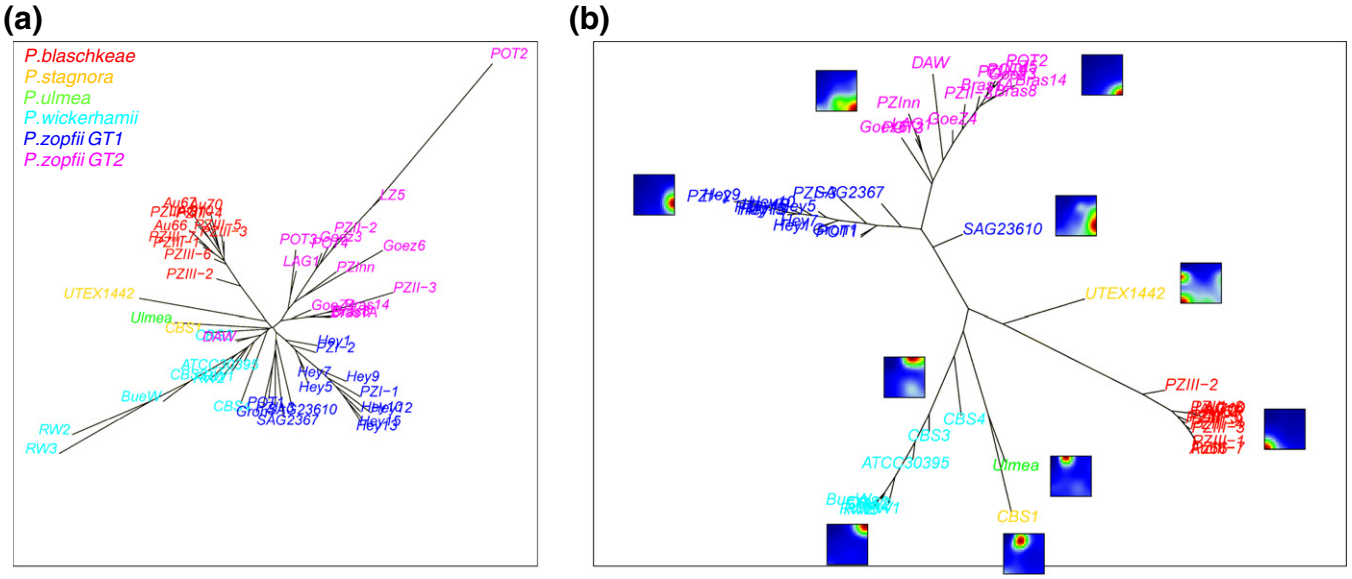

Fig. 9. The phylogenetic tree based on single spectra (panel a) features less discrimination power than the meta spectra based tree (panel b). SOM profiles are shown for selected branches in panel b.

samples (magenta) in the single peak dendrogram primarily due to an extraordinarily strong intensity of the MS-peak at ~4235.5 ± 6 Da (data not shown). It becomes mostly averaged out in the meta peak profiles which, in consequence, leads to the clearly better integration of POT2 in the cluster of the remaining *P. zopfii GT2* samples in the dendrogram of meta profiles.

3.5. Classification and biplot analysis

The dendrograms shown in the previous subsection suggest high classification accuracy even for the actual six-class analysis problem. We applied the Random Forest method which successfully separates the six algae species: The 'Out-Of-Bag' classification error obtained for both single and meta peak data was equal to 0%. Iterative classifier training shows that single peak data and metadata classifiers required at least 35 peaks of highest variance to reach perfect 100% classifier accuracy.

