# Cancer Informatics by prototype networks in mass spectrometry

Frank-Michael Schleif ,

*University Leipzig, Dept. of Medicine, 04107 Leipzig, Germany*

Thomas Villmann [a]

*University Leipzig, Dept. of Medicine, 04107 Leipzig, Germany*

Markus Kostrzewa [b]

*Bruker Daltonik GmbH, Dept. of Bioanalytics, R & D, 04318 Leipzig, Germany*

Barbara Hammer [c]

*Technical University of Clausthal, Dept. of Computer Science, Computational Intelligence Group, Julius-Albert-Str. 4, 38678 Clausthal-Zellerfeld, Germany*

Alexander Gammerman [d]

*Royal Holloway University College London, Dept. of Computer Science, Computational Research Learning Center, London, UK*

**Abstract**

Mass spectrometry has become a standard technique to analyse clinical samples in cancer research. The obtained spectrometric measurements reveal a lot of information of the clinical sample at the peptide and protein level. The spectra are high dimensional and, due to the small number of samples a sparse coverage of the population is very common. In clinical research the calculation and evaluation of classification models is important. For classical statistics this is achieved by hypothesis testing with respect to a chosen level of confidence. In clinical proteomics the application of statistical tests is limited due to the small number of samples and the high dimensionality of the data. Typically soft methods from the field of machine learning like prototype based vector quantizers [17], Support Vector Machines(SVM) [32], Self-Organizing Maps (SOMs) [17] and respective variants are used to generate such models. However for these methods the classification decision is crisp in general and no or only few additional information about the safety of the decision is available.

*Preprint submitted to Elsevier* *11 June 2008*

In this contribution the spectral data are processed as functional data by a wavelet based preprocessing [29] employing a functional metric [30,28] in the prototype based classifiers. In particular, we demonstrate applications of the weighted Euclidean metric and the weighted functional norm (based on weighted $L^p$-norm) taking the specific nature of mass-spectra into account. This also allows the detection of potential biomarker candidates. To judge the classification decisions and model accuracy we focus on a method for the estimation of confidence using prototype based networks.

We demonstrate the usefulness of the above extensions in the analysis of mass spectra in proteomics and related knowledge discovery. In particular, we give application examples for biomarker detection based on feature selection and classification of spectra.

*Key words:* clinical proteomics, cancer informatics, mass spectrometry, prototype classifiers, confidence estimation

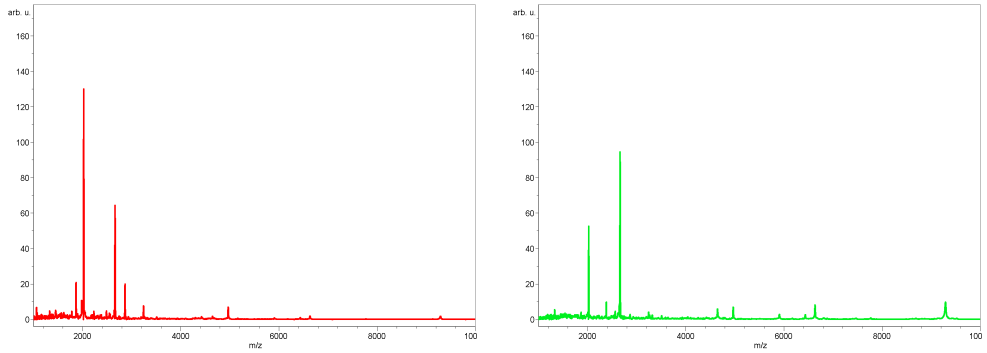
---

## 1 Introduction

Analysis of clinical proteomic spectra obtained from mass spectrometric measurements is a complicated issue [22]. One major objective is the search for potential biomarkers in complex body fluids like serum, plasma, urine, saliva, or cerebral spinal fluid [6,24,25,10]. Typically the spectra are given as high-dimensional vectors. Thus, from a mathematical point of view, an efficient analysis and visualization of high-dimensional data sets is required. Moreover, the amount of available data is restricted: usually patient cohorts are small in comparison to the dimensionality of the data.

In contrast to the widely applied multilayer perceptron [2], prototype based classification allows an easy interpretation, which is of particular interest for many (clinical) applications. One prominent prototype based classifier is the Supervised Relevance Neural Gas algorithm (SRNG)[12]. SRNG leads to a robust classifier where efficient learning of labeled high dimensional data is possible and has been already used in different types experiments [37,27,38,34].

*Email addresses:* schleif@informatik.uni-leipzig.de,+49(0)3419718955|, http://www.uni-leipzig.de/∼compint (Frank-Michael Schleif), villmann@informatik.uni-leipzig.de,+49(0)3419718868|, http://www.uni-leipzig.de/∼compint (Thomas Villmann), km@bdal.de|, http://www.bdal.de (Markus Kostrzewa), hammer@in.tu-clausthal.de|, http://cig.in.tu-clausthal.de (Barbara Hammer), alex@cs.rhul.ac.uk| (Alexander Gammerman).




(a) Cancer Spectrum (b) Control Spectrum

Fig. 1. (a) MALDI-TOF spectrum of a colorectal cancer patient and (b) a healthy subject after peptide isolation with C8 magnetic beads. On the Y-axis the relative intensity is shown. The mass to charge ratio (m/z) is demonstrated on the X-axis in Dalton. The spectra are already preprocessed (baseline correction,recalibration) using ClinProTools 2.1

In general the available approaches to model classifiers in clinical proteomics initially transform the spectra into a vector space followed by training a classifier. In this way the functional nature of the data is lost, which may lead to suboptimal classifier models. A functional representation of the data with respect to the used metric and a weighting or pruning of (priorly not known) irrelevant parts of the inputs, would be desirable. A discriminative data representation is necessary. The extraction of such discriminant features is difficult for spectral data and typically done by a parametric peak picking procedure. This peak picking is often the focus of criticism because some present peaks may not be detected and the functional nature of the data is partially lost. To avoid this difficulties we focus on the approach as given in [30,28] and apply a wavelet encoding to the spectral data to get discriminative features. The obtained wavelet coefficients are sufficient to reconstruct the signal, still containing all relevant information of the spectra in a *functional* encoding. However this better discriminating set of features is typically more complex and hence a robust approach to determine the desired classification model is needed. Taking this into account a feature selection is applied based on a statistical pre-analysis of the data and the SRNG algorithm is used to obtained predictive models.

In this contribution, we focus on the conformal prediction concept incorporated in prototype based learning vector quantizers (LVQ). The paper is organized as follows. First we briefly review the functional encoding of mass spectrometric data by means of a wavelet based encoding. Subsequently the theory of the Supervised Relevance Neural Gas (SRNG) and its equipment with a functional metric is reviewed. After these settings, the method of conformal prediction [39,9] is reviewed and we show how it can be used together with LVQ approaches. Subsequently the methodology is applied on experimen-

tal data from two clinical proteom studies. We evaluate the results not only using cross validation but also in the light of conformal prediction which allows the assessment of the classification safety by means of p-values as known from classical statistics.

## 2 Preprocessing

The classification of mass spectra involves multiple preprocessing steps. In general peak picking is used to locate and quantify positions of peaks within the spectrum and feature extraction is applied on the peak list to obtained an adequate feature matrix. In the first step a number of procedures as baseline correction, optional denoising, noise estimation and normalization are needed[16,26]. 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 baseline correction and recalibration (alignment) of multiple spectra is standard, and has been done using ClinProTools (details in [16]) [1] . Here we propose an alternative feature extraction procedure preserving 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-Toolbox [2] , due to the local analysis property of wavelet analysis the features can still 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. The test was used to identify features which show a significant ($p < 0.01$) discrimination between the two groups (cancer,control). To get valid results a p-value adjustment by means of the bonferroni-correction has been applied as well. This is done in accordance to [40] where also a generation to a multiclass experiment is given.

