Cytometry-Ontology Framework

Progress Report

March 6, 2013

Current status

So far we have achieved the following goals:

- Writing a parser that can identify markers in a given immunophenotype using R.
- Making an "ask" query for each of the markers to check its existence written in Python (part of the code was borrowed from Cliburn's).
- Preparing the final query based on the existent markers

We have produced prototype queries and executed them against the Cell Ontology to ensure suitability of the resource. Following this, we established different ways of interacting with the cl.owl file, either locally or remotely. We decided that querying a remote server was more sustainable than updating local files, and allows for greater flexibility in term of access. A triplestore has been setup and provide a SPARQL endpoint at `http://cell.ctde.`

net:8080/openrdf-sesame/repositories/CL for programatic access. We chose to rely on the OWLIM (http://www.ontotext.com/owlim) system. While the CL provides a pre-reasoned file for download, it did not work well. OLWIM provides built-in inference, allowing for example for forward chaining and therefore addition of necessary restrictions. For example, a query for those cells that *has part some CD 19* retrieves only the 4 classes for which the restriction is explicitly stated. With OWLIM, the reasoner adds the same restriction on their subclasses, yielding a result set of 95 classes instead.

We also decided to add a block to the previously designed pipeline (presented in the framework) to check for the existence of each marker in a given immunophenotype. For instance, for $CD19^{+}CD100000^{-}$, we first check the existence of $CD19$ and $CD100000$ before the final query is done. Otherwise, the results are likely to be a null set since there is no $CD100000$.

Next steps

- Ensure, in coordination with CL and PRO developers that necessary information (specifically synonyms) are present in the Cell Ontology to allow retrieval of markers based on their label

- Warn users when they are querying for a marker that can not be found in the ontology. This could result from missing information in CL or from lack of synonyms in the resource in which case we would want to feedback this to the CL developers. It could also help identify an issue on the user side, such as a typo.

- Further test the triplestore to assertain whether all needed restrictions are being added by the reasoner.
- To implement the *confidence level calculator* (previously presented in the framework) to find the part of the ontology structure that represents the cell population of interest with a high degree of confidence
- Add the *immunophenotype level* to the confidence calculator
- To change the Python code to R code to make it pure R by means of current R packages for SPARQL query