Fig. 10 shows principal component biplots of the data matrices composed of the algae samples and the lists of the 35 single or meta peaks of highest variance which ensure perfect classification. The chosen biplot-representation allows the visual appraisal of the structure of these data matrices. Particularly, it visualizes inter-sample distances and sample clusters as well as displays variances and correlations between the peak amplitudes (Gabriel, 1971). On the average, longer arrows in the biplot of the meta peaks reflect the larger variability of their amplitudes between the classes compared with that of the single peaks. Note also that the arrows referring to the meta peaks which classify *P. zopfii* spread over a wider range of mutual orientations. This result reflects a wider range of correlation coefficients between their meta peak profiles and thus a more diverse pattern of linear combinations of these meta features. In consequence, the different species aggregate into better resolved clusters than in the single-peak based biplot. Note that a given number of meta peaks typically covers a larger diversity of MS spectra than the same number of single peaks due to the benefit in representativeness of meta features (Wirth et al., 2011). We will address this advantage of the meta peaks in the next subsection.

3.6. Filtering and zoom-in

Selecting and removing peaks that carry essentially no or low information is common practice to improve class discovery in spectral analysis. Filtering in general aims at improving data by removing either noisy, biased and/or non-informative (usually weak amplitude) peaks. Random noise tends to disrupt similarity relations between samples whereas, contrarily, systematic noise, e.g. due to batch effects, can cause artificial clustering if the bias affects groups of samples in a coordinated fashion. On the other hand, extreme filtering is dangerous because it may eliminate valuable information, for example, about peaks of relatively low and thus noisy amplitudes but with important impact for sample classification. Hence, filtering is an optimization task with the requirement of removing virtually irrelevant data while preserving all information which is important in the context of the particular issue studied.

We recently showed in detail for gene expression data that lists of meta features are less sensitive to data filtering than lists of single features due to the better representativeness of the former ones. Moreover, meta features possess the better signal-to-noise characteristics as a comparable collection of single genes. Therefore meta features provide a natural choice to detect context-dependent feature patterns in complex data sets which outperform at least naïve filtering methods (e.g. by selecting a certain number of most intense or most variable peaks) as has been shown previously (Wirth et al., 2011).

One essential feature of the SOM approach is the reduction of dimensionality of the full data set from more than thousand of single peak profiles to 400 meta peak profiles. In a second unsupervised reduction step, the dimensionality is further reduced to a few high-amplitude spots which represent clusters of similar meta peak profiles showing high amplitude in, at minimum, one *Prototheca* species. Importantly, this dimension reduction does not entail the loss of primary information in contrast to simple filtering which irretrievably removes part of the data. Instead, the reduction of dimension in SOM training is achieved by the weighting of primary information in the aggregation step which essentially reduces redundant information in feature space. Importantly, the whole set of single spectra remains 'hidden' behind the meta spectra. This primary information

