

Analysis of Proteomic Spectral Data by Multi Resolution Analysis and Self-Organizing Maps

Frank-Michael Schleif¹, Thomas Villmann¹ and Barbara Hammer²

(1) University Leipzig, Dept. of Medicine, 04107 Leipzig, Germany

(2) TU-Clausthal, Dept. of Math. & C.S., 38678 Clausthal-Zellerfeld, Germany

{schleif,villmann}@informatik.uni-leipzig.de, +49(0)3419718896

{hammer}@in.tu-clausthal.de, +49(0)53237271 [86] [39]

Abstract. Analysis and visualization of high-dimensional clinical proteomic spectra obtained from mass spectrometric measurements is a complicated issue. We present a wavelet based preprocessing combined with an unsupervised and supervised analysis by Self-Organizing Maps and a fuzzy variant thereof. This leads to an optimal encoding and a robust classifier incorporating the possibility of fuzzy labels.

Key words: fuzzy visualization, clinical proteomics, wavelet analysis, biomarker, spectra preprocessing

1 Introduction

Applications of mass spectrometry (ms) in clinical proteomics have gained tremendous visibility in the scientific and clinical community [9, 5]. One major objective is the search for potential biomarkers in complex body fluids like serum, plasma, urine, saliva, or cerebral spinal fluid. For this purpose, efficient analysis and visualization of large high-dimensional data sets derived from patient cohorts is crucial. Additionally, it is necessary to apply statistical analysis and pattern matching algorithms to attain validated signal patterns. A powerful tool for faithful data mining and visualization of potential high-dimensional data is the unsupervised self-organizing map (SOM) [6] which has been recently extended to a supervised counterpart in [8]. The later allows the determination of a prototype based fuzzy classification model (FLSOM). In contrast to the widely applied multilayer perceptron [3], prototype based classification allows an easy interpretation of the classification scheme, which is of particular interest for clinical applications. FLSOM leads to a robust fuzzy classifier where efficient learning of fuzzy labeled or partially contradictory data is possible.

Hereby a discriminative data representation is necessary. The extraction of such discriminant features is critical for spectral data and typically done by a parametric peak picking procedure. This peak picking is often focus of criticism because peaks may be insufficiently detected. To avoid this difficulties we focus on a wavelet encoding of the spectral data to get discriminative features. Thereby the obtained wavelet coefficients are sufficient to reconstruct the signal, still containing all relevant information of the spectra. However this better discriminating set of features is typically more complex and hence a robust approach to determine the desired classification model is needed.

2 Bioinformatic methods

The classification of mass spectra involves in general the two steps peak picking to locate and quantify positions of peaks within the spectrum and feature extraction from the obtained peak list. In the first step a number of procedures as baseline correction, optional denoising, noise estimation and normalization must be applied. Upon these prepared spectra the peaks have to be identified by scanning all local maxima and the associated peak endpoints followed by a S/N thresholding such that one obtains the desired peak list.

The procedure of peak picking is standard, and has been done using ClinProTools for comparison in this paper (details in [5]). Here we propose an alternative which simplifies the procedure and preserves all (potentially small) peaks containing relevant information by use of the discrete wavelet transformation (DWT). The feature extraction has been done by Wavelet analysis using the Matlab Wavelet-Toolbox¹ and with ClinProTools to obtain peak lists with peak areas as final features. In that way both feature lists can be related back to original mass position in the spectral data which is essential for further biomarker analysis. In a first step a feature selection procedure using the Kolmogorov-Smirnoff test (KS-test) was applied. Thereby the test was used to identify features which show a significant ($p < 0.01$) discrimination between the two groups (cancer, control). This is done in accordance to [12] where also a generation to a multiclass experiment is given. The roughly reduced set has been further processed by FLSOM to obtain a classification model with a *small*, ranked set of features, crossvalidated by a 10-fold cross validation procedure.

2.1 Feature Extraction and Denoising with the Bi-orthogonal Discrete Wavelet Transform

Wavelets have been developed into powerful tools [1, 7] used for noise removal and data compression. The discrete version of the continuous wavelet transform leads to the concept of a multiresolution analysis (MRA). This allows a fast and stable wavelet analysis and synthesis. The analysis becomes more precise if the wavelet shape is adapted to the signal to be analyzed. For this reason one can apply the so called bi-orthogonal wavelet transform [2] which uses two pairs of scaling and wavelet functions. One is for the decomposition/analysis and the other one for reconstruction/synthesis. The advantage of the bi-orthogonal wavelet transform is the higher degree of freedom for the shape of the scaling and wavelet function. In our analysis such a smooth synthesis pair was chosen to avoid artifacts. It can be expected that a signal in the time domain can be represented by a small number of a relatively large set of coefficients from the wavelet domain. The spectra are reconstructed in dependence of a certain approximation level L of the MRA which can be considered as a hard-thresholding. The denoised spectrum looks similar to the reconstruction as depicted in Figure 1. The starting point for an argumentation is the simplest example of a MRA which can be defined by the characteristic function $\chi_{[0,1]}$. The corresponding wavelet is the so-called *Haar* wavelet. Assume that the denoised spectrum $f \in L_2(\mathbb{R})$ has a peak with

