Supporting Triage of PubMed Abstracts for mycoCLAP

Marie-Jean Meurs, Erin McDonnell, Ingo Morgenstern, Greg Butler, Justin Powlowski, Adrian Tsang
Centre for Structural and Functional Genomics, Concordia University, Canada

mycoCLAP

Searchable database of sequenced, characterized lignocellulose-active proteins of fungal origin [1]

Steps by steps:
1. PubMed abstract selection based on curator’s keywords
2. Abstract sorting by relevance stated by mycoMINE [2]
3. Full paper reading by curators for good candidates using augmented browsing integrating mycoMINE results
4. Manual curation of relevant full papers by curators
5. Record of new mycoCLAP entries

Automatic Triage of PubMed abstracts
- Relies on mycoMINE text mining system
- Input = set of PubMed abstracts
- Output = classification decisions [accepted/rejected] + extracted topics and entities
- Inference engine: first order logic rules → document topic + presence of entities or concepts

Evaluation
- 104 PubMed abstracts from 11.01.2012 to 01.30.2013
- Retrieved by keyword search for:
  - fungal oxidoreductase;
  - lignin, versatile and manganese peroxidase;
  - pyranose oxidase;
  - glyoxal oxidase.

Constraints on Topics and Entities:
- enzyme characterization;
- protein expression;
- specific activity;
- activity assay conditions;
- substrate specificity;
- fungus, enzyme.

System selection checked against manual triage

Results
- precision 0.68
- recall 0.79
- true negative rate 0.83
- accuracy 0.88

Acknowledgment
Funding: Genome Canada, Génome Québec
Collaboration: Sherry Wu, Min Wu
Technical support: Andrei Wasylyk

References
- [1] Murphy et al., Curation of characterized glycoside hydrolases of fungal origin, Database, 2011