Hierarchical PCA using Tree-SOM for the Identification of Bacteria

Stephan Simmuteit[1], **Frank-Michael Schleif**[1], **Thomas Villmann**[2] and **Markus Kostrzewa**[3]

(1) Univ. Leipzig, Dept. of Medicine, 04107 Leipzig, Germany
(2) Univ. of Appl. Sc. Mittweida, Dept. of MPI, 09648 Mittweida / Germany
(3) Department of Bioanalytics, Bruker Daltonik GmbH, 28359 Bremen, Germany
{schleif}@informatik.uni-leipzig.de,+49(0)3419718955
{stephan.simmuteit}@medizin.uni-leipzig.de,+49(0)3419718954
{thomas.villmann}@hs-mittweida.de,+49(0)372758-1328
{km}@bdal.de

Abstract. In this paper we present an extended version of *Evolving Trees* using Oja's rule. Evolving Trees are extensions of *Self-Organizing Maps* developed for hierarchical classification systems. Therefore they are well suited for taxonomic problems like the identification of bacteria. The paper focus on clustering and visualization of bacteria measurements. A modified variant of the Evolving Tree is developed and applied to obtain a hierarchical clustering. The method provides an inherent PCA analysis which is analyzed in combination with the tree based visualization. The obtained loadings support insights in the classification decision and can be used to identify features which are relevant for the cluster separation.

Key words: tree som, bacteria identification, mass spectrometry, hierarchical PCA, unsupervised feature selection

1 Introduction

The identification of bacteria in medical and biological environments by means of classical methods like gram stain is time consuming and frequently leads to mistakes in separation of species or even genus. These data are categorized in a taxonomical tree-structure. It can be expected that the supporting measurements reflect such a structure. Further its known that for some bacteria molecular finger prints exist [9]. Taking these two aspects into account we derive the *Hierarchical PCA-based Evolving Tree* to obtain an optimal compact encoding and tree-structured representation of such data based on *Evolving Trees* [13] and Oja-PCA learning [12].

The utilization of mass spectrometry (MS) provides a fast and reproducible way to receive bio-chemical information to identify bacteria cultured on nutrient solution. One task in this line is an appropriate classification of the high-dimensional mass spectra. This requires a reasonable classification structure to achieve adequate storage and retrieval performance. It is further valuable to obtain interpretable visualizations of the data for a later expert analysis. Existing approaches are based on the direct comparison of spectra with manually selected reference spectra by means of a (pre-filtered) peak matching including their intensity as well as their mass position [9, 11].

The application of MS for bacteria identification is quite new and a representation of the taxonomic (tree-) nature of bacteria is difficult. The problem of discriminating bacteria species with MS is described in [1]. Forero et al. use extracted features from images of bacteria to identify them [5]. Discrimination of bacteria can be done also by bio-markers based on MS spectra [10]. Most of those approaches are also based on the evaluation of the peak intensities. In case of bacteria even the peak intensities alone are an unsafe criterion. Further, the encoded peaks (line spectra) to be compared are huge-dimensional vectors representing a functional relation. Fast and reliable investigation of line spectra requires, on the one hand side, an adequate processing, which preserves the relevant information as good as possible. On the other hand, optimum interpretable data structures are required.

This contribution provides new aspects for efficient information-preserving representation of line spectra by a data-driven tree generation using the *Hierarchical PCA-based Evolving Trees (ET)*.

2 Evolving Trees and hierarchical PCA

As mentioned above the 'natural' identification methodology in taxonomy/analysis of bacteria is tree structured. Therefore, in context of machine learning, decision trees (DT) may come into mind. However, DTs don't integrate structural data information like data shape and density in an adequate manner during tree generation. An alternative is presented by Pakkanen et al. – the *evolving trees (ET)* for which we provide a formal definition later on. The ET-approach is an extension of the concept of *self-organizing maps* (SOMs) introduced by Kohonen [6].