¹ The Matlab Wavelet-Toolbox can be obtained from www.mathworks.com

endpoints $2^j k$ and $2^j(k+1)$, the integral of the peak can be written as

$$\int_{2^j k}^{2^j(k+1)} f(t) dt = \int_{\mathbb{R}} f(t) \chi_{[2^j k, 2^j(k+1))}(t) dt$$

Obviously the right hand side is the Haar DWT scaling coefficient $c_{j,k} = \langle f, \psi_{j,k} \rangle$ at scale $a = 2^j$ and translation $b = 2^j k$. One obtains approximation- and detail-coefficients [2]. The approximation coefficients describe a generalized peak list of the denoised spectrum encoding primal spectral information and depending on the level L which is determined with respect to the measurement procedure. For linear MALDI-TOF spectra a device resolution of $500 - 800 Da$ can be expected. This implies limits to the minimal peak width in the spectrum and hence, the reconstruction level of the Wavelet-Analysis should be able to model corresponding peaks. A level $L = 4$ is appropriate for our problem (see Figure 1). Applying this

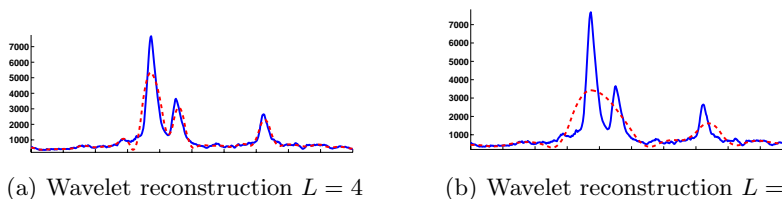


Fig. 1. Wavelet reconstruction of the spectra with $L = 4, 5$, x -mass positions, y -arbitrary unit. The original signal is plotted with the solid line. One observes that a wavelet analysis with $L = 5$ is too rough to approximate the sharp peaks.

procedure on the spectra with an initial number of 22306 measurement points per spectrum one obtains 602 wavelet coefficients and 91 peak areas used as representative features per spectrum. Subsequently, the data were processed by FLSOM in a 10-fold cross validation procedure.

2.2 Fuzzy labeled Self Organizing Map

The SOM is a popular unsupervised data mining and visualization method, mapping a given possibly high-dimensional data set non linearly onto a low-dimensional regular lattice, which is topology-preserving under certain conditions [6]. In the SOM formulation by Heskes [4] training of a SOM is a stochastic gradient descent on a cost function. FLSOM as given in [8] is obtained by adding a classification error term to the original cost function of Heskes, such that classification labels are taken into account for model adaptation. One obtains a supervised classification scheme based on SOM, whereby the visualization and topographic properties are preserved. In our analysis we use the scaled Euclidean metric whose parameters are adapted during the optimization (metric adaptation) in accordance to [8], [11]. Hence, prototypes, labels and the metric are optimized by a stochastic gradient descent on the FLSOM cost function.

3 Visualization and analysis of proteomic data

Subsequently the proposed data processing scheme is applied to clinical ms spectra taken from a cancer study. Thereby we focus on a reliable encoding, visualization and classification model generation, indicating discriminative features.

3.1 Clinical sample preparation and MS data acquisition

Sample preparation and profile spectra analysis were carried out using the CLINPROT system². Plasma samples from 45 cancer patients and 50 controls were prepared using the MB-WCX Kit. Purifications were performed according to the product description. Sample preparation onto the Anchor Chip target was done using HCCA matrix and the spectra were generated using a linear autoflex.

3.2 Analysis with SOM and FLSOM

The preprocessed set of spectra and their corresponding peak areas or wavelet coefficients were now analyzed using the SOM and the FLSOM algorithm. For complex data sets a low (2-3) dimensional visualization is complicated to achieve and simple approaches as shown in [8] are often insufficient. To overcome this, SOMs offer very powerful visualization capabilities especially for high-dimensional data. However they are not considering label information and are also not applying the concept of relevance learning which is often necessary for biomarker search [11]. The presented FLSOM algorithm allows for both directions the search for markers incorporating potentially fuzzy label information and subsequent visualization by a fuzzy labeled SOM. The results are depicted in Figure (2,3)



