Mining the Scientific Literature for and Annotating Proteomics Software using the EDAM ontology and biotoolsXSD

Over the last two decades, hundreds of academic groups and individual researchers have published or made available software for performing particular operations on proteomics data. As the field is competitive and some work is done in parallel by different groups, the publications and software vastly outnumber the different types of operations, data formats and topics. To bring some order to this chaos, the EDAM[1] (from "EMBRACE Data and Methods") ontology of established and prevalent concepts within bioinformatics, including types of data and data identifiers, data formats, operations and topics, was created.

An ontology is simply a set of terms with synonyms and definitions, organized into an intuitive hierarchy for convenient use by curators, software developers and end-users. Whereas ontologies themselves are maintained and agreed upon by experts in the field, anyone is free to use them. Currently (2016) a large number of terms from mass spectrometry, proteomics and metabolomics are being added to the EDAM ontology, enabling annotation of software tools in these fields.

While ontologies provides structure, they are not in themselves of much use. However, annotating objects, such as data or software, using an ontology infers relationships based on higher levels in the ontological hierarchy. These relationships may be of use in assisting assembly and validation of data analysis pipelines combining several tools and passing data in well-defined formats between the individual components in the pipeline. But before this can become a reality, already existing tools should be properly annotated and manually curated to ensure quality annotations. This requires mining the scientific literature and the World Wide Web for software, checking if it is still available, and annotating it with respect to programming language, software license, user interface (a controlled vocabulary for which is defined in biotoolsXSD[2]), and data formats, operations and topic (all part of EDAM). To provide reliable, consensus annotations, the students will independently annotate the software or cross-validate their respective annotations, with disagreements warranting further investigation. As the goal is for the annotations to be practically useful in building data analysis workflows, emphasis is placed on practically important details such as input and output data formats.

The project builds upon ongoing work that has produced basic annotation on some 200+ tools. The expected project output will include at least 300 new tools registered in the ELIXIR Tools & Data Services Registry[3], annotated to sufficiently high standard as to serve as an exemplar for other Registry providers. Assuming satisfactory progress is made, the work will be published in a proteomics-focussed journal.

Attention to detail, good command of the English language, and a general familiarity with proteomics bioinformatics software are required. All supervision will be in English and by Dr. Magnus Palmblad at the LUMC Center for Proteomics and Metabolomics and Jon Ison at DTU, Denmark.

1. Ison, J. *et al.*, 2013, EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10): 1325-1332.
2. https://github.com/bio-tools/biotoolsxsd
3. Ison, J. *et al.*, 2015, Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*, 44(D1): 38-47.

Contact info: Magnus Palmblad, n.m.palmblad@lumc.nl (tel: +31 71 526 9582) cc Jon Ison jison@cbs.dtu.dk.