SOMs project high-dimensional vectorial data onto a predefined low-dimensional regular grid usually chosen as a hypercube. This mapping is topology preserving under certain conditions, i.e. in case of no violations similar data points in the data space are mapped onto the same or neighbored grid nodes. For this purpose, to each node a weight vector, also called prototype, is assigned. A data point is mapped onto this node, the prototype of which is closest according to a similarity measure in the data space, usually the Euclidean distance. This rule is called winner-take all. In this sense, all data points mapped onto the same node are called *receptive field* of this node and the respective prototype is a representative of this field.

2.1 Evolving Trees

Yet, the usual rectangular lattice as output structure is only mandatory. Other choices are possible depending on the task. ETs use trees as output structures and, hence, are potentially suited for mapping of vectorial data with an inherent hierarchical structure.

Suppose we consider an ET $\mathcal{T}$ with nodes $r \in R_{\mathcal{T}}$ (set of nodes) and root r_0 which has the depth level $l_{r_0} = 0$. A node r with depth level $l_r = k$ is connected to its successors r' with level $l_{r'} = k + 1$ by directed edges $\varepsilon_{r \to r'}$ with length is unit. The set of all direct successors of the node r is denoted by S_r. If $S_r = \varnothing$

is valid, the node r is called a leaf. The degree of a node r is $\delta_r = \#S_r$, here assumed to be constant δ for all nodes except the leafs. A sub-tree $\mathcal{T}_r$ with node r as root is the set of all nodes $r' \in R_{\mathcal{T}_r}$ such that there exists a directed cycle-free path $p_{r\to r'} = \varepsilon_{r\to m} \circ \ldots \circ \varepsilon_{m'\to r'}$ with $m, \ldots, m' \in R_{\mathcal{T}_r}$ and $\circ$ as the concatenation operation. $L_{p_{r\to r'}}$ is the length of path $p_{r\to r'}$, i.e. the number of concatenations plus 1. The distance $d_{\mathcal{T}}(r, r')$ between nodes r, r' is defined as

$$d_{\mathcal{T}}(r, r') = L_{p_{\hat{r}\to r}} + L_{p_{\hat{r}\to r'}} \tag{1}$$

with paths $p_{\hat{r}\to r}$ and $p_{\hat{r}\to r'}$ in the sub-tree $\mathcal{T}_{\hat{r}}$ and $R_{\mathcal{T}_{\hat{r}}}$ contains both r and r' and the depth level $l_{\hat{r}}$ is maximum for all sub-trees $\mathcal{T}_{\hat{r}'}$ which contain r and r'. A connecting path between a node r and a node r' is defined as follows: let $p_{\hat{r}\to r'}$ and $p_{\hat{r}\to r}$ be direct paths such that $L_{p_{\hat{r}\to r'}} \cdot L_{p_{\hat{r}\to r}}$ is $d_{\mathcal{T}}(r, r')$. Then $p_{r\to r'}$ is the reverse path $p_{r'\to\hat{r}} \cdot p_{\hat{r}\to r}$ and the node set of P is denoted by $\mathcal{N}_{p_{r\to r'}}$. As for usual SOMs, each node r is equipped with a prototype $\mathbf{w}_r \in \mathbb{R}^D$, provided that the data to be processed are given by $\mathbf{v} \in V \subseteq \mathbb{R}^D$. Further, we assume a differentiable similarity measure $d_V : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The winner detection is different from usual SOM but remains the concept of winner-take-all. For a given subtree $\mathcal{T}_r$ with root r the *local winner* is

$$s_{\mathcal{T}_r}(\mathbf{v}) = \arg\min_{r\in S_r} \left(d_V(\mathbf{v}, \mathbf{w}_r)\right) \tag{2}$$