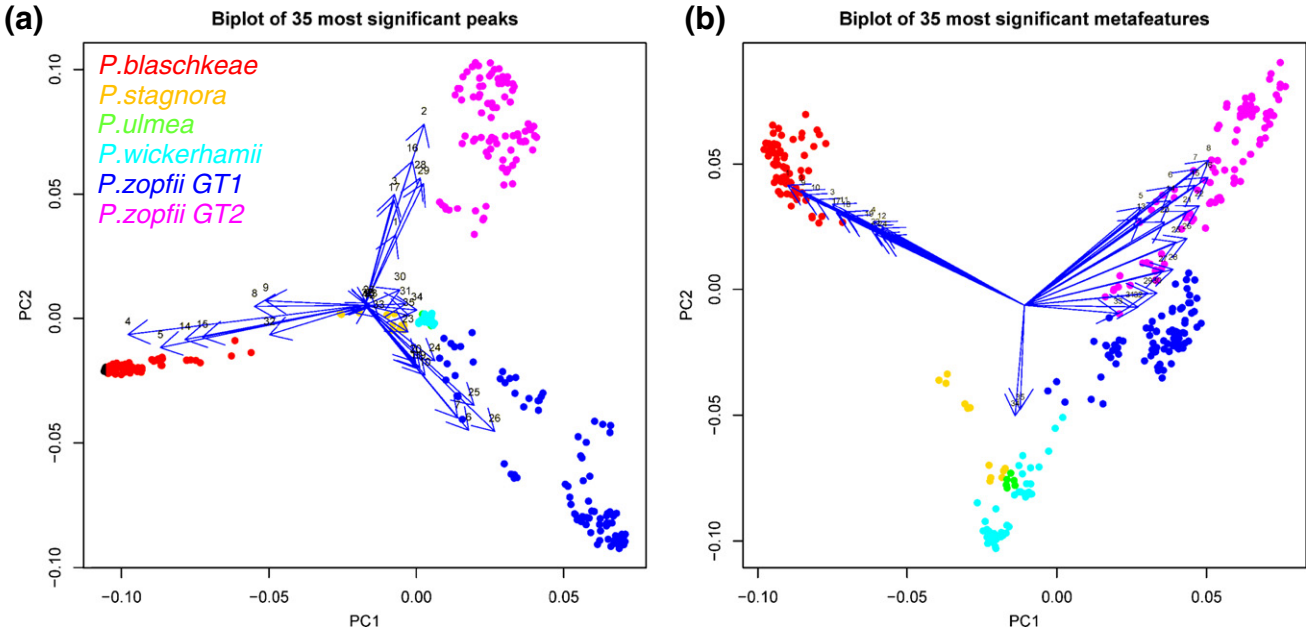

Fig. 10. Principal component biplots of the 35 peaks of highest amplitude variance ensuring 100% classification accuracy based on single (panel a) and meta peaks (panel b), respectively. The arrows numbered 1…35 refer to the peaks whereas the dots show the algae samples. The lengths of the arrows scale with the changes of the peak amplitude between the samples in terms of the standard deviation. The cosine of the angle between pairs of arrows provides the correlation coefficient of the amplitude profiles of the respective peaks.

together with the respective peak annotations can be extracted in later steps of analysis to interpret the observed SOM textures using detailed information about the single peaks and their assignment to different molecular species.

Another possible option to extract additional information is the so-called 'zoom-in' step which trains a new SOM using a reduced set of samples. SOM training thus adapts the meta peak profiles to a smaller bandwidth of feature values observed in the subensemble selected. In consequence the obtained SOM images virtually 'amplify' the features of high variability in the selected samples enabling diversification of the obtained spot pattern (see (Wirth et al., 2011)).

We applied zoom in analysis to the two genotypes of *P. zopfii*. It has been previously proposed to consider them as two subspecies based on the sequence analysis of the 18S rRNA gene (Roesler et al., 2006). Interestingly, the 1st level SOM portraits of both subtypes GT1 and GT2 still show essentially only two closely located high-amplitude spots which have already been observed in the SOM portraits before zoom-in despite the fact that the spots appear at different absolute positions in the map due to initialization effects (see Fig. 11a). This relative invariance of the spot pattern reflects a precipitous feature landscape with pronounced and stable peak characteristics of each genotype. In contrast, zoom in maps of expression data almost completely change their spot patterns after zoom in due to a much smoother feature landscape (Wirth et al., 2011).

Similarity analysis using second level SOM shows that the samples points cover nearly the full area of the map after zoom in (see Fig. 11b). Selected batches of samples, especially of the GT1 subtype, aggregate into disjunct clusters with improved resolution compared with the original map shown in Fig. 8a. After zoom-in, also the phylogenetic tree reveals more pronounced branching of the different sample types (see Fig. 11c). Zoom-in thus enhances the resolution of the maps allowing extraction of subtle differences between the different samples which are not clearly resolved before zoom-in. This enhanced resolution is mainly related to small effects which are captured by the finer granulation of the feature space.

