Chapter 2.17
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.

DOI: 10.4018/978-1-60960-561-2.ch217

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

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

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 cor-

6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/prototype-based-classification-bioinformatics/53603

Related Content

An Advanced Concept of Altered Auditory Feedback as a Prosthesis-Therapy for Stuttering Founded on a Non-Speech Etiologic Paradigm
Manuel Prado-Velascoand Carlos Fernández-Peruchena (2011). *Clinical Technologies: Concepts, Methodologies, Tools and Applications (pp. 1284-1326).*
www.irma-international.org/chapter/advanced-concept-altered-auditory-feedback/53650

Wearable Kinesthetic System for Joint Knee Flexion-Extension Monitoring In Gait Analysis
Mario Tesconi, Enzo Pasquale Scilingo, Pierluigi Barbaand Danilo De Rossi (2011). *Clinical Technologies: Concepts, Methodologies, Tools and Applications (pp. 792-800).*
www.irma-international.org/chapter/wearable-kinesthetic-system-joint-knee/53620

A Framework for Information Processing in the Diagnosis of Sleep Apnea
Udantha R. Abeyratne (2011). *Clinical Technologies: Concepts, Methodologies, Tools and Applications (pp. 295-304).*
www.irma-international.org/chapter/framework-information-processing-diagnosis-sleep/53589

Adoption of Electronic Health Records
Yousuf J. Ahmad, Vijay V. Raghavanand William Benjamin Martz Jr. (2011). *Clinical Technologies: Concepts, Methodologies, Tools and Applications (pp. 132-146).*
www.irma-international.org/chapter/adoption-electronic-health-records/53581

Risks and Benefits of Technology in Health Care
Stefane Kabeneand Melody Wolfe (2011). *Clinical Technologies: Concepts, Methodologies, Tools and Applications (pp. 13-24).*
www.irma-international.org/chapter/risks-benefits-technology-health-care/53574