If $s_{\mathcal{T}_r}(\mathbf{v})$ is a leaf then it is also the overall winner node $s(\mathbf{v})$. Otherwise, the procedure is repeated recursively for the sub-tree $\mathcal{T}_{s_{\mathcal{T}_r}}$. The *receptive field* Ω_r of a leaf r (or its prototype) is defined as

$$\Omega_r = \{\mathbf{v} \in V | s(\mathbf{v}) = r\} \tag{3}$$

and the receptive field of root r' of a sub-tree $\mathcal{T}_{r'}$ is defined as

$$\Omega_{r'} = \cup_{r''\in R_{\mathcal{T}_{r'}}} \Omega_{r''} \tag{4}$$

The adaptation of the prototypes $\mathbf{w}_r$ takes place only for those prototypes, where the nodes r of are leafs. The other nodes remain fixed. This learning for a randomly selected data point $\mathbf{v} \in V$ is neighborhood-cooperatively as in usual SOM:

$$\triangle \mathbf{w}_r = \epsilon h_{SOM}(r, s(\mathbf{v}))(\mathbf{v} - \mathbf{w}_r) \tag{5}$$

with $s(\mathbf{v})$ being the overall winner and $\epsilon > 0$ a small learning rate. The neighborhood function $h_{SOM}(r, r')$ is defined as a function depending on the tree distance $d_{\mathcal{T}}$ usually of Gaussian shape

$$h_{SOM}(r, r') = \exp\Big(\frac{-\left(d_{\mathcal{T}}(r, r')\right)^2}{2\sigma^2}\Big). \tag{6}$$

with neighborhood range σ.

Unlike for the SOM we cannot guarantee that $s(\mathbf{v})$ is the true best matching unit (bmu), because the tree model is subject of a stochastic optimization process.

The whole ET learning is a repeated sequence of adaptation phases according to the above mentioned prototype adaptation and tree growing beginning with a minimum tree of root r_0 and its δ successors as leafs. The decision, which leafs become roots of sub-trees at a certain time can be specified by the user. Subsequently for each node r a counter b_r is defined. This counter is increased if the corresponding node becomes a winner and the node is branched if a given threshold $\theta \in \mathbb{N}, \theta > 0$ is reached.

Possible criteria might be the variance of the receptive fields of the prototypes or the number of winner hits during the competition. The prototypes of the new leafs should be initialized in a local neighborhood of the root prototype according to d_V. Hence, the ET also can be taken as a special growing variant of SOM as it is known for example from [2].

Since ETs are extended variants of usual SOM one can try to transfer evaluation methods known from SOMs to ETs. Unknown samples can be identified using the ET in the following way. The ET is fully labeled by assignment of a label to each node by an analysis of the receptive fields of the corresponding sub-trees. The root node remains unlabeled. For each receptive field a common label is determined by a majority voting of the contained samples and their labels. An unknown, new item is preprocessed as described later on. For this item the *bmu* in the tree is determined in accordance to Equation (2) and $s(\mathbf{v})$ is calculated. The label of the receptive field of $s(\mathbf{v})$ defines the label of the item.

2.2 Hierarchical PCA by Evolving Tree learning using Oja's rule

In [12] a learning rule for neuron models has been proposed which inherently provides a principal component analyses of the represented data. This rule was recently used in [7] to get an optimal data encoding and proven to be effective in learning using neighborhood cooperativeness. We combine this approach with the learning of Evolving Trees such that the prototype representing a data cluster become the first eigenvector of this cluster. In this way a hierarchical PCA can be calculated. We replace the learning rule of Equation (5) by the following Oja based learning dynamic but keeping the neighborhood cooperativeness of ET:

$$\triangle \mathbf{w}_r = \epsilon h_{ET}\left(r, s\left(\mathbf{v}\right)\right) O\left(\mathbf{v} - \mathbf{O}\mathbf{w}_r\right) \tag{7}$$

$$O \qquad\qquad = < \mathbf{v}, \mathbf{w_r} > \tag{8}$$

Further the winner determination of Equation (2) is changed accordingly

$$s_{\mathcal{T}_r}\left(\mathbf{v}\right) = \arg\max_{r \in S_r}\left(< \mathbf{v}, \mathbf{w}_r >\right) \tag{9}$$

