An ontological-based knowledge organization for bioinformatics workflow management system

Antonino Fiannaca, Massimo La Rosa, Salvatore Gaglio, Riccardo Rizzo, Alfonso Urso

ICAR-CNR, National Research Council of Italy, Palermo

Motivation and Objectives

In the field of Computer Science, ontologies represent formal structures to define and organize knowledge of a specific application domain (Chandrasekaran et al., 1999). An ontology is composed of entities, called classes, and relationships among them. Classes are characterized by features, called attributes, and they can be arranged into a hierarchical organization. Ontologies are a fundamental instrument in Artificial Intelligence for the development of Knowledge-Based Systems (KBS). With its formal and well defined structure, in fact, an ontology provides a machine-understandable language that allows automatic reasoning for problems resolution. Typical KBS are Expert Systems (ES) and Decision Support Systems (DSS). ESs gather and formalize the knowledge of a human expert of a domain in order to produce inferences and recommendations given an initial query. DSSs are more interactive KBS, in the sense they offer support, rather than replacement, for the decision making process during the execution of a task, suggesting one possible strategy or tool given a set of initial conditions. DSSs are mainly adopted in the clinical field, where they are called Clinical DSS (CDSS). Ontology specification, structure and organization are then of fundamental importance for the development of a KBS.

In this paper we present an improvement of our ontological approach for knowledge organization in DSS design. In our previous publication (Fiannaca et al., 2012) we defined a paradigm for ontology specification named Data Problem Solver (DPS) and we showed how our approach can be applied to bioinformatics domain, modeling the Protein-Protein Interaction Network extraction scenario. In the proposed approach, we aim at integrating into our ontology the concept of Workflow as a set of processes. Our main objective is to provide a general schema in order to add the functionalities and capability of a DSS to the more recent Workflow Management Systems, that especially in bioinformatics, with the Taverna workbench (Hull et al., 2006), represent a powerful instrument for researchers. We called our extended ontological approach Data Problem Solver Workflow (DPSW).

Methods

DPSW ontology is shown, using UML notation, in Figure 1. The four main entities are, as the name suggests, Data, Problem, Solver and Workflow. Problem represents the set of Tasks to do in an application scenario, and it models the task decomposition from more complex goals to simpler ones. Data summarizes the type of information needed to perform a task belonging to a Problem. Data concept is specialized by Data _ Type class, representing the type of input and output data of a task, and each Data _ Type has one or more Data _ Format that encodes it. Solver concept fills the gap between a Problem to solve and the Tools that actual solve it. Each Solver is characterized by a computational Approach (probabilistic, topological, numerical approach for instance) and it models the expert knowledge (in terms of heuristics or strategies) on which Tool, or combination of Tools, are needed in order to accomplish a Task. Solver class is also characterized by a set of attributes, not shown in Figure 1, that specifies what are the pros and cons for using a solving strategy. Tool class identifies the generic entity that can be actually run, and it generalizes the concept of Algorithm, Web Service, Application, Device. Each Tool has a computational Paradigm (for example neural networks, graph analysis, etc...) and eventually a set of configuration Parameters; it requires a Data _ Format object (the input file), and of course other type of Tools can be further added. By considering separately Problem, Solver and Data, we want to clearly separate among the models of the problem itself, the way to resolve it, and the input data requested. This way we aim at enhancing the generalization, modularity and expandability features of the proposed ontology. The last main component of the proposed ontological approach is the Workflow entity. It rep-