### *2.1 Feature Extraction by Bi-orthogonal Discrete Wavelet Transform*

Wavelets have been developed as powerful tools [1,19] 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

[1] Biomarker software available at http://www.bdal.de
[2] The Matlab Wavelet-Toolbox can be obtained from www.mathworks.com

reason one can apply the so called bi-orthogonal wavelet transform[3] 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.

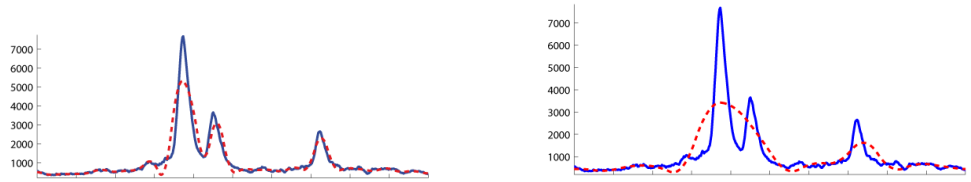
In our analysis such a smooth synthesis pair was chosen to avoid artifacts. It can be expected that a signal in the time domain can be represented by a small number of a relatively large set of coefficients from the wavelet domain. The spectra are reconstructed in dependence of a certain approximation 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 2. 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 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 [3]. The approximation coefficients describe a generalized peak list of the denoised spectrum encoding primal spectral information and depend on the level $L$ which is determined with respect to the measurement procedure. For linearly MALDI-TOF spectra a device resolution of $500 - 800Da$ 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 typically sufficient for a linear measured spectrum with $\approx 20000$ measurement points (see Figure 2). The level $L$ can be automatically determined by considering expected peak width in $Da$ and the reconstruction capabilities of wavelet analysis at a given level. Alternatively multiple levels can be tried and a standard peak picking approach can be applied on both, the original and the reconstructed spectrum. If the obtained peak lists are sufficiently similar, which means, that at least peaks with good S/N values in the original spectrum are sufficiently recovered in the reconstruction the taken level can be considered as acceptable for the experiment.

Applying this procedure including the KS-test on the spectra with an initial number of $\approx$ 4000 measurement points in a range of $1500 - 3500Da$ per spectrum one obtains 416 wavelet coefficients used as representative features per spectrum, still allowing a reliable functional representation of the data. An application of the KS-Test still keeps 101 coefficients for the final analysis of the colorectal cancer patients (CRC) data set and 40 coefficients for the




(a) Wavelet reconstruction $L = 4$ (b) Wavelet reconstruction$L = 5$

Fig. 2. Wavelet reconstruction of the spectra with $L = 4, 5$, $x$ measurement positions, $y$-arbitrary unit. The original signal is plotted with the interrupted line (blue) and the reconstruction with the solid with a white band inside. One observes that a wavelet analysis with $L = 5$ is too rough to approximate the sharp peaks.

lung cancer (LC) data set [3] .

## 3 Bioinformatic methods

The Supervised Relevance Neural Gas (SRNG) algorithm is a prototype based classification model, which will be introduced very briefly. Subsequently we extend the concept of conformal prediction as introduced in [39,9] in the context of prototype based networks which is used in the evaluation part to determine confidence values for obtained classification results.

### *3.1 Supervised Relevance Neural Gas with generalized metrics*

Supervised Neural Gas (SNG) is considered as a representative for prototype based classification approaches as introduced by KOHONEN. Different prototype classifiers have been proposed so far [17,23,14,36] as improvements of the original approach. The SNG has been introduced in [36] and combines ideas from the Neural Gas algorithm (NG) introduced in [20] with the Generalized learning vector quantizer (GLVQ) as given in [23]. Subsequently we give the basic notations and some remarks to the integration of alternative metrics into Supervised Neural Gas (SNG). Details on SNG including convergence proofs can be found in [36].

Let us first clarify some notations: Let $c_{\mathbf{v}} \in \mathcal{L}$ be the label of input $\mathbf{v}$, $\mathcal{L}$ a set of labels (classes) with $\#\mathcal{L} = N_{\mathcal{L}}$. Let $V \subseteq \mathbb{R}^{D_V}$ be a finite set of inputs $\mathbf{v}$. LVQ uses a fixed number of prototypes (weight vectors, codebook vectors) for each class. Let $\mathbf{W} = \{\mathbf{w_r}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class

[3] The ks-test is an optional data reduction step, the removed dimensions are in general neighbored, closed stripes of noise and not discriminating signals

label of $\mathbf{w_r}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w_r} | c_\mathbf{r} = c\}$ be the subset of prototypes assigned to class $c \in \mathcal{L}$.

The task of vector quantization is realized by the map $\Psi$ as a winner-take-all rule, i.e. a stimulus vector $\mathbf{v} \in V$ is mapped onto the closest $\mathbf{w}_s$,

$$\Psi^\lambda_{V \to \mathcal{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \text{argmin}_{\mathbf{r} \in A} d^\lambda(\mathbf{v}, \mathbf{w_r}) \tag{1}$$

with $d^\lambda(\mathbf{v}, \mathbf{w})$ being an arbitrary differentiable distance measure [4] which may depend on a parameter vector $\lambda$ and $A$ a (ordered) grid of neurons. Subsequently we only expect that the used distance measure is differentiable with respect to its parameters. For the moment we take $\lambda$ as fixed. The neuron $\mathbf{s}(\mathbf{v})$ is called winner or best matching unit. The subset of the input space

$$\Omega^\lambda_\mathbf{r} = \{\mathbf{v} \in V : \mathbf{r} = \Psi_{V \to A}(\mathbf{v})\} \tag{2}$$

which is mapped to a particular neuron $\mathbf{r}$ according to (1), forms the (masked) receptive field of that neuron forming a Voronoi tessellation. If the class information of the weight vector is used, the boundaries $\partial\Omega^\lambda_\mathbf{r}$ generate the decision boundaries for classes. A training algorithm should adapt the prototypes such that for each class $c \in \mathcal{L}$, the corresponding codebook vectors $\mathbf{W}_c$ represent the class as accurately as possible. This means that the set of points in any given class $V_c = \{\mathbf{v} \in V | c_\mathbf{v} = c\}$, and the union $\mathcal{U}_c = \bigcup_{\mathbf{r}|_{\mathbf{w_r} \in \mathbf{W}_c}} \Omega_\mathbf{r}$ of receptive fields of the corresponding prototypes should differ as little as possible.

Supervised Neural Gas (SNG) constitutes a method to train prototypes efficiently according to given data points. Again, let $\mathbf{W}_c = \{\mathbf{w_r} | c_\mathbf{r} = c\}$ be the subset of prototypes assigned to class $c \in \mathcal{L}$ and $K_c$ its cardinality.

