Semantic Text Mining for Lignocellulose Research

Marie-Jean Meurs, Caitlin Murphy, Ingo Morgenstern,
Nona Naderi, Greg Butler, Justin Powlowski,
Adrian Tsang and René Witte

Université Concordia University

Department of Computer Science and Software Engineering
Centre for Structural and Functional Genomics
Department of Biology and Department of Chemistry and Biochemistry
Concordia University, Montréal (QC), Canada

GENOZYMES

DTMBIO 2011

Oct. 24th, Glasgow

Introduction

- Context
- Biofuel Process
- Knowledge sources

Project Context and System Architecture

Text Mining Pipelines

Evaluation

Conclusion

mycoMINE: a semantic infrastructure supporting biofuel research

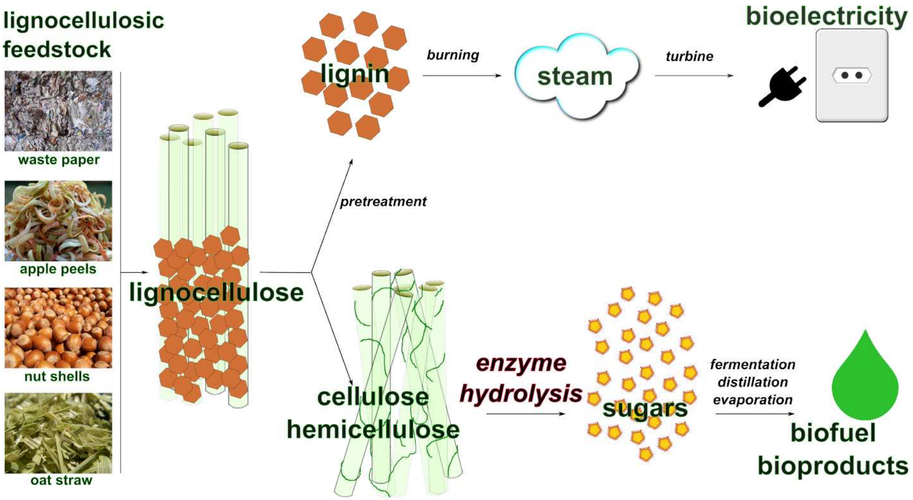
Automated curation of available knowledge on fungal enzymes

- curation of the existing literature
- development of ontological NLP pipelines
- integration through Web-based interfaces

Goals

- spending less time to mine the literature for facts
- being provided with richer and semantically linked information

Biofuel Process



Biofuel

Fuels produced from biomass (lignocellulose)

- sustainable liquid fuels with low environmental impact
- promising alternative to fossil based fuels

From cellulose to biofuel

- extraction of sugars requires to break down cellulose
- degradation of cellulose uses specific molecules called enzymes

⇒ discovering efficient enzymes is a key challenge

Knowledge sources

- ever growing documentation on fungal enzymes
- many knowledge sources

Example: the PubMed bibliographic database

- 19 million citations from over 21,000 life science journals
- linked to other databases like:
  - *Entrez Genome* provides access to genomic sequences
  - *BRENDA, The Comprehensive Enzyme Information System*, the main collection of enzyme functional data available to the scientific community

Knowledge sources

Querying PubMed

→ collecting an often long list of relevant papers

Analyzing the collection

→ reading abstracts and sometimes full papers

A time consuming task, difficult to handle

→ significant knowledge can be missed

⇒ Natural Language Processing (NLP) and Semantic Web approaches are increasingly adopted in biomedical research.

Introduction

Project Context and System Architecture

- User Groups
- Semantic Entities
- Semantic Resources
- System Architecture

Text Mining Pipelines

Evaluation

Conclusion

User Groups

Objectives ▷ Curation of fungal genes

→ guiding research and experiments

→ filling the mycoCLAP database

- DB of fungal genes encoding biochemically characterized lignocellulose-active proteins
- `http://cubique.fungalgenomics.ca/mycoCLAP/`

Users of our system:

- *curators*: manual curation of fungal genes
- *biology researchers*: decision about the experiments to conduct
- *experimenters*: execution of the experiments