Recall in this context that SOM training minimizes the *Euclidian* distance between the meta- and single peak profiles. Consequently, the discrete m/z values have been treated as independent signals even if they are part of the same peak and thus mutually dependent. The binning width of the spectra thus determines the minimum

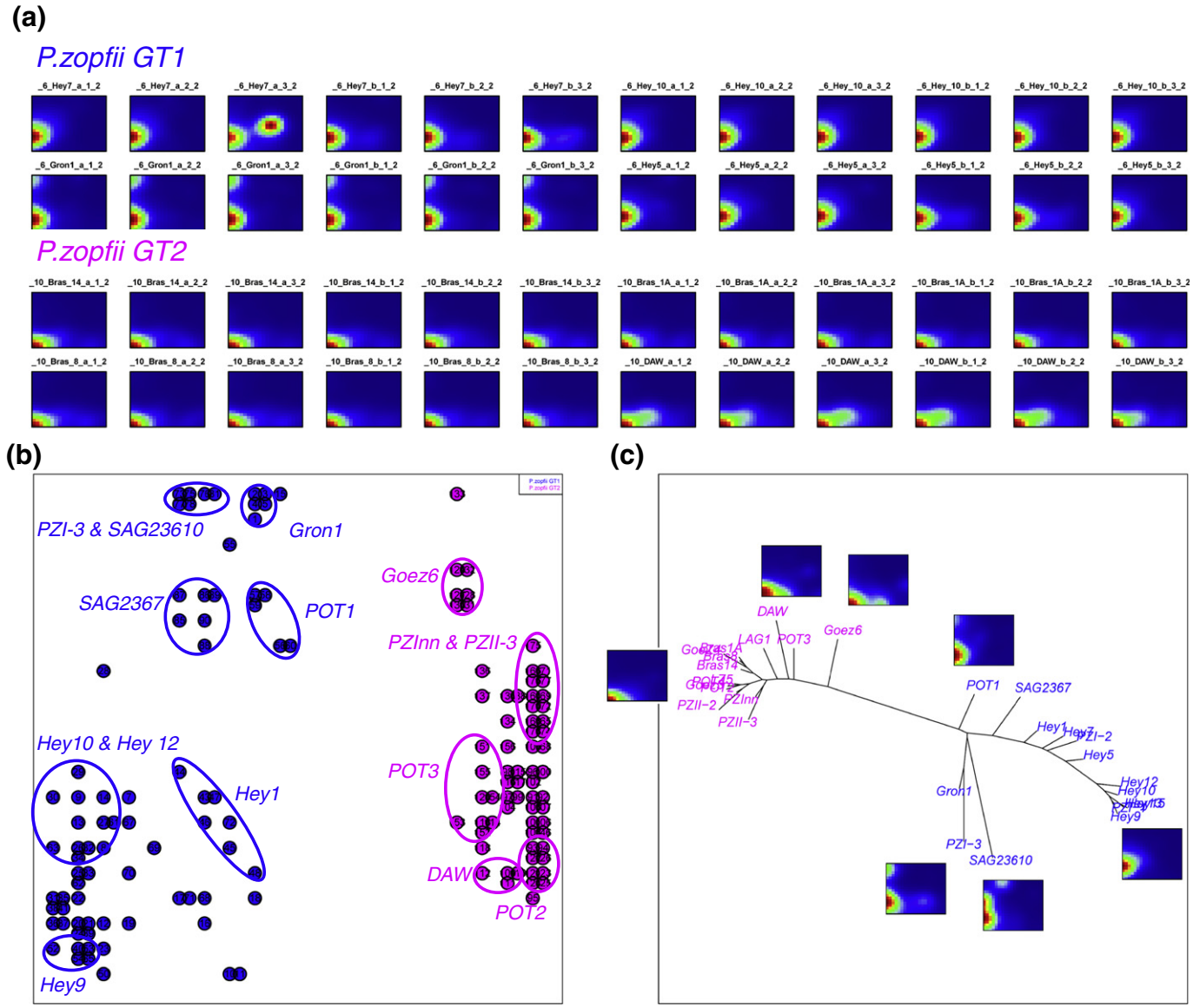

Fig. 11. Zoom-in analysis of *P. zopfii* GT1 (blue) and GT2 (magenta): Panel a shows SOM portraits of the same samples selected in Fig. 3. Panel b shows the respective 2nd level SOM and panel c the phylogenetic tree using the neighbor-joining algorithm. Selected batches of samples included in the circles are indicated in the 2nd level SOM.

separation of independent features. In this study we used the same peak lists as in our previous publication (von Bergen et al., 2009). We expect that binning width and also the peak selection algorithm affect the performance of the SOM and particularly the ability of the spot patterns to discriminate between different classes of samples. These methodical issues open additional options for the improvement of the SOM approach in future work. On the other hand, even the simple approach used here provides robust and reliable results combined with the intuitive and easy to understand visual presentation of the results in terms of individual sample portraits and similarity plots.