Further we assume to have $m$ data vectors $\mathbf{v}_i$. As pointed out in [36], neighborhood learning for a given input $\mathbf{v}_i$ with label $c$ is applied to the subset $\mathbf{W}_c$. The respective cost function is

$$Cost_{SNG}(\gamma) = \sum_{i=1}^{m} \sum_{\mathbf{r}|\mathbf{w_r} \in \mathbf{W}_{c_i}} \frac{h_\gamma(\mathbf{r}, \mathbf{v}_i, \mathbf{W}_{c_i}) \cdot f(\mu_\lambda(\mathbf{r}, \mathbf{v}))}{C(\gamma, K_{c_i})} \tag{3}$$

with $f(x) = (1 + \exp(-x))^{-1}$, $h_\gamma(\mathbf{r}, \mathbf{v}, \mathbf{W}) = \exp\left(-\frac{k_\mathbf{r}(\mathbf{v}, \mathbf{W})}{\gamma}\right)$ and $\mu_\lambda(\mathbf{r}, \mathbf{v}) = \frac{d^\lambda_\mathbf{r} - d^\lambda_{\mathbf{r}_-}}{d^\lambda_\mathbf{r} + d^\lambda_{\mathbf{r}_-}}$ and $d^\lambda_{\mathbf{r}_-}$ is defined as the squared distance to the best matching pro-

[4] A distance measure is a non-negative real-valued function, which, in contrast to a metric does not necessarily fulfill the triangle inequality and the symmetry property. For prototype algorithms of the mentioned type the used distance measure need not to be a metric. A detailed discussion of this fact with respect to the considered methods is available in [11,13]

totype but labeled with $c_{\mathbf{r}_-} \neq c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_-}$ and $d^{\lambda}_{\mathbf{r}} = d^{\lambda}(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$. Details on the corresponding update rules are given in [36].

### *3.1.1 Incorporation of a functional metric to SNG*

As pointed out before, the distance measure $d^{\lambda}(\mathbf{v}, \mathbf{w})$ is only required to be differentiable with respect to $\lambda$ and $\mathbf{w}$. The triangle inequality has not to be fulfilled necessarily. This leads to a great freedom in the choice of suitable measures and allows the usage of non-standard metrics in a natural way. We now review the functional metric as given in [18], the obtained derivations can be plugged into the above equations leading to SNG with a functional metric, the data are functions represented by vectors and, hence, the vector dimensions are spatially correlated.

Common vector processing does not take the spatial order of the coordinates into account. As a consequence, the functional aspect of spectral data is lost. For proteom spectra the order of signal features (peaks) is due to the nature of the underlying biological samples and the measurement procedure. The masses of measured chemical compounds are given ascending and peaks encoding chemical structures with a higher mass follow chemical structures with lower masses. In addition multiple peaks with different masses may encode parts of the same chemical structure and hence are correlated.

In [18] a distance measure has been proposed taking the functional structure of the data into account, involving the previous and next values of $v_i$ in the $i$-th term of the sum, instead of $v_i$ alone. $V$ can be represented as $V = (v_1, \ldots, v_D)$. Assuming a constant sampling period $\tau$, the proposed norm is:

$$\mathcal{L}_p^{fc}(\mathbf{v}) = \left( \sum_{k=1}^{D} \left( A_k(\mathbf{v}) + B_k(\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{4}$$

with

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \tag{5}$$

$$B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{6}$$

are respectively the triangles on the left and right hand sides $v_i$. Just as for $L_p$, the value of $p$ is assumed to be a positive integer. At the left and right ends of the sequence, $v_0$ and $v_D$ are assumed to be equal to zero. The derivatives for the functional metric taking $p = 2$ are given in [18].

Now we consider the scaled functional norm where each dimension $v_i$ is scaled by a parameter $\lambda_i \geq 0$ $\lambda_i \in (0,1]$ and $\sum_i \lambda_i = 1$. Then the scaled functional norm is:

$$\mathcal{L}_p^{fc}(\lambda \mathbf{v}) = \left( \sum_{k=1}^{D} \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{7}$$

with

$$A_k(\lambda \mathbf{v}) = \begin{cases} \frac{\tau}{2} \lambda_k |v_k| \text{ if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k |v_k| + \lambda_{k-1} |v_{k-1}|} \text{ else} \end{cases} \quad B_k(\lambda \mathbf{v}) = \begin{cases} \frac{\tau}{2} \lambda_k |v_k| \text{ if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k |v_k| + \lambda_{k+1} |v_{k+1}|} \text{else} \end{cases} \tag{8}$$

The prototype update changes to:

$$\frac{\partial \delta_2^2(\mathbf{x}, \mathbf{y}, \lambda)}{\partial x_k} = \frac{\tau^2}{2} (2 - U_{k-1} - U_{k+1}) (V_{k-1} + V_{k+1}) \triangle_k \tag{9}$$

with

$$U_{k-1} = \begin{cases} 0 \text{ if } 0 \leq \triangle_k \triangle_{k-1} \\ \left( \frac{\lambda_{k-1} \triangle_{k-1}}{\lambda_k |\triangle_k| + \lambda_{k-1} |\triangle_{k-1}|} \right)^2 \text{ else} \end{cases} , U_{k+1} = \begin{cases} 0 \text{ if } 0 \leq \triangle_k \triangle_{k+1} \\ \left( \frac{\lambda_{k+1} \triangle_{k+1}}{\lambda_k |\triangle_k| + \lambda_{k+1} |\triangle_{k+1}|} \right)^2 \text{ else} \end{cases}$$

$$V_{k-1} = \begin{cases} 1 \lambda_k \text{ if } 0 \leq \triangle_k \triangle_{k-1} \\ \frac{\lambda_k |\triangle_k|}{\lambda_k |\triangle_k| + \lambda_{k-1} |\triangle_{k-1}|} \text{ else} \end{cases} , V_{k+1} = \begin{cases} 1 \lambda_k \text{ if } 0 \leq \triangle_k \triangle_{k+1} \\ \frac{\lambda_k |\triangle_k|}{\lambda_k |\triangle_k| + \lambda_{k+1} |\triangle_{k+1}|} \text{ else} \end{cases}$$

and $\triangle_k = x_k - y_k$ For the $\lambda$-update one observes:

$$\begin{aligned} \frac{\partial \mathcal{L}_p^{fc}(\lambda \mathbf{v})}{\partial \lambda_k} &= \frac{\partial \left( \sum_{k=1}^{D} \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p \right)^{\frac{1}{p}}}{\partial \lambda_k} \\ &= p \left( \sum_{k=1}^{D} \left( A_{k-1}(\lambda \mathbf{v}) + A_{k+1}(\lambda \mathbf{v}) \right)^p \right)^{\frac{1-p}{p}} \frac{\partial \left[ \sum_{k=1}^{D} \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p \right]}{\partial \lambda_k} \\ &= C_p \frac{\partial \left[ \sum_{k=1}^{D} \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p \right]}{\partial \lambda_k} \\ &= C_p \frac{\sum_{k=1}^{D} \partial \left[ \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p \right]}{\partial \lambda_k} \\ &= C_p \frac{\partial \left[ \left( A_{k-1}(\lambda \mathbf{v}) + B_{k-1}(\lambda \mathbf{v}) \right)^p + \left( A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right)^p + \left( A_{k+1}(\lambda \mathbf{v}) + B_{k+1}(\lambda \mathbf{v}) \right)^p \right]}{\partial \lambda_k} \\ &= C_p \left( c_p^{k-1} \frac{\partial \left[ A_{k-1}(\lambda \mathbf{v}) + B_{k-1}(\lambda \mathbf{v}) \right]}{\partial \lambda_k} + c_p^k \frac{\partial \left[ A_k(\lambda \mathbf{v}) + B_k(\lambda \mathbf{v}) \right]}{\partial \lambda_k} + c_p^{k+1} \frac{\partial \left[ A_{k+1}(\lambda \mathbf{v}) + B_{k+1}(\lambda \mathbf{v}) \right]}{\partial \lambda_k} \right) \end{aligned}$$

