Using Ontologies in Proteus for Modeling Data Mining Analysis of Proteomics Experiments

Mario Cannataro, Pietro Hiram Guzzi, Tommaso Mazza, and Pierangelo Veltri

University Magna Græcia of Catanzaro, 88100 Catanzaro, Italy
{cannataro, hguzzi, t.mazza, veltri}@unicz.it

Abstract. Bioinformatics applications are often characterized by a combination of (pre) processing of raw data representing biological elements, (e.g. sequence alignment), and an high level data mining analysis. Developing such applications needs knowledge of both data mining and bioinformatics domains, that can be effectively achieved by combining ontology about the application domain (the problem) and ontology about the resolving approaches (the solution). We talk about using ontologies to model proteomics *in silico* experiments. In particular data mining of mass spectrometry proteomics data is considered.

1 Introduction

Bioinformatics is an emerging research field aiming to help the biologic research with informatics tools. In this sense bioinformatics is a multidisciplinary area that involves performing critical experiments (e.g. sequence alignment, structure prediction), organizing and storing of collected data (e.g. protein data banks), extracting knowledge from data and sharing it (e.g. verify hypothesis about diseases, sharing commonly agreed bio-medical practices and protocols). In a more broad perspective, biomedical experiments whose data are analyzed through bioinformatics platforms, involve different technologies such as mass spectrometry, bio-molecular profiling, nanotechnology, computational chemistry, drug design, and so on. The way in which data are produced, the possible errors affecting them, the assumptions and the approaches to analyze them, that are known to biomedical domain experts, should be taken into account when choosing a particular data mining approach or algorithm. The problem addressed here is how to enhance the design of complex *in silico* experiments combining, in a unique bioinformatics platform, both data mining and biomedical knowledge. Basic technologies used are Data Mining to analyze data, Ontologies to model knowledge, and Workflows to design experiments with several steps involving different informatics tools and spanning different domains.

In Bioinformatics, data mining is useful both in extracting knowledge from literature using Text Mining and in extracting rules and models from databases and experimental data. Ontologies have a broad range of applicability in Bioinformatics, such as classification of medical concepts and data, database integration and collaboration between different groups, and recently, enhancing of application design[1]. A workflow is a partial or total automation of a process, in which a collection of activities must be executed according to certain procedural rules. Workflow Management Systems (WfMSs) allow the design of a workflow, supporting its enactment by scheduling different activities on available entities. PROTEUS [2] is a Grid-based Problem Solving Environment that uses ontologies to model application domain, workflow techniques to compose distributed *in silico* applications, and is developed on Grid middleware. Mass Spectrometry (MS) is a widely used technique for the mass spectral identification of the thousands of proteins that populate complex biosystems such as serum and tissue. The combined use of MS with data mining is a novel approach in proteomic pattern analysis and is emerging as an effective method for the early diagnosis of diseases [5]. A main goal of the paper is to show the use of ontologies in PROTEUS to model and compose *in silico* proteomics experiments where data mining techniques are used to find patterns and classify mass spectrometry proteomic data.

The rest of the paper is organized as follows. Section 2 describes the workflow of a representative proteomic data mining application. Section 3 describes ontologies in PROTEUS and their use to compose proteomic experiments. Finally, Section 4 concludes the paper and outlines future work.

2 Workflow of a Representative Proteomic Experiment

The bio-medicine research group of our University is affording the study of the breast cancer and the overexpression of the HSP90 (heat shot proteins) in the chemo resistant patients [6]. The study is addressed to discover where this overexpression occurs in order to block the production excess of HSP90 and, consequently, to test if this hypothesis is valid and useful to making effective the chemotherapy. Mass Spectrometry is currently a hot research area and this approach is quickly becoming a powerful technique in order to identify different molecular targets in different pathological conditions.

Mass Spectrometry Analysis. *Mass Spectrometry* is a powerful tool for determining the masses of biomolecules and biomolecular fragments present in a complex sample mixture. Understanding the information contained in mass finger printing data is a new hot area of bioinformatics. Before obtaining MS data, biological samples need to be prepared and treated: *Sample Preparation*, *Proteins Extraction*, and *ICAT Protocol* refer, respectively, to the choice of samples to be analyzed (in our experiments we consider serum, tissue, and cell culture samples), the selection of proteins from samples, and their treatment before mass spectrometry. Mass Spectrometry data are represented, at a first stage, as a (large) sequence of value pairs, where each pair contains a measured intensity, which depends on the quantity of the detected biomolecule, and a mass to charge ratio (m/Z), which depends on the molecular mass of detected biomolecule. Due

[1] The myGrid Project: http://www.mygrid.org.uk/

to the large number of m/Z data contained in a mass spectra obtained by real samples, analysis by manual inspection is not feasible. Usually mass spectra are represented in a graphical form as in Fig. 1.