4. Conclusions

The MALDI-TOF peaklists of a series of algae samples in combination with SOM machine learning were used to detect and to classify Prototheca species, genotype variants and isolates. Our SOM approach decomposes the original MS-spectra into meta-spectra each of them is associated with a cluster of single peaks of similar amplitude profiles in the samples studied. The amplitudes of the meta-peaks taken from the meta-spectra in the individual samples are transformed into mosaic images visualizing the algae-specific distribution of high- and low-amplitude meta-peaks in two dimensions. The color texture of these spectral portraits allows the direct comparison of samples.

Particularly, a species-specific pattern of MS intensities were readily discernable in the obtained gallery of individual *Prototheca* portraits. As a rule of thumb, they reveal a species-specific single spot of high amplitude MS-peaks which allows the easy identification of the respective class of algae.

The SOM method compresses the original set of high-dimensional MS-spectral data in two consecutive steps: Firstly, similar profiles of MS-peak amplitudes are collected into meta-peak clusters, which reduces the number of relevant features from about 1400 single peaks to 400 meta-peaks in our application. Secondly, the textures of the obtained SOM are decomposed into a few (typically one per class) spots of high-amplitude meta-peaks. This 'double compression' sequentially applies global (similar profiles) and local (high amplitudes in part of the samples) criteria. It also combines supervised and unsupervised clustering: The first step is based on a predefined number of meta-spectra whereas the second step selects the number of spot-clusters detected. Unique spots observed only in one species collect MS fingerprint spectra of this species.

The use of meta-spectra instead of single-spectra reduces the dimension of the data and leads to an increased discriminating power in downstream analyses such as phylogenetic tree reconstruction owing to essentially two facts: Firstly, the set of meta-spectra better represents the diversity of MS-patterns inherent in the data. Secondly, the meta-spectra possess an improved signal-to-noise characteristics compared with a comparable collection of single spectra. Hence, meta-spectra can be seen as a natural choice to detect context-dependent MS-spectral patterns in large sets of samples.

Our SOM-method thus further improves the MALDI-MS based classification approach of harmless and pathogenic algae presented previously (von Bergen et al., 2009). Importantly, SOM-data compression of MS-spectra is not restricted to the examples presented here. We expect that it improves MS-based classifications and feature selection in general.

Acknowledgements

The project LIFE is financially supported by the European Funds for Regional Development (EFRE) and the State of Saxony (Ministry for Science and the Arts). HW was kindly supported by Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE).

References

Arnold, P., Ahearn, D.G., 1972. The systematics of the genus Prototheca with a description of a new species *P. filamenta*. Mycologia 64 (2), 265–275. http://www.jstor.org/stable/3757830.

Bianchi, M., Robles, A.M., Vitale, R., Helou, S., Arechavala, A., Negroni, R., 2000. The usefulness of blood culture in diagnosing HIV-related systemic mycoses: evaluation of a manual lysis centrifugation method. Med. Mycol. 38 (1), 77–80. http://www.ncbi.nlm.nih.gov/pubmed/10746231 (February).

Binder, Hans, Fasold, Mario, Hopp, Lydia, Cakir, Volkan, von Bergen, Martin, Wirth, Henry, 2011. Genomic and molecular phenotypic portraits—exploring the 'OMEs' with individual resolution. HIBIT 2011 Proceedings.

