

IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences

Bahar Sateli Marie-Jean Meurs Greg Butler
Justin Powlowski Adrian Tsang René Witte

Concordia University, Montréal, QC, Canada



Semantic Software Lab



NETTAB 2012

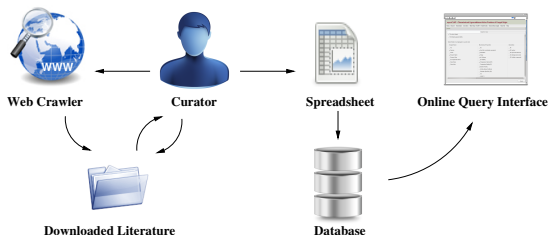
Nov. 15th, Como, Italy

Outline

- 1 Introduction
- 2 System Architecture
- 3 User Interface
- 4 Application
- 5 Evaluation
- 6 Conclusion

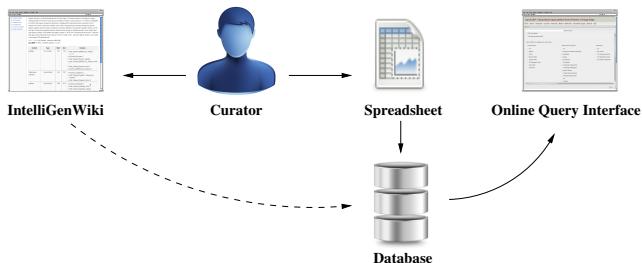
MOTIVATION: Curation of Biomedical Literature

- ▶ Finding and extracting relevant knowledge from the domain literature
- ▶ Manually refining and updating bioinformatics databases



- ▶ Manual literature curation is
 - ▶ **Expensive** → requires domain experts
 - ▶ **Labour-intensive** → ever growing amount of scientific publications
 - ▶ **Error-prone** → critical knowledge can be easily missed

APPROACH: IntelliGenWiki



Enhanced Literature Curation Workflow Using IntelliGenWiki

- ▶ Text mining techniques integrated within the wiki environment
- ▶ Novel Human-AI collaboration patterns
- ▶ Producing semantic metadata
- ▶ Transform text into knowledge base

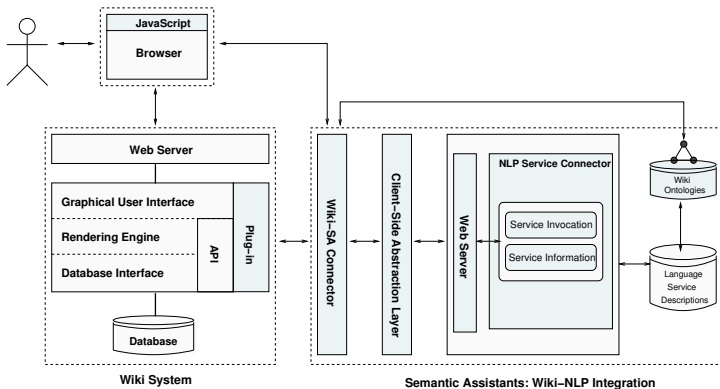
APPROACH: IntelliGenWiki

- ▶ Adopts the “Wiki” paradigm
 - ▶ Accessible via a web browser
 - ▶ Simple syntax (markup)
 - ▶ Open collaboration
- ▶ Based on the MediaWiki engine
 - ▶ Open source
 - ▶ Highly scalable
 - ▶ Extensible: Semantic MediaWiki
- ▶ Integrated Text Mining *Assistants*
- ▶ Provides semantic capabilities
 - ▶ Formalization of knowledge
 - ▶ Producing machine-readable content
- ▶ Open source software (AGPL3)

IntelliGenWiki User Interface

System Overview

- **Front-end:** Semantic MediaWiki
- **Back-end:** Wiki-NLP Integration [Sateli and Witte, 2012]
 - Comprehensive architecture based on the Semantic Assistants Framework [Witte and Gitzinger, 2008]
 - Seamless integration of various NLP capabilities *within* a wiki environment



IntelliGenWiki Pages

- ▶ Each wiki page corresponds to a literature instance, e.g., abstract of a paper
- ▶ Revision History
- ▶ Inquire text mining services via wiki toolbox

Wiki Toolbox

toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Semantic Assistants
- Browse properties

PubMed:20709852 - IntelliGenWiki - Mozilla Firefox

PubMed:20709852 - IntelliGenWiki

Wikisysop my talk my preferences my watchlist my contributions log out

page discussion edit history delete move protect watch refresh

PubMed:20709852

Title: Characterization of a cellobiohydrolase (MoCel6A) produced by *Magnaporthe oryzae*.