As pointed out in [12] the update for the weight vector $\mathbf{w}_r$ as defined by Equation (7) will, neglecting statistical fluctuations, tend to the dominant eigenvector c of the input correlation matrix C of the input data v limited to the receptive field of $\mathbf{w_r}$. Using this approach we obtain eigenvectors for each cluster, at each depth level l_r for each node of the tree. The first eigenvector as obtained from an analysis of the prototype $\mathbf{w_r}$ at l_{r_o} is the regular first principal component of the whole data set. With increasing depth of the tree the data are clustered by the Tree-SOM approach and a hierarchical PCA analysis of the sub-clusters become

available. The principal components can be used to analyse and visualize the cluster separability. Further the obtained loadings provide insights in a variance based analysis of the individual input dimensions of the clusterings such that separating features become apparent.

3 Evolving Tree applied on mass spectra of bacteria

The introduced *Hierarchical PCA-based Evolving Tree* is now applied to investigate MS-spectra for identification of bacteria. These data are spectra of different species of *Vibrio*- and *Listeria*-bacteria. Thereby we use the spectra in a pre-processed form of line spectra. The resulting identification is visualized and it is shown how the obtained hierarchical PCA model can be interpreted.

3.1 Data, Measurement and Pre-processing

The data used in the experiments are MS spectra of 56 different vibrio species and 7 different *Listeria* species. Every data-set contains about $20 - 40$ single spectra, being measurements of the same bacterium. Together there are 1452 spectra of vibrio and 231 spectra of *Listeria*. Each MS measurement is processed as described later on. Biological details on the bacteria samples can be obtained from [4].

Details on the mass spectrometry technique can be found in [8]. At the end of the measurement process one obtains for each measurement a spectrum with a mass axis in m/z respectively Dalton and an unit-less intensity for every mass. The spectrum is encoded as a high-dimensional vector (profile spectrum) of intensities, often visualized as a function of mass.

The standard pre-processing to generate a line spectrum (consisting only of peaks) is provided by the measurement system as detailed in [3]. A line spectrum typically consists of around $100 - 500$ peaks depending on the sample complexity and system mode while the profile spectra are original given as measurements with around 40 000 sample points. In order to the line spectra for our approach the input vectors of peak lists are mapped onto a global mass vector covering every appearing peak within a predefined tolerance (here 500 ppm) depending on the expected measurement accuracy.

The resulting aligned peak-lists are now located in the same data space, still very high-dimensional. For the *Listeria* data the line spectra have a dimensionality of $D = 1181$ (peak positions) whereas for the vibrio data the dimensionality is given as $D = 2382$.

3.2 Experiments

Euclidean distance is used to find the *bmu*. $\delta_r = 3$ for all nodes without leafs. The learning is done in accordance to the standard SOM approach, thereby the initial learning rate α_0 is defined as $\alpha_0 = 0.2$ which is logarithmically decreased during learning to a final value of $\alpha_{\text{end}} = 0.01$. The neighborhood cooperation value σ is initialized with $\sigma = 1$ and logarithmically decreased to $\sigma = 0.35$ in accordance to suggestions given in [15]. The total number of learning iterations I