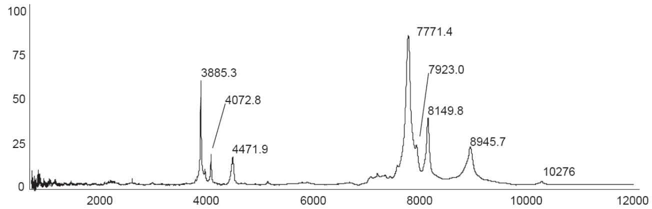

Fig. 1. Example of MS spectrum for breast cells

Data Mining Analysis. Biological data mining is an emerging research area. The high volume of mass spectrometry data is a natural application field of data mining techniques. In fact large sequences of m/Z data contain a lot of information in a implicit way. Manual inspection of experimental data is difficult despite biological relevance of conformation of peaks list. For this reason both computational method and soft-computing techniques can made automatic clustering and pattern discovery. Our aim is to build some clusters (i.e. diseased, healthy patients) in which classify each new collected spectrum. This process needs to identify the distinctive characteristics of each group and then find those in new spectra. [4] and [5] explain the application of this methods in ovarian and prostate cancer. In this way, early detection of cancer can leverage the high throughput of Mass Spectrometry and computational methods. The overall proteomics experiment is described in [6], whereas the high level workflow of the data mining application is shown in Fig. 2

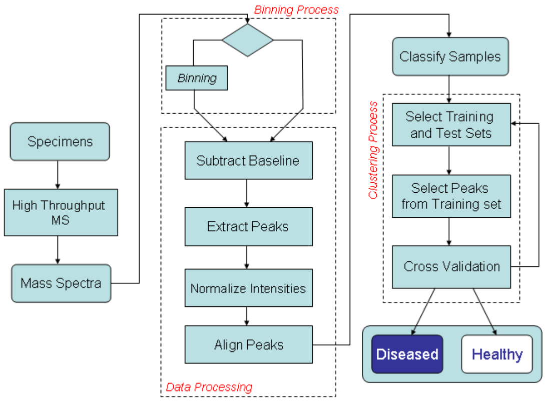

Fig. 2. Clustering Workflow

Data Preprocessing. In order to reduce dimensional complexity of the problem is important to use an efficient and biological consistent data pre-processing phase. Currently this is an open problem. The preprocessing phase involves the following steps as shown in Fig. 2: (i) Identification of Base Line; (ii) Identification and extraction of peaks; (iii) Normalization of intensities; (iv) Alignment of correspondent peaks.

Data Clustering and Classification. In our system we are using the *Matlab* version of the Q5 clustering algorithm, It works on the whole spectrum and is available at[2]. Q5 is a closed-form, exact solution to the problem of classification of complete mass spectra of a complex protein mixture. The Clinical Proteomics Program Databank has provided three set of ovarian cancer data that can be used without restriction with Q5 algorithm [3].

3 Using Ontologies to Enhance Application Design

PROTEUS [2] is a Grid-based Problem Solving Environment that allows modelling, building and executing of bioinformatics application on the Grid. It combines existing software tools and data sources by (i) adding metadata to software, (ii) modelling applications through ontology and workflows, and (iii) offering pre-packaged bioinformatics applications.

DAMON (Data Mining Ontology) ontology [1] represents the features of the available data mining software, classifying their main components and evidencing relationships and constraints among them. It allows the semantic search (concept-based) of data mining software and resources and assists the user in finding the suitable software to conduct a data mining computation.

PROTON (Proteomic Ontology) ontology models the concepts of Proteomic domain. A top-down development process that starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts has been adopted. The rationale is that a fundamental distinction between biological concepts (e.g. proteins) and non-biological concepts (e.g. data-preprocessing) exists. We show respectively the taxonomic, and the non taxonomic relations between different concepts.

PROTON taxonomic relations between biological concepts. In developing PROTON we assumed that a protein is a tangible thing that has not temporal or spatial dynamic (except its biological transformations, such as protein folding). We summarize the post-translational modification as an attribute. This concept has the evident limitation that we cannot approach the biochemical pathways. The justification of our choice is that all the algorithms operate over static representation of proteins in their native forms. In our conceptualization we introduce three primary concepts:(i) *Aminoacid*,(ii) *Protein*, (iii) *Structure*. The structure concept can be specialized in Primary, Secondary, Tertiary and Quaternary, according to biological classification. It is important to precise that

[2] Q5 clustering algorithm: http://www.cs.dartmouth.edu

[3] Clinical Proteomics Program Databank - http://clinicalproteomics.steem.com/

each protein has a unique identifier, and a unique primary structure. These determine the spatial conformation, i.e. secondary and tertiary structure, despite of quaternary one that identifies globular proteins.