Authors: Takahashi M, Takahashi H, Nakano Y, Konishi T, Terauchi R, Takeda T.

Institute: Iwate Biotechnology Research Center, Kitakami, Iwate, Japan.

PMID: 20709852

Received on March 10, 2010. Accepted on July 30, 2010.

Full Text

[edit]

Abstract

Three GH-6 family cellobiohydrolases are expected in the genome of *Magnaporthe grisea* based on the complete genome sequence. Here, we demonstrate the properties, kinetics, and substrate specificities of a *Magnaporthe oryzae* GH-6 family cellobiohydrolase (MoCel6A). In addition, the effect of cellobiose on MoCel6A activity was also investigated. MoCel6A continuously fused to a histidine tag was overexpressed in *M. oryzae* and purified by affinity chromatography. MoCel6A showed higher hydrolytic activities on phosphoric acid-swollen cellulose (PSC), β -glucan, and cellobiosaccharide derivatives than on cellulose, of which the best substrates were cellobiosaccharides. A tandemly aligned cellulose binding domain (CBD) at the N terminus caused increased activity on cellulose and PSC, whereas deletion of the CBD (catalytic domain only) showed decreased activity on cellulose. MoCel6A hydrolysis of cellobiosaccharides and sulforhodamine-conjugated cellobiosaccharides was not inhibited by exogenously adding cellobiose up to 438 mM, which, rather, enhanced activity, whereas a GH-7 family cellobiohydrolase from *M. oryzae* (MoCel7A) was severely inhibited by more than 29 mM cellobiose. Furthermore, we assessed the effects of cellobiose on hydrolytic activities using MoCel6A and *Trichoderma reesei* cellobiohydrolase (TrCel6A), which were prepared in *Aspergillus oryzae*. MoCel6A showed increased hydrolysis of cellobioacetate used as a substrate in the presence of 292 mM cellobiose at pH 4.5 and pH 6.0, and enhanced activity disappeared at pH 9.0. In contrast, TrCel6A exhibited slightly increased hydrolysis at pH 4.5, and hydrolysis was severely inhibited at pH 9.0. These results suggest that enhancement or inhibition of hydrolytic activities by cellobiose is dependent on the reaction mixture pH.

PMID: 20709852 [PubMed - indexed for MEDLINE] PMID: PMC2950481 Free PMC Article

This page was last modified on 6 November 2012, at 23:21. [IntelliGenWiki](#) [Disclaimers](#) [Privacy policy](#) [About](#)

Done

Paper
Information

Paper
Content

The NLP Interface

- ▶ The IntelliGenWiki NLP user interface offers various text mining services
- ▶ Customizing services at runtime
- ▶ Dynamically-generated interface

Text Mining Assistants inside the wiki

The screenshot displays the IntelliGenWiki NLP user interface. At the top, there is a navigation bar with tabs: 'Available Assistants', 'Results Target', 'Global Settings', and 'Console'. Below this, a message states: 'Step 1. Select the service your wish to execute on your collection. Once you add this page to your collection, you can continue browsing as your collection is saved.'

The 'Available Assistants' dialog box is open, showing a list of services: 'mycoMINE', 'IR Information Extractor', 'Information Extractor', and 'OrganismTagger'. The 'Collection' input field is empty, and the 'Add' button is visible. A purple arrow points from the 'Available Assistants' dialog box to the 'Collection' input field.

On the right side of the interface, there is a sidebar with a 'navigation' menu containing links: 'Main page', 'Community portal', 'Current events', 'Recent changes', 'Random page', and 'Help'. Below the sidebar, there is a section titled 'PubMed:20709852' with a 'Full Text' link and an 'Abstract' section. The abstract text is partially visible on the right side of the interface.

NLP Interface features

► Multi-document Analysis

Available Assistants Results Target Global Settings Console

Step 1. Select the service your wish to execute on your collection.
Once you add this page to your collection, you can continue browsing as your collection is saved.

Available Assistants Select a service

Runtime Parameters Select a service

- mycoMINE
- IR Information Extractor
- Information Extractor
- OrganismTagger

Collection <http://loompa.cs.concordia.ca/.../PubMed:19912637>
<http://loompa.cs.concordia.ca/.../PubMed:2186187>

Add Clear

► Flexible handling of results

- Writing to the same page as the resource
- Writing to a different page in the wiki
- Writing to an external wiki

► Dynamic discovery of NLP services

Available Assistants Results Target Global Settings Console

Optionally, you can select the server you would like to connect to. Select a server from the list, press "Connect" and refresh the page using your browser.