with the following expressions

$$
\begin{aligned}
c_p^j &= p \cdot \left(A_j\left(\lambda \mathbf{v}\right) + B_j\left(\lambda \mathbf{v}\right)\right)^{p-1} \\
&= p \cdot \left( \begin{cases} \frac{\tau}{2}\lambda_j \left|v_j\right| & \text{if } 0 \leq v_j v_{j-1} \\ \frac{\tau}{2}\frac{\lambda_j^2 v_j^2}{\lambda_j |v_j| + \lambda_{j-1}|v_{j-1}|} & \text{if } 0 > v_j v_{j-1} \end{cases} + \begin{cases} \frac{\tau}{2}\lambda_j \left|v_j\right| & \text{if } 0 \leq v_j v_{j+1} \\ \frac{\tau}{2}\frac{\lambda_j^2 v_j^2}{\lambda_j |v_j| + \lambda_{j+1}|v_{j+1}|} & \text{if } 0 > v_j v_{j+1} \end{cases} \right)^{p-1}
\end{aligned}
$$

putting all together and with some minor mathematical transformations one obtains:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_p^{fc}\left(\lambda \mathbf{v}\right)}{\partial \lambda_k} = C_p & \begin{cases} 0 + c_p^k \left(\frac{\tau}{2}\left|v_k\right|\right) & \text{if } 0 \leq v_{k-1} v_k \\ \frac{1}{2}\tau \frac{\lambda_k^2 c_p^k v_k^2 |v_k| - c_p^{k-1}|v_k| v_{k-1}^2 \lambda_{k-1}^2 + 2\lambda_k c_p^k v_k^2 \left|v_{k-1}\right| \lambda_{k-1}}{\left(\lambda_k |v_k| + \left|v_{k-1}\right| \lambda_{k-1}\right)^2} & \text{if } 0 > v_{k-1} v_k \end{cases} \\
+ C_p & \begin{cases} c_p^k \left(\frac{\tau}{2}\left|v_k\right|\right) + 0 & \text{if } 0 \leq v_{k+1} v_k \\ \frac{1}{2}\tau \frac{\lambda_k^2 c_p^k v_k^2 |v_k| - c_p^{k+1}|v_k| v_{k+1}^2 \lambda_{k+1}^2 + 2\lambda_k c_p^k v_k^2 \left|v_{k+1}\right| \lambda_{k+1}}{\left(\lambda_k |v_k| + \left|v_{k+1}\right| \lambda_{k+1}\right)^2} & \text{if } 0 > v_{k+1} v_k \end{cases}
\end{aligned}
$$

Using this parametrization one can emphasize/neglect different parts of the function for classification. This distance measure can be put into SNG as shown above and has been applied subsequently in the analysis of clinical proteom spectra. SNG with metric adaptation is subsequently referred as SRNG.

## 4 Evaluation of Prototype based classifier models

Advanced prototype based classification models show typically high regularisation capabilities [11]. Nevertheless also the results of prototype networks need a thoroughly analysis by cross validation to get practical measures to rate the prediction capabilities of the current model. Beside these generic measures of confidence in the results obtained by a classification model a more fine grained confidence analysis would be desirable. Classical statistics typically allows a judgment on the classification accuracy of a single item by means of p-values [15] but are not applicable (in a valid sense) for these type of data, in general. Also Gaussian Mixture Models allow to determine the probability of a classification decision, but make additional constraints on the considered type of data [15]. This techniques are well understood but in general not available for soft methods like SVM or prototype networks. Only few attempts were made to give reliability estimate for these soft methods (see e.g. [4,5]). Thereby the reliability estimate can be helpful to judge on the reliability of a decision but also in a more generic framework to improve the overall performance of the classifier. Reliability sometimes also referred as confidence, has been subject of a quite new theory called conformal prediction as introduced in [39] which fills this gap under some moderate constraints. Here we show how the concept of conformal prediction can be applied to prototype networks and allows the

determination of statistical significance values as needed in clinical studies and cancer informatics.

### *4.1 Conformal Prediction for Prototype based Networks*

Conformal predictors aim at the estimation of confidence of a given classification decision. They remain automatically valid (in average) under the randomness assumption [39,9]. It is assumed, that the objects and their labels are generated independently from the same probability distribution. This appears to be a strong assumption but in fact it is a much weaker assumption than assuming a parametric statistical model. Conformal predictors never overrate the accuracy and reliability of their predictions [39,9]. When the stochastic mechanism significantly deviates from the model, conformal predictors remain valid but their efficiency inevitably suffers [39,9]. As conformal predictors are provably valid, efficiency with respect to computational performance as well as with respect to the effort to extend a classifier to a conformal predictor, are the only things which we need to worry about. First we will give some basic notations and review the main concepts of conformal prediction as given in [39,9].

#### *4.1.1 Conformal prediction a brief overview*

We now briefly review the concepts of conformal prediction as presented [7] and the tutorial given in [31]. The basics of conformal prediction rely on confidence intervals from classical statistics and are well theoretically founded [8]. Here we focus on classification and deal with labeled data. The task is: predict each label after seeing its object:

- from $x_1$ predict $y_1$
- from $(x_1, y_1), x_2$, predict $y_2$
- from $(x_1, y_1), (x_2, y_2), x_3$ predict $y_3$ and so on

Here we assume *randomness*. In reality we choose the examples independently from some probability distribution $\mathbb{Q}$ on $\mathbb{Z} = \mathbb{D} \times \mathbb{Y}$. The samples are independent and identically distributed. And we do not make any assumptions about $\mathbb{Q}$. Usually independence can be weakened to exchangeability [31]. To do the prediction with confidence we write $\mathbb{Z}^*$ for the set of all finite sequences of elements of $\mathbb{Z}$ such that:

$$\mathbb{Z}^* = \cup_{n=0}^{\infty} Z^n$$

A level $(1-\epsilon)$ confidence predictor is a mapping

$$\Gamma : \mathbb{Z}^* \times \mathbb{D} \to 2^{\mathbb{Y}}$$

after observing old examples $z_1, \ldots, z_{n-1}$ and the new object $x_n$, we predict that the label of $(x_n, y_n)$ will be in the subset

$$\Gamma(z_1, \ldots, z_{n-1}, x_n)$$

of the label space $Y$. A $(1-\epsilon)$ confidence predictor is exactly valid if its hits are independent and all happen with probability $(1-\epsilon)$. It is conservatively valid if the probability that the predictions on rounds $n_1, \ldots, n_k$ are all hits is always at least $(1-\epsilon)^k$. This does not depend on a specific probability function. Valid confidence predictors are constructed from nonconformity measures by means of real values functions A: $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x, y)$ as a measure of how different $(x, y)$ is from $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. Here one predicts values of $y_n$ that make $x_n, y_n$ differ minimally from the rest. From a given nonconformity measure, we construct a $(1-\epsilon)$ confidence predictor $\Gamma^\epsilon$ for every $\epsilon \in [0, 1]$, and they are nested in the natural way: $\Gamma^{\epsilon_1}(z_1, \ldots, z_n, x_n) \subseteq \Gamma^{\epsilon_2}(z_1, \ldots, z_n, x_n)$ when $\epsilon_1 \geq \epsilon_2$. The more confident one wants to be, the larger the region must be chosen. So e.g. $\Gamma^{0.05}$ the prediction region, is a set that contains the true labeling with a probability of at least 95%. Typically $\Gamma^{0.05}$ also contains the prediction $\hat{y}$. We call $\hat{y}$ the point prediction. In case of classification $\Gamma^{0.05}$ may consist of a few of these values or, in the best case, just one [31]. Given a nonconformity measure, the conformal prediction algorithm produces a prediction region $\Gamma^\epsilon$ for every probability of $\epsilon$. The region for $\Gamma^\epsilon$ is a $1-\epsilon$ prediction interval which contains $\hat{y}$. with a probability of at least $1-\epsilon$. The regions for different $\epsilon$ are nested: when $\epsilon_1 \leq \epsilon_2$: so that $1-\epsilon_1$ is a lower level of confidence than $1-\epsilon_2$ , we have $\Gamma^{\epsilon_1} \subset \Gamma^{\epsilon_2}$. If $\Gamma^\epsilon$ consists of only one entry (label) we may ask our self how small $\epsilon$ can be made until the cardinality changes, the obtained $1-\epsilon$ is the level of confidence.

