

Information Extraction for Bioinformatics: Semi-automatic discovery, disambiguation and storage of protein-related abbreviations in scientific literature

- **finding and identifying abbreviations** (for now, our scope is restricted to proteins):
 - within scientific literature (articles publicly available on web repositories, or via the Faculty's licenses);
 - on the web in generalSemi-automatic process:
 - we could start from a well-known abbreviation catalogue (if there is one);
 - we then need to discover abbreviations within the **full text** of scientific articles and identify their nature
 - with some metrics, like their "distance" from a certain term referring to a protein, or similar/more advanced techniques;
 - ontologies might come in handy and must be considered;
 - the protein name/terms can be found within the **protein databases**, so we'll need to access them as well;
 - we need to **disambiguate** identical abbreviations each referring to a different protein (how can this be achieved? Via some proper techniques, as in the case above)
 - we need to allow users to provide **feedback** during the automatic steps, in order to double-check the correct identification and disambiguation of the discovered abbreviations or to **manually specify** the corresponding proteins, if known, when the automatic processes fail;
- **storing the identified abbreviations in a database**
 - we need to create a repository for protein-related abbreviations, which could theoretically/hopefully be improved by subsequent/periodic runs of the application and/or by the scientific community
- **technologies to be used** (at the very least):
 - Java
 - PostgreSQL/MySQL
 - ...