Contribution ID: **161** Type: **not specified**

A BioInformatics Framework for NGS Immunogenetics Data (BIND)

Wednesday, 18 September 2013 14:50 (10 minutes)

IG/TR repertoire analysis can offer biological and clinical information with important implications for understanding normal and pathologic immune conditions. BIND will delineate a methodology and make available tools that allow for detailed repertoire analysis based on NGS data. Additionally, BIND will set the basis for the detection of disease-specific IG/TR molecular properties that may later be used for early diagnosis of diseases, starting from cancer, but also extensible to autoimmune, and other diseases. The methodological approach and the toolbox that will be produced aim to constitute a reference in the field, applicable in other scenarios as well. Additionally, the toolbox will be made available for research and educational applications in the field, and will be in a form allowing extensibility and further advancements.

In particular, we will focus on the systematic and quantitative processing of the output produced from the analysis of the NGS data by algorithms and tools of the International ImMunoGeneTics information system (IMGT), the global reference on Immunoinformatics. IMGT/HighV-QUEST is the high-throughput tool for the analysis of thousands of immunoglobulin (IG) and T cell receptor (TR) rearranged nucleotide sequences (up to 150,000 sequences) per run (http://www.imgt.org/). The standard output is a series of text files with information per sequence on the recognition of genes involved as well as the aminoacids and nucleotides of specific areas of interest.

The huge amount and the complexity of the data involved establish a set of requirements ideally covered through a Gridinfrastructure. To this end, the design and implementation of the process have the particular attributes of the Grid environment in mind, especially as regards the execution of task pipelines on massive data, employing opensource programming tools. Our proposed approach is as follows:

• A reading and information management framework of the IMGT-VQuest report. The challenge here pertains to the large amount of data, as well as to the heterogeneity of data types, ranging from simple values, and strings, to DNA sequences, that need to be simultaneously handled, tackling big data issues.

• Preprocessing and Filtering of the IMGT report data, to exclude potentially erroneous information. This will include a series of rules that can be standardized and reused as a tool, based on existing and ongoing experience as regards the NGS readings.

• Intrasubject Repertoire and statistical analysis of the IMGT report data, for the identification of clonality and diversity. This will be based on state-of-art methods.

• Intersubject clustering based on the dominant repertoire expression. Additional clustering methods based on a variety of similarity measures from the biomedical informatics domain will be explored.

• Integration of the proposed functionality and deployment as a toolbox.

• Testing and evaluation of the proposed functionality with expert knowledge and existing data made available via the EuroClonality group (http://www.euroclonality.org/).

Presenters: HADZIDIMITRIOU, Anastasia; CHOUVARDA, Ioanna; STAMATOPOULOS, Kostas

Session Classification: Scaling up life sciences with grids and clouds - stories and recommendations