PROTON taxonomic relations between non biological concepts. In this area we start from an intangible fact: *Analysis.* In medical research, literature citations of a method define a criterium of evaluation, so the Analysis concept models theoretic study, as literature articles. This class can be specialized in the following subclasses: (i) *Mass-Fingerprinting Analysis*; (ii) *Primary Structure Analysis*; (iii) *Secondary Structure Analysis*; (iv) *Tertiary Structure Analysis*; (v) *Quaternary Structure Analysis.* It is evident that a simple relation exists between each kind of analysis. A *Task* is a concrete problem that a researcher has to solve. Specialization of this concept gives the following sub-classes: (i) *Interpretation of MS data.*, e.g. identifying a protein [7], or recognizing a disease; (ii) *Alignment* that comprises Sequence Alignment and Structural Alignment; (iii) *Prediction* that comprises secondary or tertiary structure prediction starting by primary sequence. A *Method* is a way to perform a task and it is implemented by one or more software tools. A *Software* is an implementation of a method through a computer program or a (Web) service.

PROTON non taxonomic relations. Main non taxonomic relations between two concepts of different classes are: (i) *is Chain of*, explaining that a protein is a sequence of amminoacids; (ii) *Has A* that links protein and its own structure; (iii) *Studies* that links a particular analysis and the corresponding explained protein structure; (iv) *Implements* that links a software to an implemented method.

3.1 Ontology-Based Design of Proteomic Experiments

In order to perform a complete in silico experiment cooperation between biologic and bioinformatics group is necessary. Ontology can give the needed unified semantic. The design and execution of an application on PROTEUS comprises the following steps:

1. **Ontology-based component selection**. PROTON and DAMON are integrated because there exist some concepts (i.e. clustering method) that are included in both ontologies, so using *OnBrowser* [3] a user can browse from an ontology to another on the basis of the current focus, e.g. informatics or biological aspects of the process under investigation. Considering the example described so far (clustering of SELDI-TOF data), a user first browses PROTON and finds in Analysis taxonomy the related literature. Then he/she can explore the Methods taxonomy finding, among others, Soft computing, Probabilistic, and Preprocessing techniques. From there the Software (e.g. Q5 clustering tool) concept can be reached. At this point all Q5 related metadata (e.g. Grid node location, parameters, etc.) can be found using DAMON as shown Fig. 3.
2. **Workflow design**. Selected components are combined producing a workflow schema that can be translated into a standard language, such BPML or

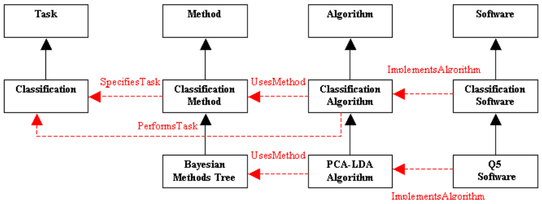

Fig. 3. Damon modelling for Q5

BPEL. Fig. 2 shows the workflow of the mass spectrometry data mining analysis described so far.

3. **Application execution on the Grid**. The workflow is scheduled by a workflow engine on the Grid.
4. **Results visualization and storing**. After application execution and result collection, the user can enrich and extend the PROTEUS ontologies.

4 Conclusions and Future Work

We discussed how to enhance the modelling and design of mass spectrometry data analysis applications in PROTEUS using ontologies, which combine both data mining and biomedical knowledge. Future work will regard the development of a methodology for the use and composition of different domain ontologies addressing different aspects of complex applications development, such as the domain of the problem, the domain of the suitable solutions, and the domain of the available resources.

References

1. M. Cannataro and C. Comito, *A Data Mining Ontology for Grid Programming*, Workshop SemPGrid-2004 (in conj. with WWW2003) (Budapest-Hungary), 2003.
2. M. Cannataro, C. Comito, F. Lo Schiavo, and P. Veltri, *Proteus, a Grid based Problem Solving Environment for Bioinformatics: Architecture and Experiments*, IEEE Computational Intelligence Bulletin **3** (2004), no. 1, 7–18.
3. M. Cannataro, A. Massara, and P. Veltri, *The OnBrowser Ontology Manager: Managing Ontologies on the Grid*, Workshop on "Semantic Intelligent Middleware for the Web and the Grid", ECAI-2004 (Valencia, Spain), 2004.
4. BL. Pdam et al., *A: Proteomic approaches to biomarker discovery in prostate and bladder cancers*, Proteomics **1** (2001), no. 1264-1270.
5. EF. Petricoin et al., *Use of proteomic patterns in serum to identify ovarian cancer*, Lancet **359** (2002), no. 572-577.
6. G. Cuda et al., *Proteomic Profiling of Inherited Breast Cancer: Identification of Molecular Targets for Early Detection, Prognosis and Treatment, and Related Bioinformatics Tools*, WIRN 2003, LNCS, vol. 2859, Springer Verlag, 2003.
7. V. Dancik et al., *De novo peptide sequencing via tandem mass spectrometry*, Journal of Computational Biology **6** (1999), 327–342.