

## Using ontology to mine and classify Li-Fraumeni Syndrome patients

Ricardo Moura Sekeff Budaruiche and Renata Wassermann (Advisor)

Institute of Mathematics and Statistics (IME) – University of Sao Paulo , Brazil  
{sekeff,renata}@ime.usp.br

The Li-Fraumeni Syndrome (LFS) is a syndrome that causes multiple primary tumors in children and young adults. The main motivation of this work is to create a single integrated system that allows doctors and researchers from the A.C. Camargo Cancer Center to relate family histories, clinical and molecular data present in different databases through an innovative data integration methodology in order to improve the existing LFS diagnose criteria, or even to propose a new set of clinical criteria.

The idea is to create a computational environment that enables integration between the various heterogeneous databases at A.C. Camargo Cancer Center. This will be done through a set of ontologies developed to represent knowledge about LFS, and patients' family relationships (genealogy). The ontologies will serve as a knowledge base for a data mining system which will extract information to form a new incremental knowledge base on LFS.

The A.C. Camargo Cancer Center has several databases, some of them are legacy databases from other EHR (Electronic Health Record), while others are the result of alternative methods for solving specific problems. All these database contain medical data of patients, which can be decisive in the diagnosis of the LFS and in the discovery of new clinical criteria. Thus, it is important to extract the complete medical data of patients by fetching data in all possible databases. Moreover, one of the goals to be achieved is to discover new knowledge about the clinical criteria of the syndrome, allowing to improve the existing criteria or to discover other factors that can result in a new set of clinical criteria.

Despite some efforts to integrate medical data using ontologies or even extract knowledge from heterogeneous databases have been reported, there are few solutions already developed for this purpose. [3] used some classifiers for mining breast cancer data of patients in the Oncology Institute of Lithuanian University of Health Sciences and answer some questions such as "what factors influence the diagnosis of BRCA1 gene mutation". The methodology used only data mining algorithms to test the prediction of BRCA1 mutation for breast cancer. In [2], the proposal was to mine blood biomarkers data from multiple data sources (the Oncomine, the Ingenuity Pathway Analysis program - IPA and data from blood samples) in order to predict six common types of cancer. This solution does not use ontologies to store the knowledge acquired. [4] suggests the use of ontologies for integrating heterogeneous data by means of a semiautomatic text search tool in metadata, and which will subsequently be human validated. Finally, [1] used a hybrid approach to demonstrate the feasibility of integrating heterogeneous databases of cancers involving p53.

In contrast, our proposal makes use of ontologies, both to integrate the databases and to aid the process of data mining and knowledge discovery.

In the next two years, the following steps are expected:

1. Mapping of information about LFS available in the A.C. Camargo Cancer Center databases and defining a solution architecture to the problem of data integration.
2. Building the genealogy and Li-Fraumeni ontologies.
3. Development of the computer environment and triplestore for database integration.
4. Definition of classifiers and the ontology based data mining tool.
5. Data loading and execution of data mining.
6. Evaluation of the results and preparation for the final reports. Involvement of medical staff in the validation of results.

At this moment, we are constructing the Li-Fraumeni Ontology, the Genealogy Ontology and mapping some clinical data which will be used for mapping the patients' data from database to the triplestore. At the same time, we are preparing the computational environment that will host the triplestore and all the machinery involved in the inference and knowledge extraction. Thus, we are currently working in steps 1, 2 and 3 of the schedule. Some issues we are facing at the moment are: (1) decide when two patients of different families are the same person; (2) what data are useful to be imported (the more data we import, the more accurate the inference although it will take longer to import) and; (3) decide which reasoner will be used for inference and data mining.

The main contribution of this work will be a set of ontologies that will form a knowledge base about LFS and that will help researchers of A.C. Camargo Cancer Center to classify patients for experiments concerning the LFS. With family members and LFS data, we will be able to extract new knowledge about the clinical criteria used in the diagnosis of this syndrome in order to provide a faster and more reliable diagnosis. Also, as the LFS is an inherited disease, the genealogy ontology will be easily adapted to help the study of other hereditary syndromes.

## References

1. Bichutskiy, Vadim Y. and Colman, Richard and Brachmann, Rainer K. and Lathrop, Richard H.: Heterogeneous biomedical database integration using a hybrid strategy: a p53 cancer research database. *Cancer Informatics* 949, 277-287 (2006)
2. Yang, Yongliang and Pospisil, Pavel and Iyer, Lakshmanan K and Adelstein, S James and Kassis, Amin I: Integrative genomic data mining for discovery of potential blood-borne biomarkers for early diagnosis of cancer. *PLoS One* 11 (3), issn 1932-6203, e3661 (2008)
3. Niakšu, O and Gedminait, J and Kurasova, Olga: Data mining approach to predict BRCA1 gene mutation. *Computational Science and Techniques* 2 (1), (2013)
4. Gelernter, Judith and Lesk, Michael: Use of Ontologies for Data Integration and Curation. *International Journal of Digital Curation* 1 (6), issn 1746-8256, 70-78 (2011)