**Summer Project: Validating PattArAn triples using literature imprints (sentences)**

**For the bioscientist: Use PattArAn to explore your research question.**

**For the computer scientist: Refine the PattArAn validation procedure.**

The exploration process begins with the bioscientist identifying a set of genes and/or controlled vocabulary terms of interest. Typically these would be used to generate a set of triplets X as follows:

(GO term1, gene, GO term2)

(GO term, gene, PO term)

(MeSH term1, gene, MeSH term2), etc.

The bioscientist helps an undergraduate research assistant to identify a set of documents that relate as closely as possible to the ideas being explored, and therefore to the set of triplets X. The bioscientist will help the student to mark up the documents. The markup is at the sentence level and identifies sentences that are most relevant and important to each triplet in X or key components of a triplet, for example sentences about (GO term1, gene, GO term 2). We would also be interested in sentences about a fragment of the triplet (GO term, gene) or (gene, PO term). In general, the sentences should be those that are most closely related to the underlying biological phenomenon.

Next, the computer scientist will identify new potentially related documents (where possible fulltext) and predict relevant sentences from these documents. Ideally we will find sentences that mirror the ideas (if these indeed exist) or that contribute as much as possible to the further development of the ideas. These will be returned to the bioscientist for validation. This process may iterate as long as deemed appropriate by the bioscientist.

**Outcomes:** Under the best of circumstances, a) the bioscientist benefits from the sentences retrieved by PattArAn b) our methods for identifying related documents and sentences improves because of the training data provided by the bioscientist.

**Responsibilities for the bioscientist:**

Recruit an undergraduate student who will spend 4 weeks in the summer, possibly from June 11 2012. The PattArAn project will provide financial support for the student assuming that they are UMD undergraduate students. We can also support an under-represented student from outside campus.

Meet with the student and the PattArAn team at the beginning of the exercise; this will be a virtual meeting.

Provide the PattArAn team with the set X and corresponding initial documents. Provide some initial feedback to the student on identifying relevant markup sentences.

Be available by email in June to answer student questions that the PattArAn team may not be able to cover.

Spend about 3 hours at the end of the 4 week period to review the training data (sentences).

**PattArAn CS Team**

Joseph Benik [jvb@umiacs.umd.edu](mailto:jvb@umiacs.umd.edu)

Louiqa Raschid [louiqa@umiacs.umd.edu](mailto:louiqa@umiacs.umd.edu)

Padmini Srinivasan [padmini-srinivasan@uiowa.edu](mailto:padmini-srinivasan@uiowa.edu)

Xiao-Ning Zhang [xzhang@sbu.edu](mailto:xzhang@sbu.edu)