To summarize these points, the most useful prediction is those containing exactly one label. Therefore two error rates are of particular interest, $\epsilon_1$ being the smallest $\epsilon$ and $\epsilon_2$ being the greatest $\epsilon$ so that $|\Gamma^\epsilon| = 1$. $\epsilon_2$ is the p-value of the best and $\epsilon_1$ is the p-value of the second best label $y$. So the prediction can be summarized as

$$(\text{confidence}) = 1 - \epsilon_1 = 1 - p_{y_{2\text{nd}}} \tag{10}$$
$$(\text{credibility}) = \epsilon_2 = p_{y_{1\text{st}}} \tag{11}$$

Confidence says something about being sure that the second best label and all worse ones are wrong. Credibility says something about to be sure that the best label is right respectively that the data point is (un)typical and not an outlier.

As further pointed out in [39,9] there are two approaches to construct conformal predictors by means of inductive or transductive learners, here we focus on transductive learners (for details see [39,9]). While the just sketched theoretical framework of conformal prediction is a generic statistical approach, the concrete utilization needs a so called nonconformity measure which is individual for each type of algorithm.

**Definition 1 (Nonconformity measure)** *A nonconformity measure is a function $A : B \times Z \to \mathcal{R}$ With $B$ as the set of all finite bags of elements in $Z$.*

In practical applications $A$ is chosen such that large values of $A(B, z)$ indicate that $z$ is strange relative to $B$. As an example for classification suppose $\mathbb{D} = \mathcal{R}^k$ and $Y$ finite. Then, a useful nonconformity measure is:

$$A(z_1, \ldots, z_n, z) = \frac{\min_{i:y_i=y} d(x_i, y)}{\min_{i:y_i \neq y} d(x_i, y)}$$

where $d$ refers to an arbitrary distance measure. A 95% confidence region for $y_n$ is constructed by a nonconformity measure $A$, old examples $z_1, \ldots, z_{n-1}$ and a new object $x_n$. The procedure can be summarized as follows. We consider each labeling $y \in Y$ with $B$ a bag consisting of $z_1, \ldots, z_{n-1}$ together with $x_n, y$. Let now $B^{-i}$ the bag obtained by removing $z_i$, further define $W_i = A(B^{-i}, z_i)$ with $i = 1, \ldots, n$ and set:

$$p_y = \frac{\#i = 1, \ldots, n | W_i \geq W_n}{n}$$

Then $p_y$ is the $p$-value for the current labeling $y$. It is the fraction of the elements in $B$ that are at least as strange relative to the others as $(x_n, y)$. Finally we include $y$ in the confidence region if and only if $p_y > 0.05$.

### *4.1.2 Conformal prediction with prototype based classifiers*

GLVQ and variants are successful prototype based learning algorithms with a winner rule in accordance or similar to the Eq. 1 used in the corresponding cost function. Multiple variants of this scheme have been presented but their common property is the existence of the distances $d^+$ and $d^-$ (closest winner with the same (+) labeling or closest prototype with a different label (−)) used in the cost function to optimize the prototype positions. To transform GLVQ variants into conformal predictors a nonconformity measure has to be determined which is of the form of Def. 1

For prototype based networks one natural measure of non-conformity ($C(\mathbf{v_i}, c_i)$ for a given sample $\mathbf{v_i}$ and a given (crisp) labeling $c_i$ is the sample margin as

the distance of the data point to the closest prototype with the same label (+) normalized by the distance of this item to the closest prototype with an alternative labeling (−):

$$C(\mathbf{v_i}, c_i) = d^{+}_{\min,\lambda}(\mathbf{w_r}, \mathbf{v_i}) / d^{-}_{\min,\lambda}(\mathbf{w_r}, \mathbf{v_i}) = d^{+}_{\min,\lambda}(\mathbf{x_r}, \mathbf{y_i}) / d^{-}_{\min,\lambda}(\mathbf{x_r}, \mathbf{y_i}) \quad (12)$$

Here, $\lambda$ is some parametrization of the underlying distance measure $d$ and the classifier decision is considered to be safe if the obtained non-conformity score is small - by means of a small distance of the datapoint to its closest prototype with the same labeling.

#### *4.1.3 Confidence estimates within clinical studies*

Conformal predictors require the definition of a valid nonconformity measure of the used modeling approach. In the former section such as measures have been presented for GLVQ networks which is applicable for SRNG as well. The estimation of confidence and credibility based on conformal prediction can be done by either induction or transduction. While the former is very common it has the drawback, that multiple splits in the data into hold-out-subsets are necessary. The transductive method avoids additional splits but is computationally expensive if the number of samples or the number of labels becomes (very) large. In clinical proteomics the number of samples is typically small, in general around $50 - 500$ samples per class with a number of classes below 10. Hence a transductive approach is still applicable, avoiding unnecessary splittings of the data while keeping computations reliably effective. The number $n$ used in the modeling should, however not become to small. Otherwise the validity of the conformal prediction will be decreased, or more precise the confidence bounds getting worse.

## 5 Clinical Data

Serum protein profiling is a promising approach for classification of cancer versus non-cancer samples. The data used in this paper are taken from a colorectal cancer (CRC) study and patients from healthy individuals [5]. Here it should be mentioned only that for each profile a mass spectrum is obtained within an analyzed mass-to-charge-ratio of 1500 to 3500Da. Two sample spectra are depicted in Figure 1. The data have been preprocessed as explained before

[5] Details about the data source can be obtained via Bruker Daltonik GmbH, 04109, Leipzig, Deutscher Platz 5d, Germany (km@bdal.de)

using the approach published in [28]. The spectra are encoded by 416 wavelet-coefficients which leads to a data reduction of $\approx 95\%$ using the rawdata and is approximately twice the range of the number of peaks as obtained by the standard peak picking approach as proposed in [16] The preprocessing step has to be included in the crossvalidation procedure to avoid overfitting. For the considered data set it could be observed that the discriminating wavelet coefficients (with respect to the ks-test) at $p \leq 0.01$ including a p-value adjustment in accordance to bonferroni, reduce further to 101 (CRC) or 40 (LC) significant coefficients in a $5-$fold double cross validation. The wavelet method was used as mentioned in the previous section with $L = 4$.

