Prototype Based Classification in Bioinformatics

Frank-M. Schleif
University of Leipzig, Germany

Thomas Villmann
University of Leipzig, Germany

Barbara Hammer
Technical University of Clausthal, Germany

INTRODUCTION

Bioinformatics has become an important tool to support clinical and biological research and the analysis of functional data, is a common task in bioinformatics (Schleif, 2006). Gene analysis in form of micro array analysis (Schena, 1995) and protein analysis (Twyman, 2004) are the most important fields leading to multiple sub *omics*-disciplines like pharmacogenomics, glycoproteomics or metabolomics. Measurements of such studies are high dimensional functional data with few samples for specific problems (Pusch, 2005). This leads to new challenges in the data analysis. Spectra of mass spectrometric measurements are such functional data requiring an appropriate analysis (Schleif, 2006). Here we focus on the determination of classification models for such data. In general, the spectra are transformed into a vector space followed by training a classifier (Haykin, 1999). Hereby the functional nature of the data is typically lost. We present a method which takes this specific data aspects into account. A wavelet encoding (Mallat, 1999) is applied onto the spectral data leading to a compact *functional* representation. Subsequently the Supervised Neural Gas classifier (Hammer, 2005) is applied, capable to handle functional metrics as introduced by Lee & Verleysen (Lee, 2005). This allows the classifier to utilize the functional nature of the data in the modelling process. The presented method is applied to clinical proteome data showing good results and can be used as a bioinformatics method for biomarker discovery.

BACKGROUND

Applications of mass spectrometry (ms) in clinical proteomics have gained tremendous visibility in the scientific and clinical community (Villanueva, 2004) (Ketterlinus, 2005). One major objective is the search for potential classification models for cancer studies, with strong requirements for validated signal patterns (Ransohoff, 2005). Primal optimistic results as given in (Petricoin, 2002) are now considered more carefully, because the complexity of the task of biomarker discovery and an appropriate data processing has been observed to be more challenging than expected (Ransohoff, 2005). Consequently the main recent work in this field is focusing on optimization and standardisation. This includes the biochemical part (e.g. Baumann, 2005), the measurement (Orchard, 2003) and the subsequently data analysis (Morris, 2005)(Schleif 2006).

PROTOTYPE BASED ANALYSIS IN CLINICAL PROTEOMICS

Here we focus on classification models. A powerful tool to achieve such models with high generalization abilities is available with the prototype based Supervised Neural Gas algorithm (SNG) (Villmann, 2002). Like all nearest prototype classifier algorithms, SNG heavily relies on the data metric d, usually the standard Euclidean metric. For high-dimensional data as they occur in proteomic patterns, this choice is not adequate due to two reasons: first, the functional nature of the data should be kept as far as possible. Second the noise present in the data set accumulates and likely disrupts the classification when taking a standard Euclidean approach. A functional representation of the data with respect to the used metric and a weighting or pruning of especially (priory not known) irrelevant function parts of the inputs, would be desirable. We focus on a functional distance measure as recently proposed in (Lee, 2005) referred as functional metric. Additionally a feature selection is applied based on a statistical pre-analysis of the data. Hereby a discriminative data representation is necessary. The extraction of such discriminant features is crucial for spectral data and typically done by a parametric peak picking procedure (Schleif, 2006). This peak picking is often spot of criticism, because peaks may be insufficiently detected and the functional nature of the data is partially lost. To avoid these difficulties we focus on a wavelet encoding. The obtained wavelet coefficients are sufficient to reconstruct the signal, still containing all relevant information of the spectra, but are typically more complex and hence a robust data analysis approach is needed. The paper is structured as follows: first the bioinformatics methods are presented. Subsequently the clinical data are described and the introduced methods are applied in the analysis of the proteome spectra. The introduced method aims on a replacement of the classical three step procedure of denoising, peak picking and feature extraction by means of a compact wavelet encoding which gives a more natural representation of the signal.

BIOINFORMATIC METHODS

The classification of mass spectra involves in general the two steps peak picking to locate and quantify positions of peaks and feature extraction from the obtained peak list. In the first step a number of procedures as baseline correction, denoising, noise estimation and normalization are applied in advance. Upon these prepared spectra the peaks have to be identified by scanning all local maxima. The procedure of baseline correction and recalibration (alignment) of multiple spectra is standard, and has been done here using ClinProTools (Ketterlinus, 2006). As an alternative we propose a feature extraction procedure preserving all (potentially small) peaks containing relevant information by use of the discrete wavelet transformation (DWT). The DWT has been done using the Matlab Wavelet-Toolbox (see http://www.mathworks.com). 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. For feature selection the Kolmogorov-Smirnoff test (KS-test) (Sachs, 2003) has been applied. The test was used to identify features which show a significant ($p <$ 0.01) discrimination between the two groups (cancer, control). In (Waagen, 2003) also a generalization to a multiclass experiment is given. The now reduced data set has been further processed by SNG to obtain a classification model with a *small* ranked set of features. The whole procedure has been cross-validated in a 10-fold cross validation.

WAVELET TRANSFORMATION IN MASS SPECTROMETRY

Wavelets have been developed as powerful tools (Rieder, 1998) used for noise removal and data compression. The discrete version of the continuous wavelet transform leads to the concept of a multi-resolution 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 (Cohen, 1992), which uses two pairs of scaling and wavelet functions. One is for the decomposition/analysis and the other one for reconstruction/synthesis, giving a higher degree of freedom for the shape of the scaling and wavelet function. In our analysis such a smooth synthesis pair was chosen. 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. The denoised spectrum looks similar to the reconstruction as depicted in Figure 1.

One obtains approximation- and detail-coefficients (Cohen, 1992). The approximation coefficients describe a generalized peak list, encoding primal spectral information. For linear 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 appropriate for our problem (see Figure 1). Applying this procedure including the KS-test on the spectra with an initial number of 22306 measurement points per spectrum one obtains 602 wavelet coefficients