☒ Predefined Servers

☐ Custom Server

minion.cs.concordia.ca:8879
 minion.cs.concordia.ca:8879
 minion.cs.concordia.ca:2011

Connect

Information Extraction

- ▶ Automatically extracting knowledge from text
- ▶ Various IE services
 - ▶ mycoMINE
 - ▶ OrganismTagger
 - ▶ Open Mutation Miner
 - ▶ ...
- ▶ Enrichment of literature content with semantic markup

Example:

`[[hasType::Enzyme|cellobiohydrolase]]`

severely inhibited at pH 9.0. These results suggest that enhancement or inhibition of hydrolytic activities by cellobiose is dependent on the reaction mixture pH.

PMID: 20709852 [PubMed](#) - indexed for MEDLINE] PMCID: PMC2950481 [Free PMC Article](#)

mycoMINE on PMID_20709852_Abstract [View](#)

Content	Type	Start	End	Features
cellobiohydrolase	Enzyme	103	120	<ul style="list-style-type: none"> ■ enzyme_alias: cellobiohydrolase ■ BRENDA_SystematicName: oligoxyloglucan reducing-end cellobiohydrolase ■ BRENDA_EcNumber: 3.2.1.150 ■ abbreviation_alias: - ■ google_search: http://www.google.com/search?q=cellobiohydrolase ■ BRENDA_RecommendedName: oligoxyloglucan reducing-end-specific cellobiohydrolase ■ SwissProt_ID: - ■ BRENDA's page: http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.2.1.150
Magnaporthe oryzae	Organism	143	161	<ul style="list-style-type: none"> ■ NCBI_Taxonomy_WebPage: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=318829&mode=info ■ organism_scientific_name: Magnaporthe oryzae ■ organism_alias: Magnaporthe oryzae ■ google_search: http://www.google.com/search?q=Magnaporthe+oryzae ■ NCBI_Taxonomy_ID: 318829

Found Entity

Entity Type

Entity Location

NLP-Provided Additional Information

Semantic Entity Retrieval

- ▶ Unadorned wikis offer only keyword-based search
- ▶ What if we want to *discover* what's contained in the wiki?
 - ▶ e.g., “Which papers in this wiki mention an enzyme entity in their text?”
- ▶ **Solution:** Querying the semantic metadata in the wiki
 - ▶ Search the wiki by semantic properties, e.g., entity *type*, generated by NLP services
 - ▶ Using special Semantic MediaWiki markup, called *inline queries*

```
{{#ask: [[hasType::Enzyme]]
|?Enzyme = Enzyme Entities Found
|format = table
|headers = plain
|default = No pages found!
|mainlabel = Page Name
}}
```

Property:Enzyme

Page Name	Enzyme Entities Found
PMID: 20709852	Cellobiohydrolase Cellulases endoglucanases β-glucosidases Invitrogen DNA polymerase

User Study

- ▶ Is the integration of text mining assistants in a wiki environment actually effective?
- ▶ User study within the Genozymes project context (www.fungalgenomics.ca)
 - ▶ **Goal:** Identifying and characterizing fungal enzymes
 - ▶ **Dataset:** 30 documents
 - ▶ **Users:** 2 expert biocurators
 - ▶ **NLP Service:** mycoMINE [Meurs et al, 2012]
 - ▶ **Measure:** Time spent on curation
 - ▶ **Method:** Comparison against time spent on manual curation

Average Curation Time

- ▶ Results:

Abstract Selection		Full Paper Curation	
no support	IntelliGenWiki	no support	IntelliGenWiki
1 min.	0.3 min.	37.5 min.	30.6 min.

- ▶ **Conclusion:** IntelliGenWiki was indeed efficient and reduced the paper selection and curation time by almost **70%** and **20%**, respectively.

Conclusion

What you can do now

- ▷ Install MediaWiki and Semantic MediaWiki extension
- ▷ Download and deploy the Wiki-NLP integration
- ▷ Use the existing text mining services in our public server
- ▷ Alternatively, setup your own Semantic Assistants services developed based on the GATE framework

What is next

- ▷ Cover other tasks, e.g.,
 - ▶ Quality assessment
 - ▶ Paper recommendation
 - ▶ Personalization
- ▷ Develop services for automatic import of literature, e.g., from PubMed
- ▷ Query the RDF in wiki from external applications

More Information

<http://www.semanticsoftware.info/intelligenwiki>

Acknowledgment

- ▶ Funding for this work was provided by NSERC, Genome Canada and Génome Québec.
- ▶ Caitlin Murphy and Sherry Wu, biocurators at the Centre for Structural and Functional Genomics (CSFG) at Concordia University, are acknowledged for their participation in the evaluation task.