Semantic Entities

Semantic Entity	Level	Definition	Example
ActivityAssayConditions	*S*	conditions at which the activity assay is carried out	disodium hydrogen phosphate, citric acid, pH 4.0, 37℃
Assay	*W*	name of the experimental assay	Dinitrosalicylic Acid Method (Somogyi-Nelson)
Enzyme	*W*	enzyme name	alpha-galactosidase
Gene	*W*	gene name	mel36F
Glycosylation	*S*	enzymatic process attaching glycans to organic molecules	N-glycosylation
Host	*W*	organism used to produce the recombinant protein	Escherichia coli
KineticAssayConditions	*S*	buffer, pH, temp. for the kinetic parameters determination	0.1M (disodium hydrogen phosphate, citric acid), pH 4.0, 37℃
Organism	*W*	organism name	Gibberella sp.
pH	*S*	pH mentions	The enzyme retained greater than 90% of its original activity between pH 2.0 and 7.0 at room temperature for 3h.
ProductAnalysis	*S*	products formed from the enzyme reaction and identification method	HPLC, glucose, galactose
SpecificActivity	*S*	specific activity of the enzyme on the substrate	11.9U/mg
Strain	*W*	strain name	F75
Substrate	*W*	substrate name	stachyose
SubstrateSpecificity	*S*	substrate specificity mentions	The Endogluccanase from Pyrococcus furiosus had highest activity on cellopentaose.
Temperature	*S*	temperature mentions	The enzyme stability at different pH values was measured by the residual activity after the enzyme was incubated at 25℃ for 3h.

Semantic annotation types defined by the curators, applicable level (sentence, *S* or word(s), *W*), definitions and examples

Semantic Resources

– organisms –

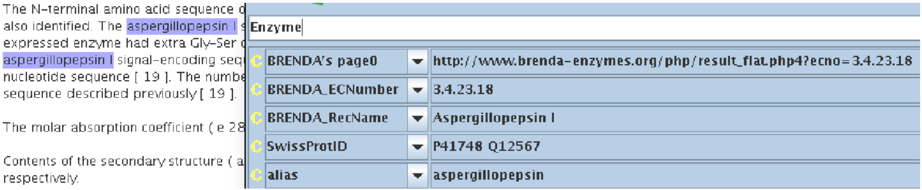
- the NCBI Taxonomy database
  - http://www.ncbi.nlm.nih.gov/Taxonomy

– enzymes –

- BRENDA
  - http://www.brenda-enzymes.org
- the UniProtKB/SwissProt database
  - http://www.uniprot.org

Semantic Resources

References to the original sources integrated into the curated data:



The N-terminal amino acid sequence o
also identified. The aspergillopepsin I s
expressed enzyme had extra Gly-Ser o
aspergillopepsin I signal-encoding seq
nucleotide sequence [19]. The numbe
sequence described previously [19].

The molar absorption coefficient (e 28

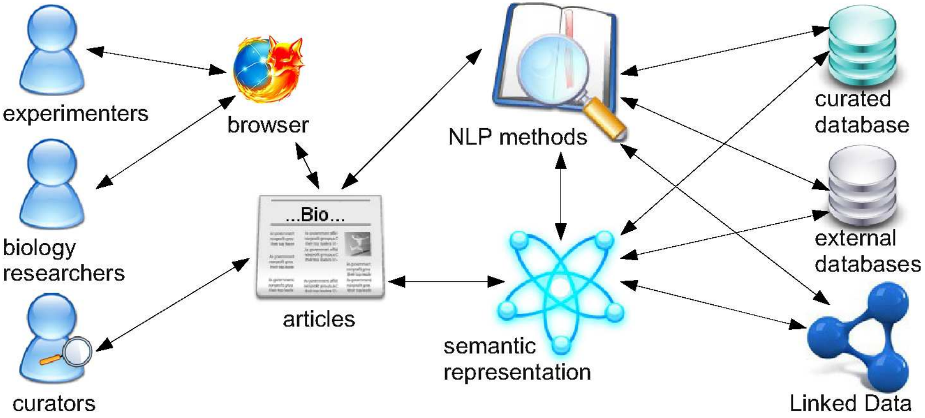
Contents of the secondary structure (a
respectively.

Enzyme

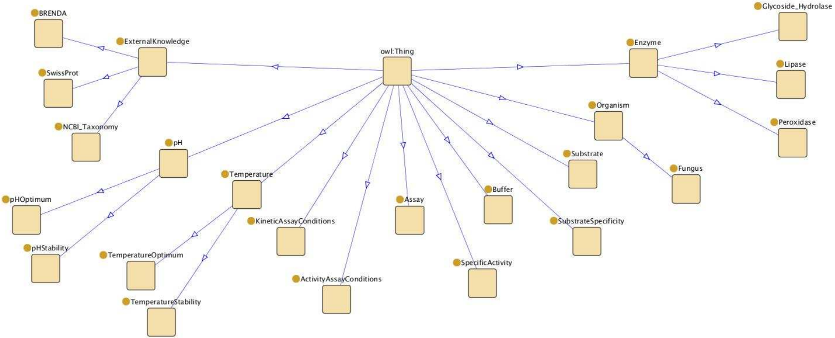
BRENDA's page0	http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.4.23.18
BRENDA_ECNumber	3.4.23.18
BRENDA_RecName	Aspergillopepsin I
SwissProtID	P41748 Q12567
alias	aspergillopepsin

