Robust Peak Detection and Alignment of nanoLC-FT Mass Spectrometry Data

Marius C. Codrea[1], Connie R. Jiménez[2], Sander Piersma[2], Jaap Heringa[1], and Elena Marchiori[1]

[1] Centre for Integrative Bioinformatics VU (IBIVU)
Department of Computer Science
mcodrea@few.vu.nl, heringa@few.vu.nl, elena@cs.vu.nl
[2] Cancer Center Amsterdam
Vrije Universiteit Amsterdam, The Netherlands
c.jimenez@vumc.nl, S.Piersma@vumc.nl

Abstract. In liquid chromatography-mass spectrometry (LC-MS) based expression proteomics, samples from different groups are analyzed comparatively in order to detect differences that can possibly be caused by the disease under study (potential biomarker detection). To this end, advanced computational techniques are needed. Peak alignment and detection are two key steps in the analysis process of LC-MS datasets. In this paper we propose an algorithm for LC-MS peak detection and alignment. The goal of the algorithm is to group together peaks generated by the same peptide but detected in different samples. It employs clustering with a new weighted similarity measure and automatic selection of the number of clusters. Moreover, it supports parallelization by acting on blocks. Finally, it allows incorporation of available domain knowledge for constraining and refining the search for aligned peaks. Application of the algorithm to a LC-MS dataset generated by a spike-in experiment substantiates the effectiveness of the proposed technique.

1 Introduction

Computational analysis of proteomic datasets is becoming of crucial relevance for discovery of reliable and robust candidate biomarkers. In particular, quantitation of changes in protein abundance and/or state of modification is the most promising, yet most challenging aspect of proteomics. In recent years label-free LC-MS methods that quantify absolute ion abundances of peptides and proteins have emerged as promising approaches for peptide quantitation and profiling of large numbers of clinical samples [1].

Briefly, peptides are subjected to (multi-dimensional) liquid chromatography for separation. Each peptide fraction is then analyzed on an LC-MS system. Each LC-MS run of a sample generates a pattern of very high input dimension consisting of one intensity (relative abundance) measurement for each pair of molecular mass-to-charge ratio (m/z) and retention time (RT) values. Ideally, the same molecules detected in the same LC-MS instrument should have the

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 35–46, 2007.

same retention time, molecular weight, and signal intensity. However, in practice this does not happen due to experimental variations. As a consequence, patterns generated by LC-MS runs need to undergo a number of processing steps before they can be comparatively analyzed. Such processing steps include normalization [5,17], background subtraction [8], alignment [4,13,16,17], and peak detection (e.g., [9]). Several tools and algorithms for processing and for difference analysis of LC-MS datasets have been introduced (e.g., [2,3,7,9,11,15,20,17]).

In this paper we focus on peak detection and alignment. As well explained in the overview paper by Listgarten et al [10], alignment algorithms involve either (i) the maximization of an objective function over a parametric set of (generally linear) transformations, or (ii) non-parametric alignment based on dynamic programming, or (iii) combination of these methods like piecewise transformations. They act either on the full pattern or on features (peaks) selected beforehand; they may or may not use the signal intensity and they may or may not incorporate scaling. Most of alignment algorithms require a reference template, to which all time series are aligned. Peak detection is usually performed in an ad-hoc manner [10], involving either a comparison of intensities with neighbours along the m/z axis [19] or detection of coinciding local maxima [18].

Whether one should perform peak detection before [14,17] or after [12] alignment has not been clearly established. In this paper we circumvent this issue by performing both tasks at the same time by means of a novel clustering algorithm. The motivations for a new algorithm rely also on the desire to overcome drawbacks of alignment algorithms, such as the need for a reference template, or the assumption of a given (local or global, usually linear) transformation in RT dimension. The versatility of clustering for simultaneous alignment and peak detection has been already recognized by Tibshirani et al. in [16]. They align MALDI (a technique that generates two dimensional patterns of m/z and intensity values from each sample) data along the m/z axis by applying one-dimensional hierarchical clustering with complete linkage for constructing the dendrogram and a specific cutoff for extracting clusters representing peaks.

The algorithm proposed here, called Peak Detection and Alignment (`PDA`), acts on blocks of m/z values. Blocks are obtained by splitting the runs along the m/z axis (e.g., in blocks of equal size). Each block is processed individually. The input of `PDA` is a set of (blocks of) runs described by a list of triplets (m/z, RT, run id) consisting of 'm/z' value, 'RT' value and run identifier. The output of `PDA` is a set of clusters of (m/z, RT) pairs representing peaks, together with information about their signal intensity.

Novel features of `PDA` include:

1. a similarity measure for comparing features, where a feature is a triplet (m/z, RT, id). Weights are associated to each of the three attributes to specify their relevance;
2. a cluster merging strategy;
3. a cluster refinement procedure for handling peaks occurring near the boarder of blocks.