The data set consist of 100 - colorectal cancer (CRC) and 90 - lung cancer (LC) data points. For the colorectal cancer and lung cancer study, 50 samples are taken from patients suffering from colorectal or lung cancer and the remaining samples are taken from a matched healthy control group. Colorectal cancer (CRC) is among the most common malignancies and remains a leading cause of cancer-related morbidity and mortality. It is well recognized that CRC arises from a multistep sequence of genetic alterations that result in the transformation of normal mucosa to a precursor abdomen and ultimately to carcinoma. Given the natural history of CRC, early diagnosis appears to be the most appropriate tool to reduce disease-related mortality. Currently, there is no early diagnostic test with sufficient diagnostic quality, which can be used as a routine screening tool. Therefore, there is a need for new biomarkers for colorectal cancer that can improve early diagnosis, monitoring of disease progression and therapeutic response and detect disease recurrence. Furthermore, these markers may give indications for targets for novel therapeutic strategies.

## 6 Experiments and Results

We focus on a supervised data analysis and reduce the dimensionality of the data by use of a problem specific wavelet analysis combined with a statistical selection criterion. We avoid statistical assumptions with respect to the underlying data sets, but take only measurement specific knowledge into account.

Hence we have a 101 and a 40 dimensional space of wavelet coefficients and we use multiple algorithms and metrics to determine classification models. We focus on the presented SRNG algorithm.

We trained in a first investigation a SRNG with 1 prototype per class which has been initialized as the mean of 30 randomly selected points from the training data, labeled by a post labeling procedure. The prototype optimization was done until convergence with an upper limit of 2000 iterations and a learning rate of $\alpha = 0.01$ using the strategy as proposed in [35] and [12]. The relevance

parameters $\lambda_i$ of the scaled Euclidean metric are adapted in parallel. This leads to a ranking of the input dimensions according to their importance for classification. A typical relevance profile using scaled Euclidean metric is depicted in Figure 3. The most important frequencies are indicated by high spiked (absolute) values. The depicted frequencies contribute substantially to classification accuracy and, therefore, are important for distinction of the classes. In all analyses we used a 5-fold CV in accordance to the suggestions in [21] because the number of sample is not so small and they are reliable homogeneous per group.

Considering the CRC study the SRNG models obtained at least $\approx 78\%$ cross validation accuracy in a 5-fold cross-validation. The usage of relevance learning typically improved the results by 10% such that a good prediction accuracy of around 90% could be achieved. The LC data set was found to be more complicated and the best obtained predictions are close to 80%. Considering the relevance profiles, looking for high ranked features, the data show the following picture. For the CRC study both metrics scaled functional and scaled Euclidean metric show similar profiles as depicted in Figure 3, the most significant features are consistent with findings as obtained by a standard peak based analysis. For the LC data set the situation is different. For the profile with scaled Euclidean metric most features are ranked as equally important with some minor exceptions. The most significant feature is encoding a peak not picked by the standard approaches and gives a cross validation accuracy of $\approx 78\%$ for its own using a kNN ($k = 3$) classifier on that feature. This shows that the wavelet encoding may help to reveal discriminative features and peaks not identified so far. The relevance profile on the LC data using the functional metric is a bit more diverse. The feature rankings are still similar with respect to the Euclidean profile but some features are pruned. Here different explanations are possible. For one position in the profile at around $2660Da$ a closer inspection with respect to the original data shows that this peak is the main peak of a quintet of closely located peaks. In the Euclidean relevance profile each peak got some relevance and the main peak obtained a higher relevance. In the functional metric only the right neighbor of the main peak is weighted high while the remaining neighbored peaks are pruned out. Further a correlation analysis of the intensities of the associated peak at $2670Da$ shows, that the discrimination power of this peak is similar to that of the new peak at around $2790Da$ which was pruned out in the functional metric but was most significant using the Euclidean metric. Hence the data representation of the functional metric is more sparse but similar discriminative as also visible in the crossvalidation results which are slightly better using the scaled functional norm on the LC data set. A comparison of the SRNG results using the different metrics and alternative algorithms is given in Table 1. It should be mentioned that for SVM the presented functional metric can not be applied directly because the generalized $L^p$ distance has no inner product. A potential alternative would be the use of a Sobolev metric which

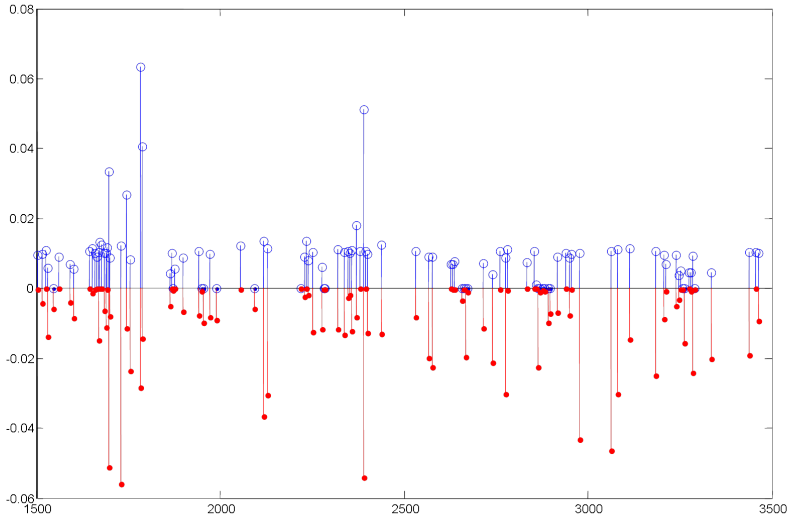



Fig. 3. Visualization of a typical relevance profile obtained by SRNG using scaled Euclidean metric (upper part) and the functional norm (lower negative part of the plot) on the CRC study. Features with larger values indicate higher relevance with respect to the classification task. The x-axis indicates the relative mass position of the corresponding wavelet coefficient in the original spectrum. The y-axis is a relevance measure $\in [0, 1]$. Here relevances for the functional norm are indicated by negative values for illustration purposes.

mimics the functional nature of $L^p$ distance but supports an inner product making the generation of a functional kernel possible [33].

One observes that the results are competitive with respect to other classifiers. The wavelet prepared data perform similar than a standardized peak picking approach with other parameters fixed but allow also the usage of features with complicated peak shape or smaller S/N level, which may be overseen by a standard peak picking approach. Considering the cross validation results for each data set in Table 1 it can be observed, that similar results were obtained using the different metrics. However the metrics itself show different properties. The relevance profile of the scaled Euclidean metric indicates most important data features in a univariate interpretation whereas the generalized $L^p$ norm takes local neighborhoods or correlations in the data space into account while keeping the functional nature of the MS spectra. Therefore also descents in the function and not just peaks as well as correlative effects can be interpreted as relevant features. This trace of information can be further analysed by e.g. LC/MS techniques to test if a potential useful pattern can be observed which in the current linear measurement has not been sufficiently resolved so far. Beside of these good results the LVQ based approaches generates models which can be interpreted very easily by clinicians because the primal model parameters (prototypes) are representative for their receptive field. This is