Fig. 2. Visualizations of FLSOM (left), SOM (right) using bar plots. The plots are obtained using wavelet coefficients and show that the maps are well ordered with respect to the labeling. This is also the case for SOM - due to the classwise data similarity. Within the cells the first column denotes the possibility for cancer and the second for control. While the SOM has been crisp post labeled and contains multiple empty cells, the FLSOM shows the degree of class responsibility learned during the optimization. Overlapping data regions can be identified with respect to the classification task.

using the obtained maps for SOM and FLSOM with peak areas or wavelet coefficients. One finds that the obtained FLSOM is well ordered with respect to the two classes separating cancer from control. In addition by calculating the

² Devices and chemical processing by Bruker Daltonik GmbH, Bremen, Germany

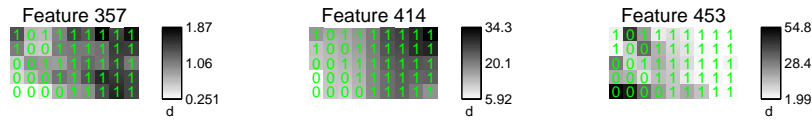
first term of the topographic function [10] we found that our map is topology preserving and hence neighborings on the map indeed corresponds to neighborings in the high-dimensional feature space. The class labels for each map cell are indicated by bar plots, whereby the first bar give a possibilistic measure for class 1 (cancer) and the second bar for control respectively. For SOM these values have been obtained by a post labeling showing a clear labeling of the cells with cancer data in the lower left and control data mapped to the upper right region. However one also observes a larger number of empty map cells. For FLSOM we found that the prototypes successfully learned a clear labeling from the data and hence the FLSOM map is well representing the underlying class information. By use of a principal component analysis we projected the obtained FLSOM together with the data into the space of the two principal components (c.f. Figure (3)). There we also get a clear separation of the classes and a reliable good spreading of the FLSOM into the data space. Thereby the prototypes of the FLSOM (huge o or \diamond at the corners of the lines) are well positioned into its corresponding classes. Considering the component planes for FLSOM some relevant input dimensions are identified which are separating the two classes. This is depicted in Figure (3). The feature 414 constitutes a small peak in the original data which has not been recognized by the standard peak picking procedure but which shows good separation capabilities. Comparing the component planes for peak features with those obtained for wavelet coefficients for SOM and FLSOM similarities and dissimilarities can be observed. Especially for the wavelet coefficients a larger number of relevant feature dimensions is observed. By analyzing their origin on the original mass axis we found that multiple neighbored wavelet coefficients encoded in fact the same peak region or related peaks³, but also new mass positions with small peaks have been found to be separating using the wavelet coefficients which were not detected using peak areas only.

The FLSOM model has been further evaluated in a 10-fold crossvalidation procedure. The models with peak areas as well as with wavelet coefficients were capable to discriminate between the two classes. For peak areas we found a recognition (training error) of 82% and a prediction (test error) of 71% whereas for wavelet coefficients the results were 78% and 75%. Hence the wavelet approach showed better generalization.

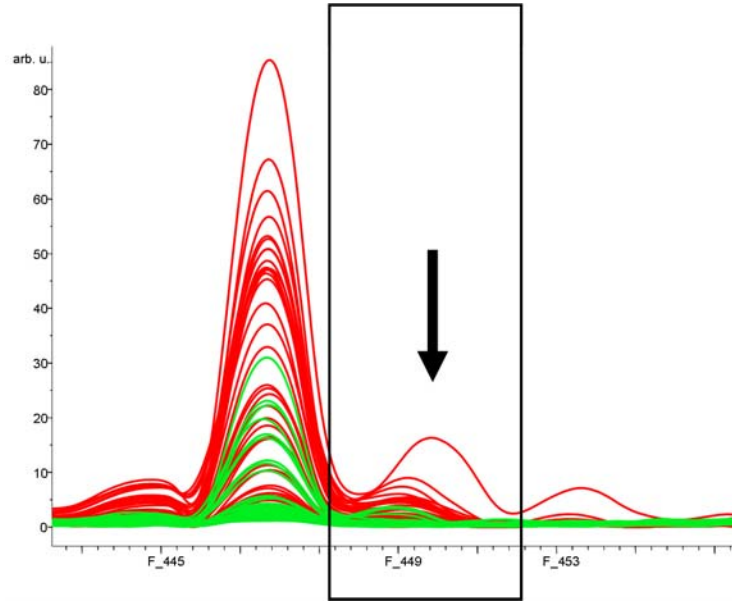
4 Conclusions

The presented initial data interpretation of proteom data demonstrate that the combined visualization and model generation of the FLSOM in combination with a wavelet based data preprocessing provides an easy and efficient detection of biomarker candidates and a very good visualization of the high-dimensional data. The usage of wavelet encoded spectra features is especially helpful in detection of small differences which maybe easily ignored by standard approaches. FLSOM was able to process high-dimensional wavelet features and observed good regularization, avoiding typical overfitting effects occurring by standard approaches. Moreover, the FLSOM visualization gives a planar representation of the high-dimensional data. If the obtained map is topological preserving one is able to identify sub-cluster of the original data space considering the receptive fields of the FLSOM cells. By use of the relevance learning we found a ranking