Bishop, C.M., Svensén, M., Williams, C.K.I., 1998. GTM: the generative topographic mapping. Neural Comput. 10 (1), 215–234. http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017953.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) 19 (2), 185–193 http://www.ncbi.nlm.nih.gov/pubmed/12538238 (January).

Breiman, L., 2001. Random forests. Machine learning. http://www.springerlink.com/index/U0P06167N6173512.pdf.

Bright, John J., Claydon, Martin A., S., Majeed, Gordon, Derek B., 2002. Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. J. Microbiol. Methods 48 (2–3), 127–138. http://www.ncbi.nlm.nih.gov/pubmed/11777563 (February).

Buckhaults, Phillip, Zhen, Zhen, Chen, Yu-Chi, Wang, Tian-Li, Croix, Brad St, Saha, Saurabh, Bardelli, Alberto, et al., 2003. Identifying tumor origin using a gene expression-based classification map. Cancer Res. 63 (14), 4144–4149. http://www.ncbi.nlm.nih.gov/pubmed/12874019 (July).

Campbell, P.M., 2005. Species differentiation of insects and other multicellular organisms using matrix-assisted laser desorption/ionization time of flight mass spectrometry protein profiling. Syst. Entomol. 30 (2), 186–190. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3113.2004.00279.x/full.

Covell, David G., Wallqvist, Anders, Rabow, Alfred A., Thanki, Narmada, 2003. Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. Mol. Cancer Ther. 2 (3), 317–332. http://www.ncbi.nlm.nih.gov/pubmed/12657727 (March).

Demirev, P.A., Ho, Y.P., Ryzhov, V., Fenselau, C., 1999. Microorganism identification by mass spectrometry and protein database searches. Anal. Chem. 71 (14), 2732–2738. http://www.ncbi.nlm.nih.gov/pubmed/10424165 (July 15).

Díaz-Uriarte, Ramón, Alvarez de Andrés, Sara, 2006. Gene selection and classification of microarray data using random forest. BMC Bioinforma. 7, 3. doi:10.1186/1471-2105-7-3http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1363357&tool=pmcentrez&rendertype=abstract (January).

Eichler, Gabriel S., Huang, Sui, Ingber, Donald E., 2003. Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics (Oxford, England) 19 (17), 2321–2322. http://www.ncbi.nlm.nih.gov/pubmed/14630665 (November).

Erlich, H.A., Gelfand, D., Sninsky, J.J., 1991. Recent advances in the polymerase chain reaction. Science (New York, N.Y.) 252 (5013), 1643–1651 (June 21).

Feltens, Ralph, Görner, Renate, Kalkhof, Stefan, Gröger-Arndt, Helke, von Bergen, Martin, 2010. Discrimination of different species from the genus Drosophila by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. BMC Evol. Biol. 10, 95. doi:10.1186/1471-2148-10-95http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2858148&tool=pmcentrez&rendertype=abstract (January).

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (3), 453. http://biomet.oxfordjournals.org/content/58/3/453.short.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.) 286 (5439), 531–537. http://www.ncbi.nlm.nih.gov/pubmed/10521349 (October).

Guo, Yuchun, Eichler, Gabriel S., Fing, Ying, Ingber, Donald E., Huang, Sui, 2006. Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers. J. Biomed. Biotechnol. 2006 (5), 69141. doi:10.1155/JBB/2006/69141http://www.ncbi.nlm.nih.gov/pubmed/17489018 (January).

Helm, D., Labischinski, H., Schallehn, G., Naumann, D., 1991. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. J. Gen. Microbiol. 137 (1), 69–79. http://www.ncbi.nlm.nih.gov/pubmed/1710644 (January).

Irenge, L.M., Durant, J.-F., Tomaso, H., Pilo, P., Olsen, J.S., Ramisse, V., Mahillon, J., Gala, J.-L., 2010. Development and validation of a real-time quantitative PCR assay for rapid identification of *Bacillus anthracis* in environmental samples. Appl. Microbiol. Biotechnol. 88 (5), 1179–1192. doi:10.1007/s00253-010-2848-0http://www.ncbi.nlm.nih.gov/pubmed/20827474 (November).