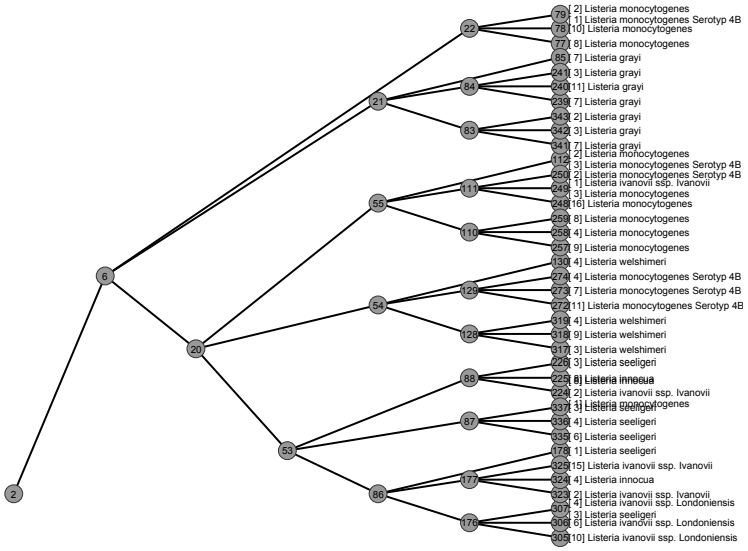

Fig. 1. Evolving Tree of seven *Listeria* species

is determined depending on the number of training samples, the desired number of clusters #C, δ_r and θ as shown in [14].

We apply the proposed methodology on the two data sets. In the first experiment an hierarchical PCA using the ET is generated for the *Listeria* and the *Vibrio* data. This is a simplified example of a bacteria identification on the genus level. In a second experiment we consider the *Listeria* bacteria only. Thereby we assume that the genus of the considered bacteria is already identified as *Listeria* using the first tree and the remaining task consists in an identification and visualization of the species and subspecies level. For both settings we generate the tree, analyze the hierarchical, local PCA visualizations and identify relevant mass positions (features) by means of PCA loadings. For simplicity we provide the plots for the *Listeria* data, only. In Figure 1 the Evolving-Tree for the *Listeria* data is shown[1]. We observe a quite clear separation of the different *Listeria* species in the tree, but also some mixed clusters occur. Especially for the *monocytogenes* data subclusters can be identified, this however is a intended effect because the *monocytogenes* group is known to be diverse. Here a single taxonomical label does not perfectly reflect the biochemical picture[2]. In Figure 5 we analyze the loadings of the local PCA of node 6, as obtained in the hierarchical PCA using the ET. We observe a cut in the loadings histogram such that $2 - 7$ dimensions can be considered to be relevant. Taking these input dimensions into account a Pseudo-Gelview can be generated as depicted in Figure 3 showing a top/gelview of the spectra restricted to the peaks

[1] Here we show the subtree from the *Listeria/Vibrio*-Tree, but an individual generated tree is actually very similar, ignoring permutations.

[2] This effect becomes even more explicit for e.g. bacillus data - which are in fact multiple subgroups (genera) (not distinctly labeled in the taxonomy of bacteria)

intensities at the masses indicated by the PCA. Some peaks differentiate between *Listeria* groups by means of intensity variations, as e.g. in the first peak with moderate intensity values for the *ivanonvii*, a missing peak situation for the *grayi* and high intensities else. We noticed that in general a 0/1 encoding of the peak intensities (peak absent/present) is sufficient but for some species and subspecies the incorporation of intensity information is valuable. In addition a box plot of the projected data on the principal component as depicted in Figure 2 may provide further information on the separation potential of the hierarchical PCA based clustering. Doing a traversal through the feature loadings over the different tree levels the approach identified the following masses most relevant $4276.4Da, 4278.0Da, 5181.0Da, 9751.0Da$. The first three dimensions are relevant to separate the vibrio data from the *Listeria* and to get separations with the vibrio genus, separating different (but not all) vibrio species. The last dimension is a clear indicator for the presents of *Listeria*.