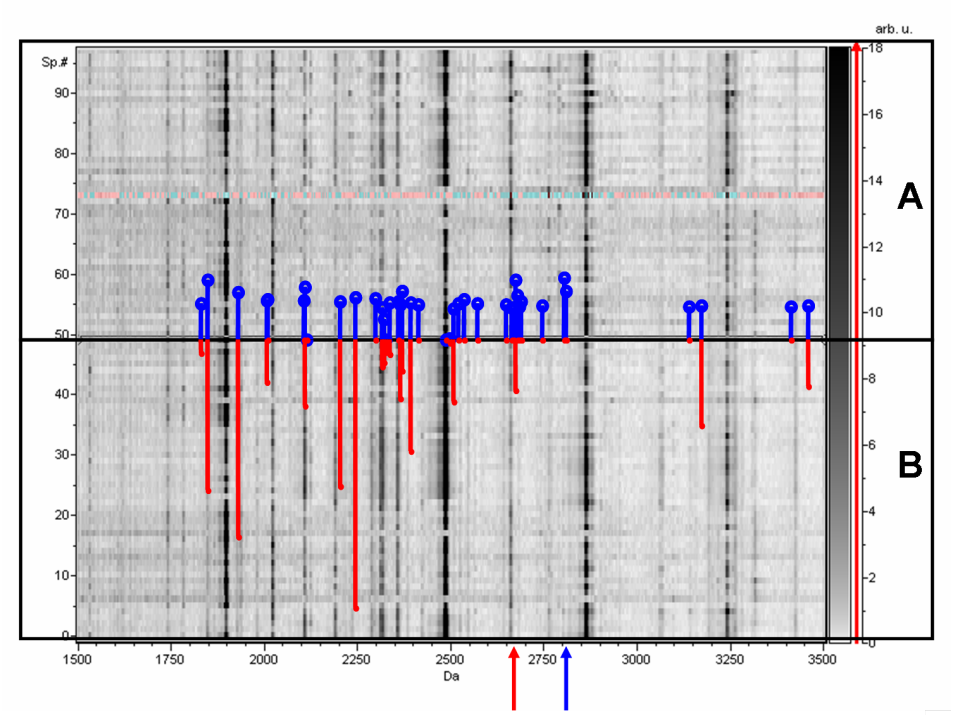



Fig. 4. A gel view of the two classes (LC study) with the control class (region A) and cancer class (region B). The relevant mass positions are indicated by arrows (bottom) using the relevance profile of SRNG with scaled Euclidean metric (top overlaid plot) or functional norm (bottom overlaid plot).

similar to the concept of a prototypical patient.

In Figure 5 an illustration of conformal prediction results for 20 samples of the lung cancer data set is given. The conformal prediction was done using the SRNG with the parametrized functional metric and the parameter settings as mentioned above. To interpret the shown values one should remember that high (e.g. 100%) confidence means, that all labels except the predicted one are unlikely. If say, the 10th example where predicted wrongly, this would mean that a rare event (of probability arround 1%) had occured; therefore, we expect the prediction to be correct which it is. In the case of the item 8 the confidence is also quite high (arround 90%), but we can see that the credibility is low arround 30%. From the confidence we can conclude, that the alternative label is excluded at the 10% level, but the predicted label itself is excluded at a level of around 30%. This shows, that the prediction algorithm was unable to extract from the training set enough information to allow us to confidently classify this example: the strangeness of the labeling different from the predicted label may be due to the fact, that the object itself is strange; perhaps the spectrum is very different from all examples in the training set. Unsurprisingly, the prediction for this example is wrong. In general, high confidence shows that all alternatives to the predicted label are unlikely. Low credibility means that the whole situation is suspect. In summary

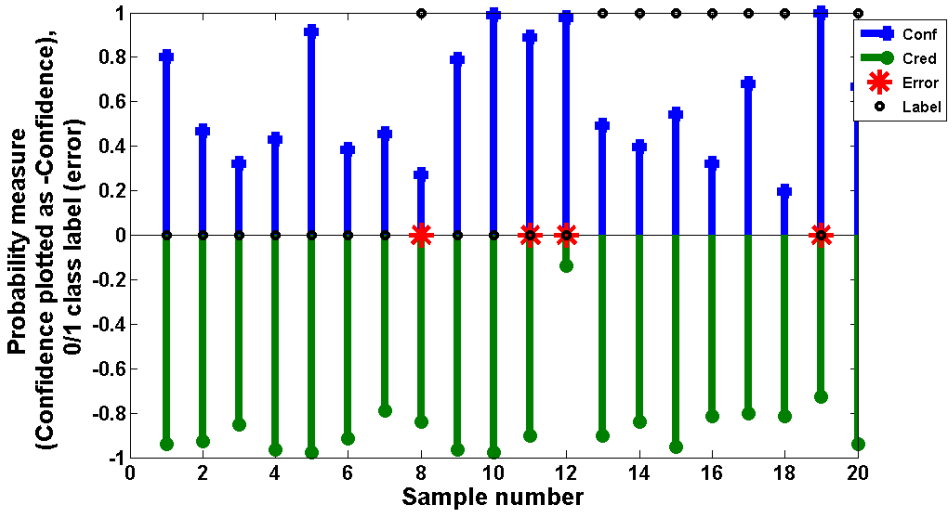



Fig. 5. Visualization of conformal prediction results for 20 samples of the lung cancer data set using the parametrized functional metric. Positive entries show the values for the credibility in an obtained class prediction and negative values indicate the confidence of the single results. The predicted class labels $0, 1$ are given by black circles at 0 or 1 respectively. Miss classifications are indicated by red stars at the 0-level.

we can trust a prediction if the confidence is close to 100% and the credibility is not low (e.g. not less than 5%) [39,9]. Taking this advice into account (with a confidence threshold of 95%) and reanalyzing the results shown in figure 5, only the items $\{4, 5, 9, 10, 15\}$ would be consideres as trusty results with high confidence and moderate or high credibility and indeed the labels for these items are correct predicted. Lowering the confidence level to 90% gives 10 trusty results, but for item 11 the prediction is wrong which means, that for this item a rare event has occurred. An analysis of further samples sets, in the way as shown in Figure 5 reveals that in general very low credibility or low confidence with high credibility, for a single item is indeed a good indicator for miss classifications, motivating the rejection of this item or assignment to the *reject* class. Which in our case of two classes should be interpreted as an unclear classification, where the considered item may belong to non of the two classes. Using the methodology of conformal prediction classification results can be judged not only on the basis of averaged cross validation accuracies but also in a fine granular single item analysis.

Initial results using the conformal prediction approach are promising. The conformal prediction on the test data sets give similar accuracy than with the standard classifiers but in addition for each datapoint a confidence and credibility measure becomes available which allows a judgment of the classification decision for each single patient in a statistical manner.

| Dataset | CRC data | | | LC data | | |
|---|---|---|---|---|---|---|
| Method | CV-$\bar{\text{Rec}}$ | CV-$\bar{\text{Conf}}$ | CV-$\bar{\text{Cred}}$ | CV-$\bar{\text{Rec}}$ | CV-$\bar{\text{Conf}}$ | CV-$\bar{\text{Cred}}$ |
| SNG-EUC | 77.89% | 89.28% | 60.15% | 75.00% | 85.07% | 64.15% |
| SNG-$L^p$ | 78.95% | 89.41% | 60.00% | 75.00% | 85.06% | 64.20% |
| SRNG-EUC | 90.53% | 95.86% | 53.48% | 74.00% | 88.52% | 63.74% |
| SRNG-$L^p$ | 89.47% | 95.68% | 56.42% | 78.00% | 88.80% | 59.67% |
| SVM-Linear | 88.42% | *n.a.* | *n.a.* | 67% | *n.a.* | *n.a.* |
| SVM-RBF | 90.53% | *n.a.* | *n.a.* | 72% | *n.a.* | *n.a.* |
| SVM-CPT | 86.00% | *n.a.* | *n.a.* | 74.00% | *n.a.* | *n.a.* |
| SNN-CPT | 85.78% | *n.a.* | *n.a.* | 72.00% | *n.a.* | *n.a.* |

Table 1
Cross validated prediction accuracies, and corresponding mean confidence and credibility values for SRNG using conformal prediction and different distance measures in comparison to alternative standard approaches on wavelet encoded data. The last two rows are for comparison with the standard peak picking based approach as available in ClinProTools using default settings for SVM and SNN (a prototype classifier approach similar to SRNG).

