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]

***myco*CLAP - Characterized Lignocellulose-Active Proteins of Fungal Origin**

Home | Search | Downloads | Correction | New Entry | BLAST | Useful Links | Internal User (Login) | About Us | Help

Search results of ''

Entry Name	Species	Enzyme Name	Host For Recombinant Expression	Substrates	Assay	Specific Activity	pH Optimum	Temperature Optimum (°C)	EC number	Protein ID (Genbank)
ABF51A_ASPAW	Aspergillus awamori	alpha-arabinofuranosidase	native	pNP-alpha-L-arabinofuranoside	pNP-releasing Assay	391 U/mg	4	60	3.2.1.55	BAB21568
			native	arabinan	Somogyi-Nelson Assay	active				
			native	debranched arabinan	Somogyi-Nelson Assay	active				
			native	arabinoxylan	Somogyi-Nelson Assay	active				
			native	arabinogalactan	Somogyi-Nelson Assay	active				
			native	gum arabic	Somogyi-Nelson Assay	active				
ABF51A_ASPKA	Aspergillus kawachii	alpha-arabinofuranosidase	native	pNP-alpha-L-arabinofuranoside	pNP-releasing Assay	active	4	55	3.2.1.55	BAB96816
ABF51A_ASPNG	Aspergillus niger	alpha-arabinofuranosidase	native	pNP-alpha-L-arabinofuranoside	pNP-releasing Assay	active	3.4	46	3.2.1.55	AAC41644
			Aspergillus nidulans	pNP-alpha-L-arabinofuranoside	pNP-releasing Assay	active				

> Manual curation <

Curation Process

Steps by steps:

1. PubMed abstract selection based on **curator**'s keywords

↓

2. Abstract sorting by relevance stated by **mycoMINE** [2]

↓

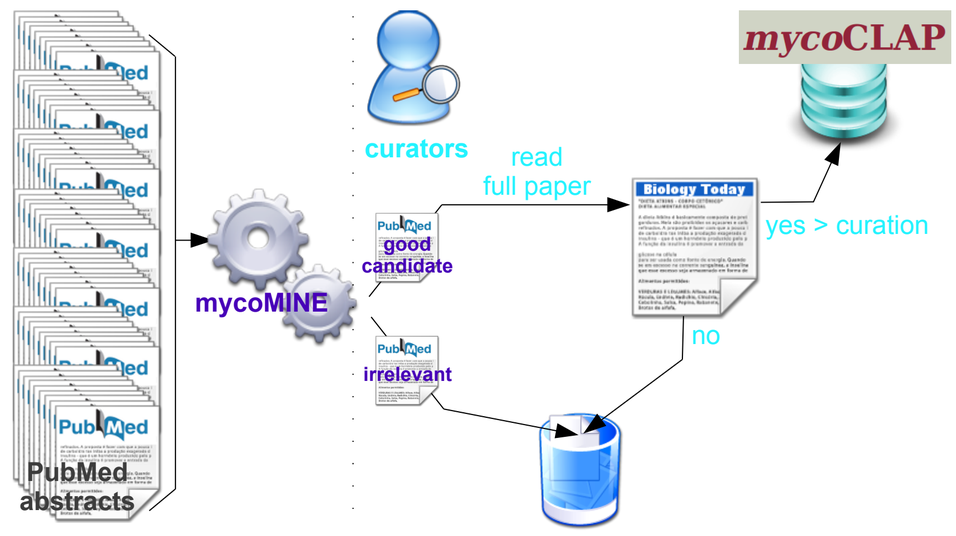
3. Full paper reading by **curators** for good candidates using augmented browsing integrating **mycoMINE** results

↓

4. Manual curation of relevant full papers by **curators**

↓

5. Record of new *myco*CLAP 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

Curator Support

	manual curation	candidate	fungus	enzyme	characterization	expression	specific activity	activity assay conds	substrate specificity
1		10742277	Aspergillus oryzae, Geotrichum candidum	peroxidase, alpha-amylase	-	v	-	-	v
2		11590604	Trametes versicolor	peroxidase	-	-	v	-	-
3		12396114	Trametes versicolor	manganese-independent peroxidase, laccase	-	-	v	-	-
4		12396115	Pleurotus ostreatus, Phanerochaete chrysosporium	laccase, peroxidase	-	-	-	-	v
5		12698279	Termitomyces albuminosus	oxidase, peroxidase, transcription polymerase	-	-	v	-	-
6		14684913	Thanatephorus cucumeris, Aspergillus oryzae	peroxidase	-	v	-	-	-
7		15158509	Phanerochaete chrysosporium, Coriolus versicolor	laccase, peroxidase	-	-	v	-	-
8		15313183	Thanatephorus cucumeris, Aspergillus oryzae	peroxidase	v	v	-	-	v

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

[2] Meurs et al., *Semantic text mining support for lignocellulose research*, BMC MIDM, 2012