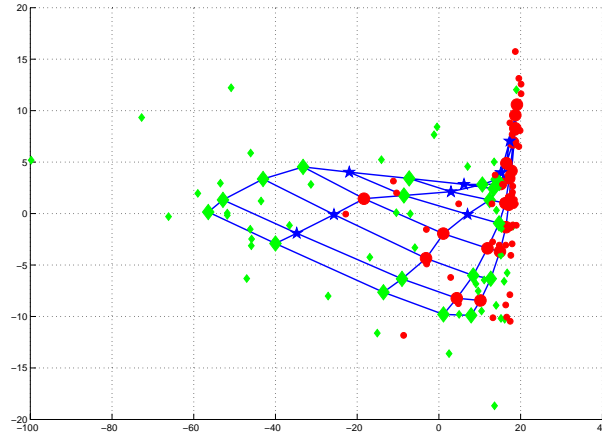
³ Related peaks maybe chemically related as encodings of multiple fragments.



(a) Component planes



(b) Relevant region



(c) PCA plot

Fig. 3. Visualizations of FLSOM component planes using three relevant wavelet coefficients identified by the obtained relevance profile. Map cells for cancer are labeled with 0 and for control with 1, respectively. The feature 357 has been observed to be relevant for SOM (visual inspection) as well as for FLSOM. The mass position encoded by feature 453 shows that also small peaks (shown in 3(b)) are still available by the DWT. In the PCA the data and the SOM are projected in PCA space using the first two principal components. Data for controls are depicted as ' \diamond ', its prototypes by huge ' \diamond ', cancer data plotted as ' o ' with huge o for prototypes, empty fields shown as '*'.

of the features related to mass positions in the original spectrum which allows for identification of most relevant feature dimensions which was used for the identification of biomarker candidates. In future analysis it could be interesting to reduce the feature set by incorporating knowledge about fractionation of the chemical compounds leading to multiple related peaks of the same source.

Potential biomarker peaks detected e.g. by FLSOM can be selected for in-depth analysis and analyzed by tandem ms (TOF/TOF) analysis. Conclusively, wavelet based spectra encoding in combination with FLSOM is an interesting alternative to standard approaches allowing more flexibility in problem modeling as well as the control of the data processing task. It combines efficient visualization with automated data pretreatment and intuitive analysis⁴.

References

1. A. Rieder A.K. Louis, P. Maaß. *Wavelets: Theory and Applications*. Wiley, 1998.
2. A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45(5):485–560, 1992.
3. Simon Haykin. *Neural Networks. A Comp. Found.* Macmillan, New York, 1994.
4. T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–316. Elsevier, Amsterdam, 1999.
5. R. Ketterlinus, S-Y. Hsieh, S-H. Teng, H. Lee, and W. Pusch. Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotocols software. *Bio techniques*, 38(6):37–40, 2005.
6. Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ext. Ed. 1997).
7. A. Leung, F. Chau, and J. Gao. A review on applications of wavelet transform techniques in chemical analysis: 1989-1997. *Chem. and Int. Lab. Sys.*, 43(1):165–184(20), 1998.
8. F.-M. Schleif, T. Elssner, M. Kostrzewa, Th. Villmann, and B. Hammer. Analysis and visualization of proteomic data by fuzzy labeled self organizing maps. In *Proc. of CBMS 2006*, pages 919–924, 2006.
9. J. Villanueva, J. Philip, D. Entenberg, and C.A. Chaparro et al. Serum peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry. *Anal. Chem.*, 76:1560–1570, 2004.
10. 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.
11. Th. Villmann, F.-M. Schleif, and B. Hammer. Comparison of relevance learning vector quantization with other metric adaptive classification methods. *Neural Networks*, 19(15):610–622, 2005.
12. D.E. Waagen, M.L. Cassabaum, C. Scott, and H.A. Schmitt. Exploring alternative wavelet base selection techniques with application to high resolution radar classification. In *Proc. of the 6th Int. Conf. on Inf. Fusion (ISIF'03)*, pages 1078–1085. IEEE Press, 2003.

⁴ **ACKNOWLEDGMENT:** The authors are grateful to M. Kostrzewa and T. Elssner for providing the clinical proteom data (both Bruker Daltonik Leipzig, Germany)