Jehmlich, Nico, Schmidt, Frank, Taubert, Martin, Seifert, Jana, von Bergen, Martin, Richnow, Hans-Hermann, Vogt, Carsten, 2009. Comparison of methods for simultaneous identification of bacterial species and determination of metabolic activity by protein-based stable isotope probing (Protein-SIP) experiments. Rapid Commun Mass Spectrom. 23 (12), 1871–1878. doi:10.1002/rcm.4084http://www.ncbi.nlm.nih.gov/pubmed/19449321 (June).

Kohonen, T., 1982. Self-organizing formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.

Lass-Flörl, Cornelia, Mayr, Astrid, 2007. Human protothecosis. Clin. Microbiol. Rev. 20 (2), 230–242. doi:10.1128/CMR.00032-06http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1865593&tool=pmcentrez&rendertype=abstract (April).

Lass-Flörl, Cornelia, Fille, Manfred, Gunsilius, Eberhard, Gastl, Günther, Nachbaur, David, 2004. Disseminated infection with Prototheca zopfii after unrelated stem cell transplantation for leukemia. J. Clin. Microbiol. 42 (10), 4907–4908. doi:10.1128/JCM.42.10.4907-4908.2004http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=522359&tool=pmcentrez&rendertype=abstract (October).

Liaw, Andy, Wiener, Matthew, 2002. Classification and regression by randomForest. R news 2 (3), 18–22. http://www.webchem.science.ru.nl/PRiNS/rF.pdf.

Maier, T., Kostrzewa, Markus, 2007. Fast and reliable MALDI-TOF MS-based microorganism identification. Chem. Today 25, 68–71.

Matsuda, T., Matsumoto, T., 1992. Protothecosis: a report of two cases in Japan and a review of the literature. Eur. J. Epidemiol. 8 (3), 397–406. http://www.ncbi.nlm.nih.gov/pubmed/1397204 (May).

Mazzeo, Maria Fiorella, De Giulio, Beatrice, Guerriero, Giulia, Ciarcia, Gaetano, Malorni, Antonio, Russo, Gian Luigi, Siciliano, Rosa Anna, 2008. Fish authentication by MALDI-TOF mass spectrometry. J. Agric. Food Chem. 56 (23), 11071–11076. doi:10.1021/jf8021783http://www.ncbi.nlm.nih.gov/pubmed/19007297 (December 10).

Meinicke, Peter, Lingner, Thomas, Kaever, Alexander, Feussner, Kirstin, Göbel, Cornelia, Feussner, Ivo, Karlovsky, Petr, Morgenstern, Burkhard, 2008. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. Algorithms Mol. Biol. 3, 9. doi:10.1186/1748-7188-3-9http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2464586&tool=pmcentrez&rendertype=abstract (January).

Nikkilä, Janne, Törönen, Petri, Kaski, Samuel, Venna, Jarkko, Castrén, Eero, Wong, Garry, 2002. Analysis and visualization of gene expression data using self-organizing maps. Neural Netw. 15 (8–9), 953–966. http://www.ncbi.nlm.nih.gov/pubmed/12416686.

Pore, R.S., 1985. Prototheca taxonomy. Mycopathologia 90 (3), 129–139. http://www.springerlink.com/index/M6N027K125G2484G.pdf.

Pozhitkov, Alex E., Bailey, Kyle D., Noble, Peter A., 2007. Development of a statistically robust quantification method for microorganisms in mixtures using oligonucleotide microarrays. J. Microbiol. Methods 70 (2), 292–300. doi:10.1016/j.mimet.2007.05.001http://www.ncbi.nlm.nih.gov/pubmed/17553581 (August).

Pozhitkov, Alex E., Beikler, Thomas, Flemmig, Thomas, Noble, Peter A., 2011. High-throughput methods for analysis of the human oral microbiome. Periodontol. 55 (1), 70–86. doi:10.1111/j.1600-0757.2010.00380.xhttp://www.ncbi.nlm.nih.gov/pubmed/21134229 (February).

