Title: **Supporting Triage of PubMed Abstracts for mycoCLAP**

Authors: Marie-Jean Meurs, Erin McDonnell, Ingo Morgenstern, Greg Butler, Justin Powlowski and Adrian Tsang

Institution: Centre for Structural and Functional Genomics, Concordia University, Montréal, QC, Canada

Abstract:

Fungi secrete a variety of enzymes that work efficiently to degrade lignocellulosic biomass. Since the breakdown of lignocellulose is vital for a number of industrial processes that stand to be improved, interest in these enzymes is high.

Fungal lignocellulose-degrading enzymes are numerous and display a wide range of characteristics and properties. As such, the curation of those that have been characterized into a searchable database is essential for supporting further research into their potential applications and uses. To date, several different types of fungal lignocellulose-degrading enzymes have been manually curated and deposited into an online, searchable database called mycoCLAP (Characterized Lignocellulose-Active Proteins of Fungal Origin) [http://mycoclap.fungalgenomics.ca].

The curation process for mycoCLAP involves a number of steps. Typically the most time-consuming one involves finding appropriate papers to curate. Literature searches often recover a large number of candidate articles which then need to be manually screened for relevance by the curator (triage task).
A reliable automated screening and filtering approach of the candidate articles would save precious time.

The work we present supports the automatic triage of PubMed abstract candidates.

The developed processing resource relies on the GATE-based mycoMINE text mining system.

It takes PubMed abstracts as input, then classifies these abstracts according to their potential for full paper curation for mycoCLAP. The inference engine used for making the classification decision is based on first order logic rules combining constraints based on the document topic with constraints based on presence of entities or concepts in smaller units of text.

A first evaluation was performed on 104 PubMed abstracts published from 11.01.2012 to 01.30.2013 and retrieved by keyword search for fungal oxidoreductase; lignin, versatile and manganese peroxidase; pyranose oxidase; glyoxal oxidase. The system selection was checked against manual triage, validating our approach with these results: precision=0.68, recall=0.79, true negative rate=0.83, and accuracy=0.88.