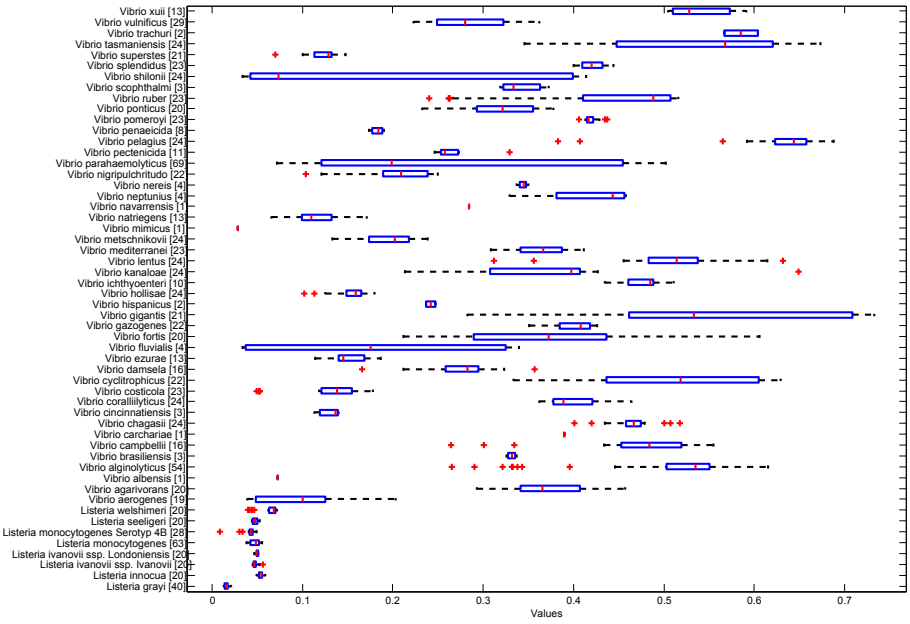

Fig. 2. Box plot of the pc's at the 2 node, branching the *Listeria* and most of the *Vibrio* data in an ET on the bacteria data.

In the Table 4 the most relevant dimensions for the *Listeria* experiment identified by the hierarchical PCA at node 6 are depicted. Similar analyses can be done for the other nodes as well. It should be noted that the identified masses at a specific node are interpreted as those dimensions explaining the largest variance of the data presented in the underlying clustering. This is an unsupervised interpretation, hence the relevant dimensions may not be relevant with respect to

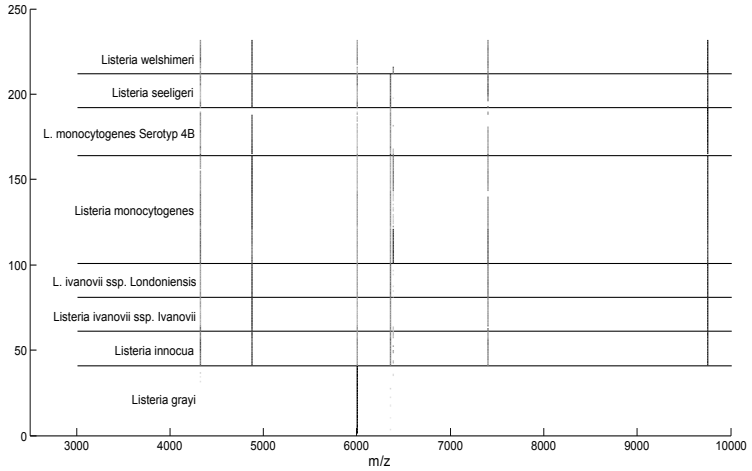

Fig. 3. Gelview of the *Listeria* data restricted to the identified most relevant masses.

a provided labeling. For bacteria data however we observed, that the highlighted dimensions are in general meaningful for the taxonomy as well.

Rank	Contribution	Dim.	Relevant Mass
1	0.6859	2243	9751.11
2	0.5860	897	4876.13
3	0.2190	1764	7402.22
4	0.1987	1441	6362.80
5	0.1879	664	4323.25
6	0.1323	1449	6388.37
7	0.1074	1307	6006.82

Fig. 4. Relevant masses contributing to the first principal component in the tree node (6) pooling all *Listeria* subspecies

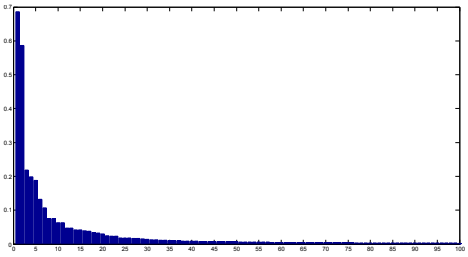

Fig. 5. Analysis of the loadings (truncated to 100) of the local PCA for node 6.