→ facilitates semantic connections through Linked Data techniques

System Architecture

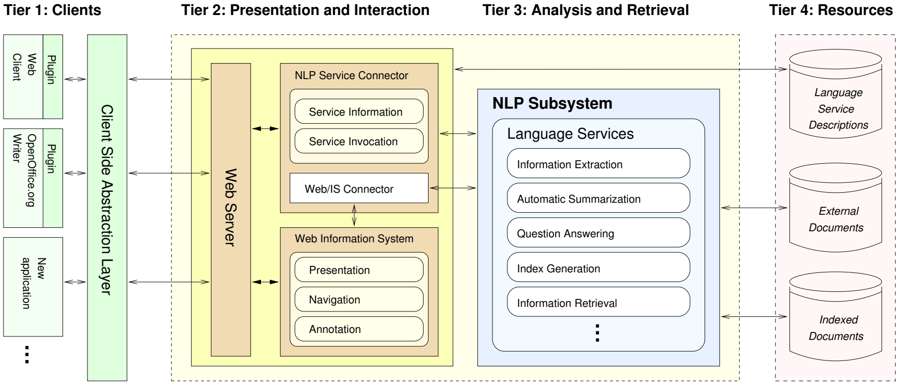


mycoMINE Ontology



Semantic Assistants

NLP services are provided by the Semantic Assistants architecture.



Introduction

Project Context and System Architecture

Text Mining Pipelines

- Text Mining Pipelines based on GATE
- Examples
- Pipeline in GATE Developer
- System Output

Evaluation

Conclusion

Text Mining Pipelines based on GATE

The *General Architecture for Text Engineering* (GATE):

- mature framework, more than 10 years of development
- development team at University of Sheffield, UK (`gate.ac.uk`)

Preprocessing steps:

- tokenization
- sentence splitting
- part-of-speech tagging

Custom pipelines:

- extract the semantic entities
- populate the OWL ontology using the OwlExporter component

Ex1. The GATE pipeline for Organism Recognition

Modules for organism entity detection:

- based on pattern matching to the NCBI reference taxonomy
- provide scientific names and NCBI Taxonomy Identifiers

Modules for extraction of strain mentions:

- specific text tokenization + machine learning (svm) based approach

Resources

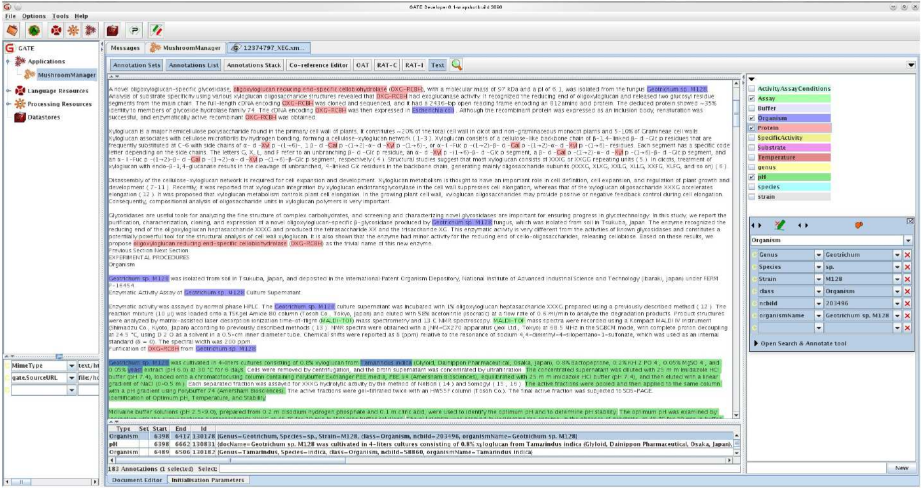
- external resources automatically translated for reuse in the system

→ ability to update the installation when the Taxonomy DB changes

Ex2. The GATE pipeline for Enzyme Recognition

1. Tokens with the *-ase* enzyme suffix:
   - enzyme-specific text tokenization
   - grammar rules [JAPE language]
2. Modules for enzyme entity detection:
   - rely on automatically extracted knowledge from BRENDA
   - based on pattern matching
   - provide enzyme *EC number*, *Recommended Name*, *Systematic Name* and *URL* on the BRENDA website.

