A new approach for the analysis of mass spectrometry data for biomarker discovery

Barbarini N, Magni P, Bellazzi R

Department of Computer Science And Systems, University of Pavia, Pavia

Motivation

The recent developments in sample preparation and mass spectrometry allow to measure simultaneously the expression level of thousands of proteins. For this reason, in the last few years an increasing interest has been devoted to the analysis of the body fluids proteome, mainly for diagnostic purposes. In particular the SELDI/MALDI-TOF techniques represent promising tools for the discovery of biomarkers, i.e. the protein signatures associated to a particular disease. However, the identification of such biomarkers is not straightforward due to the presence of several sources of complexity; moreover a well-established procedure for data analysis is not yet available. In the present work, we will propose a new strategy for the analysis of SELDI/MALDI-TOF data based on a three steps procedure for i) data-preprocessing, ii) feature (mass/charge ratio, m/z) reduction and selection and iii) for the association of the selected features to a list of known proteins.

Methods

The proposed methodology for the analysis of the mass spectrometry data consists of three steps. In the first step, many algorithms for the preprocessing of mass spectra are considered. We search for the best sequence of preprocessing algorithms, which is able to maximize the classification accuracy calculated with a simple classifier, based on the difference between the over and underexpressed m/z peaks. The search strategy follows a stepwise approach. In the second step, to decrease the data complexity and to increase the information associated to each feature, the original mass/charge data (e.g. about 300000 in SELDI/TOF high resolution data) are reduced by grouping together the m/z values corresponding to the same protein. Our algorithm exploits the available knowledge on the mass spectrometry technique (e.g. routine resolution) and the chemical properties of proteins (e.g., isotopic distribution). In particular, the algorithm: i) computes the median spectrum of the preprocessed spectra of all subjects; ii) smoothes the median spectrum with a moving window (whose width is equal to the resolution of the spectrometer); iii) finds the maxima of the smoothed curve; iv) computes the sum of the m/z intensities of the isotopic distribution of each maximum (taking into account the routine resolution). The output of this step is therefore a suitable binning of the spectra which is then applied to the data of each subject. In this way, we obtain a reduced number of features that can be used for further analysis. In particular, it is then possible to apply any of the available algorithms to select the most differentially expressed or to define a classifier for diagnostic/prognostic purposes. In the third step, every feature computed and eventually selected in the previous step are associated to a list of proteins that could generate the isotopic distribution. To this aim, a local database composed of proteins and fragments annotated in the Entrez protein database has been created. A list of proteins can be associated to each feature by simply selecting in the local database the entries with molecular weight around the mass of the feature of interest. To reduce the length of such list, it is possible to consider only the proteins that contain at least a peptide discovered in human serum by Plasma Proteome Project.

Results

The proposed methodology has been applied to a public dataset regarding ovarian cancer (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). It consists of 216 mass spectra (121 ovarian cancer patients and 95 healthy women) obtained from serum samples by mean of the SELDI-QqTOF technology with WCX2 ProteinChip. The first step of our procedure selected three algorithms for the data preprocessing phase: baseline correction and two smoothing filters (lowess and Sawitzky-Golay). In the second step the initial 373401 m/z were reduced to 3282 features, that correspond to 3282 different isotopic distributions. These features were used as input of several classification algorithm (e.g, decision trees, k-nearest neighbours, etc.) to discriminate between the

two classes (cancer and healthy women). In the third step, the most differentially expressed feature (that classifies the subjects with an accuracy of 90%) was associated to a short list of proteins, among which a possible biomarker for ovarian cancer can be found. The validity of this approach was preliminary tested with success on the protein identification problem of an isotopic distribution reported and experimentally validated in a previous study.

Contact email: nicola.barbarini01@ateneopv.it