4 Conclusions

A method for an unsupervised hierarchical PCA based analysis of bacteria spectra from mass spectrometry has been presented. One obtains a hierarchical representation of the bacteria by means of a Evolving Tree with local principal components in a hierarchical manner. This can be used to get a better interpretation of the underlying clustering. The approach is unsupervised but nicely reflects the expected taxonomical ordering of the data. The approach can be used to identify masses which are relevant for the clustering in a hierarchical way, e.g. by traversing through the different levels of the tree. If the clustering

fits to an added set of meta information, as in our case, the taxonomy of bacteria the identified dimensions could be interpreted in a supervised scheme as well. The approach can be used to get highly interpretable representations of bacteria spectra and to get quick identifications with a logarithmic number of comparisons. [3]

References

1. S. B. Barbuddhe, T. Maier, G. Schwarz, M. Kostrzewa, H. Hof, E. Domann, T. Chakraborty, and T. Hain. Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology*, 74(17):5402–5407, 2008.
2. H.-U. Bauer and Th. Villmann. Growing a Hypercubical Output Space in a Self–Organizing Feature Map. *IEEE TNN*, 8(2):218–226, 1997.
3. Bruker Daltonik GmbH. Bruker BioTyper 2.0, user manual. available via http://www.bdal.de, 2008.
4. Bruker Daltonik GmbH. Bruker listeria and vibrio spectra. available via http://www.bdal.de (Dr. Markus Kostrzewa), 2008. personal communicated.
5. M.G. Forero, F. Sroubek, and G. Cristobal. Identification of tuberculosis bacteria based on shape and color. *Real-time Imaging*, 10(4):251–262, 2004.
6. Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ed. 1997).
7. K. Labusch, E. Barth, and T. Martinetz. Learning data representations with sparse coding neural gas. In *Proc. of ESANN'08*, pages 233–238, 2008.
8. Daniel C. Liebler. *Introduction to Proteomics*. Humana Press, 2002.
9. T. Maier and M. Kostrzewa. Fast and reliable maldi-tof ms-based microorganism identification. *Chemistry Today*, 25:68–71, 2007.
10. M. F. Mazzeo, A. Sorrentino, M. Gaita, G. Cacace, M. Di Stasio, A. Facchiano, G. Comi, A. Malorni, and R. A. Siciliano. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for the discrimination of food-borne microorganisms. *Applied and Environmental Microbiology*, 72(2):1180–1189, 2006.
11. A. Mellmann, J. Cloud, T. Maier, U. Keckevoet, I. Ramminger, P. Iwen, J. Dunn, G. Hall, D. Wilson, P. LaSala, M. Kostrzewa, and D. Harmsen. Evaluation of matrix-assisted laser desorption/ionization time-of-flight-mass spectrometry MALDI-TOF MS in comparison to 16s rrna gene sequencing for species identification of nonfermenting bacteria. *J. Clinical Microbiology*, 46:1946–1954, 2008.
12. Erkki Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.
13. Jussi Pakkanen, Jukka Iivarinen, and Erkki Oja. The evolving tree—a novel self-organizing network for data analysis. *Neural Process. Lett.*, 20(3):199–211, 2004.
14. Stephan Simmuteit. Effizientes Retrieval aus Massenspektrometriedatenbanken, Diplomarbeit, Technische Universität Clausthal. February 2008.
15. Th. Villmann, R. Der, M. Herrmann, and Th. Martinetz. Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.

[3] Acknowledgment: The authors are grateful to S. Klepel and T. Maier for providing the bacteria data and helpful discussions (both Bruker Daltonik Leipzig, Germany)