## 7 Conclusions

We presented a specific pre-processing for mass spectrometric data analysis combined with an extension of the SRNG by a functional metric and integration of conformal prediction. The presented processing of the spectra aims on a natural compact encoding of the signals by means of a functional representation, while the classification model is especially suited to deal with high dimensional sparse data and allows strong regularizations to reduce overfitting effects.

In an initial setup the presented scenario has been embedded into a conformal prediction approach which allows the determination of clinical relevant confidence measures. The extension of conformal prediction for multiple types of prototype based classifiers has been presented.

Beside of the good results the problem of high dimensionality is still remaining. An analysis of proteomic spectra based on peak lists is in general easier to handle, e.g. it is easy to apply multiple different classification models. The wavelet based approach leads to a compact but still high dimensional representation of the data and overfitting may be a stronger issue than in contrast to a standard peak picking approach.

In future research a stronger integration of domain specific knowledge will be

tried to overcome these problems and to make the approach more robust and easier to apply [6]. We will also apply the method using the priorly motivated Sobolev-Kernel[33] to improve the functional encoding using SVM.

## References


[1] A. Rieder A.K. Louis, P. Maaß. *Wavelets: Theory and Applications*. Wiley, 1998.

[2] C Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, New York, NY, 2006.

[3] A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45(5):485–560, 1992.

[4] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Reliability parameters to improve combination strategies in multi-expert systems. *Pattern Analysis and Applications*, 2(3):205–214, 1999.

[5] C. de Stefano, C. Sansone, and M. Vento. To reject or not to reject: that is the question: an answer in case of neural classifiers. *IEEE Transactions on Systems, Man and Cybernetics Part C*, 30(1):84–93, 2000.

[6] G.M. Fiedler, S. Baumann, A. Leichtle, A. Oltmann, J. Kase, J. Thiery, and U. Ceglarek. Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clinical Chemistry*, 53(3):421–428, 2007.

[7] A. Gammerman, I. Nouretdinov, B. Burford, A. Chervonenkis, V. Vovk, and Z. Luo. Clinical mass spectrometry proteomic diagnosis by conformal predictors. *Statistical applications in genetics and molecular biology*, (accepted), 2008.

[8] A. Gammerman and V. Vovk. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 287:209–217, 2002.

[9] A. Gammerman and V. Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.



[6] ACKNOWLEDGMENT: The authors are grateful to T. Elssner and M. Gerhard for useful discussions and support in interpretation of the results (both Bruker Daltonik GmbH Leipzig/Bremen, Germany). Further we would like to thank Luo Zhiyuan for helpful discussions on Hedging predictions (Computer Learning Research Center (CLRC), Royal Holloway, University of London, UK). Frank-Michael Schleif would also like to thank Beate Müller (Ritsumeikan University, Japan) for an effective working atmosphere during preparation of this paper.

[10] N. Guerreiro, B. Gomez-Mancilla, and S. Charmont. Optimization and evaluation of seldi-tof mass spectrometry for protein profiling of cerebrospinal fluid. *Proteome science*, 4:7, 2006.

[11] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Proc. Letters*, 21(2):109–120, 2005.

[12] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Proc. Letters*, 21(1):21–44, 2005.

[13] B. Hammer and Th. Villmann. Generalized relevance learning vector quantization. *Neural Netw.*, 15(8-9):1059–1068, 2002.

[14] Barbara Hammer, Marc Strickert, and Thomas Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, February 2005.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[16] R. Ketterlinus, S-Y. Hsieh, S-H. Teng, H. Lee, and W. Pusch. Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotools software. *Bio techniques*, 38(6):37–40, 2005.

[17] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ext. Ed. 1997).

[18] J. Lee and M. Verleysen. Generalizations of the lp norm for time series and its application to self-organizing maps. In Marie Cottrell, editor, *5th Workshop on Self-Organizing Maps*, volume 1, pages 733–740, 2005.

[19] 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.

[20] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[21] A.M. Molinaro, R. Simon, and R.M. Pfeiffer. Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[22] W. Pusch, M. Flocco, S.M. Leung, H. Thiele, and M. Kostrzewa. Mass spectrometry-based clinical proteomics. *Pharmacogenomics*, 4:463–476, 2003.

[23] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[24] E. Schäffeler, U. Zanger, M. Schwab, and M. Eichelbaum et al. Magnetic bead based human plasma profiling discriminate acute lymphatic leukaemia from non-diseased samples. In *52st ASMS Conference 2004*, page TPV 420, 2004.

[25] R. Schipper, A. loof, J. de Groot, L. Harthoorn, W. van Heerde, and E. Dransfield. Salivary protein/peptide profiling with seldi-tof-ms. *Annals of the New York Academy of Science*, 1098:498–503, 2007.

[26] F.-M. Schleif. *Prototype based Machine Learning for Clinical Proteomics*. PhD thesis, Technical University Clausthal, Technical University Clausthal, Clausthal-Zellerfeld, Germany, 2006.

[27] F.-M. Schleif, U. Clauss, Th. Villmann, and B. Hammer. Supervised relevance neural gas and unified maximum separability analysis for classification of mass spectrometric data. In *Proceedings of ICMLA 2004*, pages 374–379. IEEE Press, December 2004.

[28] F.-M. Schleif, B. Hammer, and Th. Villmann. Supervised neural gas for functional data and its application to the analysis of clinical proteom spectra. In *Proc. of IWANN 2007*, pages 1036–1044, 2007.

[29] F.-M. Schleif, M. Lindemann, P. Maass, M. Diaz, J. Decker, T. Elssner, M. Kuhn, and H. Thiele. Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. *Computing and Visualization in Science*, page in press, 2008.

[30] F.-M. Schleif, T. Villmann, and B. Hammer. Analysis of proteomic spectral data by multi resolution analysis and self-organizing-maps. In *Proc. of CIBB 2007*, pages 563–570, 2007.

[31] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. http://alrw.net (20.10.2007), 2007.

[32] V Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[33] T. Villmann. Sobolev metrics for learning of functional data - mathematical and theoretical aspects. *Machine Learning Reports*, 1(MLR-03-2007), 2007. ISSN:1865-3960 http://www.uni-leipzig.de/compint/mlr/mlr_03_2007.pdf.

[34] T. Villmann, F.-M. Schleif, B. Hammer, and M. Kostrzewa. Exploration of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.

[35] T. Villmann, M. Strickert, C. Brüß, F.-M. Schleif, and U. Seiffert. Visualization of fuzzy information in in fuzzy-classification for image sagmentation using MDS. In *Proc. of ESANN 2007*, pages 103–108, 2007.

[36] Th. Villmann and B. Hammer. Supervised neural gas for learning vector quantization. In D. Polani, J. Kim, and T. Martinetz, editors, *Proc. of the 5th German Workshop on Artificial Life (GWAL-5)*, pages 9–16. Akademische Verlagsgesellschaft - infix - IOS Press, Berlin, 2002.

[37] Th. Villmann, F.-M. Schleif, and B. Hammer. Supervised neural gas and relevance learning in learning vector quantisation. In *Proceedings of WSOM 2003*, pages 47–52, 2003.

[38] 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.

[39] V. Vovk, A. Gammerman, and G. Shafer. *Alorithmic Learning in a Random World*. Springer, New York, 2005.

[40] 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.