MS-Analyzer: Composing and Executing Preprocessing and Data Mining Services for Proteomics Applications

Cannataro M, Veltri P

University Magna Graecia of Catanzaro, Catanzaro, Italy

Motivation

Mass Spectrometry (MS) proteomics produces huge datasets, said spectra, that contain large set of measures (intensity, m/Z), representing the abundance of biomolecules having certain mass to charge ratios. MS data hides a lot of information about cell functions and disease conditions and can be used for various analysis, e.g. biomarker discovery, peptide/protein identification, and sample classification. The discovering of such information needs the combined use of bioinformatics and data mining, and requires the efficient access to huge spectra datasets and various software tools for to the loading, management, preprocessing, and mining of spectra, as well as the interpretation and visualization of discovered knowledge models. The increasing use of MS in clinical studies causes the collection of spectra data from large sample populations, e.g. to control the progression of a disease. In addition, the comparative study of a disease may require the analysis of spectra produced in different laboratories, so it is possible to envision that in few years biomedical researchers will need to collect and analyze more and more spectra data. Since spectra have a high dimensionality and are often affected by errors and noise, specialized spectra databases and preprocessing techniques are needed. Finally, MS involves different technological platforms, such as sample treatments, MS techniques, spectra processing, data mining analysis, and results visualization. Choosing the right methods and tools requires multidisciplinary knowledge from MS specialists to biologists and computer scientists, thus, modelling the semantic of processes, tools, and data is a key issue to simplify application design.

Methods

Ontologies constitute a well established tool to model the steps of data mining applications and support the application design, while Grid technology may provide: efficient storage space where maintaining on line large spectra datasets, broadband infrastructure needed to collect in a secure and efficient way proteomics data coming from remote laboratories, computational power needed by preprocessing and mining algorithms. To address the key issues of spectra data management and analysis we provide the basic MS bioinformatics tools as Web Services and propose the combined use of ontologies for the modelling of protemics tools, workflow techniques to compose in a seamless way basic proteomics services, and the Grid as the deployment infrastructure. The proposed bioinformatics platform, named MS-Analyzer, offers to the biologist a set of high level services, namely: spectra management services, providing spectra format conversion and efficient spectra storage through a specialized spectra database; moreover experimental data are modelled through the dataset concept, i.e. a set of spectra that can be in raw, preprocessed or prepared stage; pre-processing services, that implement common spectra pre-processing algorithms, such as base line subtraction, smoothing, normalization, binning, peaks extraction, and peaks alignment; data preparation services, that provide the spectra reorganization needed when applying data mining tools (e.g. Weka tools require spectra dataset formatted in a unique input file having a specific metadata header); data mining services, obtained by wrapping Weka tools, a popular data mining suite; moreover, tools for knowledge models visualization are provided. An Ontology-based Workflow Editor allows the concept-based browsing/searching of such services modelled through the MS-Analyzer ontologies: WekaOntology, that models the Weka data mining tools and is enriched by the description of relevant spectra concepts and pre-processing algorithms, and ProtOntology, that models concepts, methods, algorithms, tools, and databases relevant to the proteomics domain, and provides the biological background and perspective to the data mining analysis.

Results

The paper presented MS-Analyzer, a Grid-based software platform that supports the semantic composition of spectra preprocessing algorithms, efficient spectra management techniques based on a specialized spectra database, and off-the-shelf data mining services, to analyze mass spectra data on the Grid. By using MS-Analyzer a user can easily design a data mining application with the help and the constraint checks provided by WekaOntology and ProtOntology, and without worrying of software details, having a suite of specialized spectra management services that simplify and automate the path to knowledge discovery. The availability of different services allows to produce in few time many workflows of the same application employing different combination of preprocessing, preparation and data mining techniques. The produced knowledge models as well as the execution performance of the scheduled workflows can be easily visualized, allowing to compare the effect of different preprocessing and data mining techniques and to evaluate the best strategies to analyze mass spectra data.

Contact email: cannataro@unicz.it