Roesler, Uwe, Möller, Asia, Hensel, Andreas, Baumann, Daniela, Truyen, Uwe, 2006. Diversity within the current algal species *Prototheca zopfii*: a proposal for two *Prototheca zopfii* genotypes and description of a novel species, *Prototheca blaschkeae* sp. nov. Int. J. Syst. Evol. Microbiol. 56 (Pt 6), 1419–1425. doi:10.1099/ijs.0.63892-0http://www.ncbi.nlm.nih.gov/pubmed/16738123 (June).

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425. http://www.ncbi.nlm.nih.gov/pubmed/3447015 (July).

Schmidt, F., Schmid, M., Jungblut, P.R., Mattow, J., Facius, A., Pleissner, K.P., Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis, 2003. J. Am. Soc. Mass Spectrom. 14 (9), 943–956. doi:10.1016/S1044-0305(03)00345-3http://www.ncbi.nlm.nih.gov/pubmed/12954163 (September).

Stenner, V.J., Mackay, B., King, T., Barrs, V.R.D., Irwin, P., Abraham, L., Swift, N., et al., 2007. Protothecosis in 17 Australian dogs and a review of the canine literature. Med. Mycol. 45 (3), 249–266. doi:10.1080/13693780601187158http://www.ncbi.nlm.nih.gov/pubmed/17464846 (May).

Suna, Teemu, Salminen, Aino, Soininen, Pasi, Laatikainen, Reino, Ingman, Petri, Mäkelä, Sanna, Savolainen, Markku J., et al., 2007. 1H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising maps. NMR Biomed. 20 (7), 658–672. doi:10.1002/nbm.1123http://www.ncbi.nlm.nih.gov/pubmed/17212341 (November).

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. U. S. A. 96 (6), 2907–2912. http://www.ncbi.nlm.nih.gov/pubmed/10077610 (March).

Tautz, Diethard, Arctander, Peter, Minelli, Alessandro, Thomas, Richard H., Vogler, Alfried P., 2002. DNA points the way ahead in taxonomy. Nature 418 (6897), 479. doi:10.1038/418479ahttp://www.ncbi.nlm.nih.gov/pubmed/12152050 (August 1).

van Veen, S.Q., Claas, E.C.J., Kuijper, Ed J., 2010. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. J. Clin. Microbiol. 48 (3), 900–907. doi:10.1128/JCM.02071-09http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832429&tool=pmcentrez&rendertype=abstract (March).

Villmann, Thomas, Schleif, Frank-Michael, Kostrzewa, Markus, Walch, Axel, Hammer, Barbara, 2008. Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. Brief. Bioinform. 9 (2), 129–143. doi:10.1093/bib/bbn009http://www.ncbi.nlm.nih.gov/pubmed/18334515 (March).

von Bergen, M., Wirth, H., Binder, H., et al., 2009. Identification of harmless and pathogenic algae of the genus Prototheca by MALDI-MS. http://www3.interscience.wiley.com/journal/122505848/abstract.

Wang, Junbai, Delabie, Jan, Aasheim, Hans, Smeland, Erlend, Myklebost, Ola, 2002. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinforma. 3, 36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=138792&tool=pmcentrez&rendertype=abstract (November).

Wirth, Henry, Loffler, Markus, von Bergen, Martin, Binder, Hans, 2011. Expression cartography of human tissues using self organizing maps. BMC Bioinforma. 12 (1), 306. doi:10.1186/1471-2105-12-306http://www.biomedcentral.com/1471-2105/12/306.

Wongravee, Kanet, Lloyd, Gavin R., Silwood, Christopher J., Grootveld, Martin, Brereton, Richard G., 2010. Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling. Anal. Chem. 82 (2), 628–638. doi:10.1021/ac9020566http://www.ncbi.nlm.nih.gov/pubmed/20038089 (January).