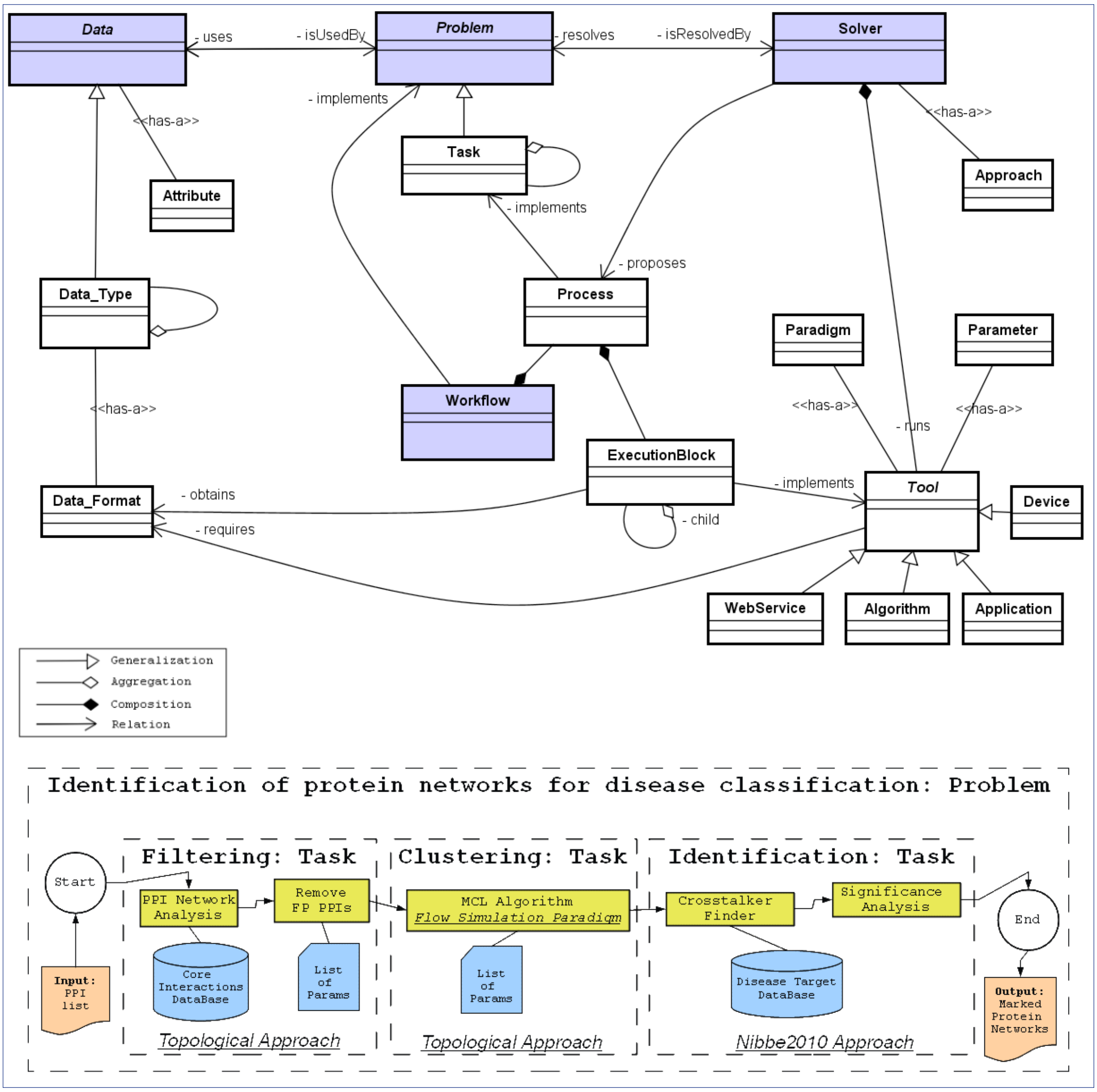

Figure 1 - The proposed DPSW ontology and its case of study.

resents the graphical view of a problem and its solving components. A Workflow is composed of one or more Processes, that can be seen as part of the global workflow implementing a specific Task, and each Process, in turn, is composed of ExecutionBlocks, that are the visual representations of an executed Tool. By embedding the concept of Workflow into our ontological structure, we want to provide a full Knowledge Base specification that can be used as building block of a DSS whose suggestions during a bioinformatics experiment can immediately be translated into an executive workflow.

Results and Discussion

In order to show how the proposed ontology can match with a real bioinformatics issue, we have taken into account a key challenge of cancer research, i.e., the detection of protein sub-networks that identifies markers correlated with metastasis. In facts, each protein complex is suggestive of a distinct functional pathway, that can provide novel hypotheses in organisms analysis (Sharan et al., 2007). A workflow related to this case of study is reported in the bottom of the Figure 1. In this example, we consider the "identification of protein networks for disease classification", that, according to DPSW ontology, represents the problem concept; the implementation of this problem (the experiment) matches with the workflow concept. Here, we take as data input a list of protein-protein interactions (PPIs) and produces as data output a list of marked protein network, that could be responsible for some specific diseases. According to the related literature, this problem could be arranged in three main tasks: filtering, clustering and identification. For instance, the first task has been handled by some authors (Ucar et al., 2005) with a topological approach; in facts, they developed some graph-based algorithms in order to eliminate redundant false positive interactions from the original PPI dataset. This preprocessing strategy points to increase the reliability of PPI-Network. As regarding the second task, i.e. finding meaningful groups of biological units, a number of approaches have been proposed and a lot of them are based on clustering. A well-know algorithm is Markov Clustering Algorithm (MCL) (Enright et al., 2002), that divides the graph by means of "flow simulation paradigm". In facts, it separates the graph into different segments, with an iteration of simulated random walks within a graph. Once sub-networks are obtained, it is possible to identify those complexes that demonstrate a differential expression with respect to carcinogenesis phenotype, by means of an integrative -omics approach proposed in (Nibbe et al., 2010). Using these elements, we could obtain some putative disease protein sub-networks. Ultimately, in order to face with this case of study, we propose to use three tasks, two different approaches and six tools (both algorithms and applications). Each executed tool, with its proper input/output file and parameters, is stored into an instance of the execution block concept, whereas a set of execution blocks that complete a single task are stored as an instance of process concept. Notice that workflow in Figure 1 has been defined using some different approaches that, we suppose, are contained into the knowledge base arranged according the DPSW ontology. Using the proposed ontology, the experimentalist can generate some novel workflows composed of both piece of well know techniques and some processes previously stored as instances of DPSW ontology. As future work, we will use this ontology for building an expert system for making reasoning in the analyzed case of study.

References

1. Chandrasekaran B, Josephson J, et al. (1999) What are ontologies, and why do we need them?. IEEE Intelligent Systems and Their Applications 14(1):20-26
2. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584. doi: 10.1093/nar/30.7.1575
3. Fiannaca A, La Rosa M, et al. (2012) An ontology design methodology for Knowledge-Based systems with application to bioinformatics. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) :85. doi: 10.1109/CIBCB.2012.6217215
4. Hull D, Wolstencroft K, et al. (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34:W729-32
5. Hibbe RK, Koyuturk M, Chance M R (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput. Biol. 6(1)
6. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol. 3:88. doi: 10.1038/msb4100129
7. Ucar D, Parthasarathy S, Asur S, Wang C (2005) Effective Pre-processing Strategies for Functional Clustering of a Protein-Protein Interactions Network. 5th IEEE Symposium on Bioinformatics and Bioengineering 129: 371-382. doi: 10.1109/BIBE.2005.25