Pipeline in GATE Developer



MimeType	text/ht
gate.SourceURL	file:/hc

A novel oligoxyloglucan-specific glycosidase, oligoxyloglucan reducing end-specific cellobiohydrolase (OXG-RCBH), with a molecular mass of 97 kDa and a pI of 6.1, was isolated from the fungus Geotrichum sp. M128. Analysis of substrate specificity using various xyloglucan oligosaccharide structures revealed that OXG-RCBH had exoglucanase activity. It recognized the reducing end of oligoxyloglucan and released two glucosyl residue segments from the main chain. The full-length cDNA encoding OXG-RCBH was cloned and sequenced, and it had a 2436-bp open reading frame encoding an 812amino acid protein. The deduced protein showed ~35% identity to members of glycoside hydrolase family 74. The cDNA encoding OXG-RCBH was then expressed in Escherichia coli . Although the recombinant protein was expressed as an inclusion body, renaturation was successful, and enzymatically active recombinant OXG-RCBH was obtained.

Xyloglucan is a major hemicellulose polysaccharide found in the primary cell wall of plants. It constitutes ~20% of the total cell wall in dicot and non-graminaceous monocot plants and 5-10% of Gramineae cell walls. Xyloglucan associates with cellulose microfibrils by hydrogen bonding, forming a cellulose-xyloglucan network (1-3). Xyloglucan consists of a cellulose-like backbone chain of β-1,4-linked β- d -Glc p residues that are frequently substituted at C-6 with side chains of α- d -Xyl p -(1→6)-, 1 β- d -Gal p -(1→2)-α- d -Xyl p -(1→6)-, or α- l -Fuc p -(1→2)-β- d -Gal p -(1→2)-α- d -Xyl p -(1→6)- residues. Each segment has a specific code letter depending on the side chains. The letters G, X, L, and F refer to an unbranching β- d -Glc p residue, an α- d -Xyl p -(1→6)-β- d -Glc p segment, a β- d -Gal p -(1→2)-α- d -Xyl p -(1→6)-β- d -Glc p segment, and an α- l -Fuc p -(1→2)-β- d -Gal p -(1→2)-α- d -Xyl p -(1→6)-β-Glc p segment, respectively (4). Structural studies suggest that most xyloglucan consists of XXXG or XXGG repeating units (5). In dicots, treatment of xyloglucan with endo-β-1,4-glucanase results in the cleavage of unbranched, 4-linked Glc residues in the backbone chain, generating mainly oligosaccharide subunits (XXXG, XLXG, XXLG, XLLG, XXFG, XLFG, and so on) (6).

Disassembly of the cellulose-xyloglucan network is required for cell expansion and development. Xyloglucan metabolism is thought to have an important role in cell definition, cell expansion, and regulation of plant growth and development (7-11). Recently, it was reported that xyloglucan integration by xyloglucan endotransglycosylase in the cell wall suppresses cell elongation, whereas that of the xyloglucan oligosaccharide XXXG accelerates elongation (12). It was proposed that xyloglucan metabolism controls plant cell elongation. In the growing plant cell wall, xyloglucan oligosaccharides may provide positive or negative feedback control during cell elongation. Consequently, compositional analysis of oligosaccharide units in xyloglucan polymers is very important.

Glycosidases are useful tools for analyzing the fine structure of complex carbohydrates, and screening and characterizing novel glycosidases are important for ensuring progress in glycotechnology. In this study, we report the purification, characterization, cloning, and expression of a novel oligoxyloglucan-specific β-glycosidase produced by Geotrichum sp. M128 fungus, which was isolated from soil in Tsukuba, Japan. The enzyme recognized the reducing end of the oligoxyloglucan heptasaccharide XXXG and produced the tetrasaccharide XX and the trisaccharide XG. This enzymatic activity is very different from the activities of known glycosidases and constitutes a potentially powerful tool for the structural analysis of cell wall xyloglucan. It is also shown that the enzyme had minor activity for the reducing end of cello-oligosaccharides, releasing cellobiose. Based on these results, we propose oligoxyloglucan reducing end-specific cellobiohydrolase (OXG-RCBH) as the trivial name of this new enzyme.
Previous Section Next Section
EXPERIMENTAL PROCEDURES
Organism

Geotrichum sp. M128 was isolated from soil in Tsukuba, Japan, and deposited in the International Patent Organism Depository, National Institute of Advanced Industrial Science and Technology (Ibaraki, Japan) under FERM P-16454.
Enzymatic Activity Assay of Geotrichum sp. M128 Culture Supernatant

Enzymatic activity was assayed by normal phase HPLC. The Geotrichum sp. M128 culture supernatant was incubated with 1% oligoxyloglucan heptasaccharide XXXG prepared using a previously described method (12). The reaction mixture (10 μl) was loaded onto a TSKgel Amide 80 column (Tosoh Co., Tokyo, Japan) and eluted with 58% acetonitrile (isocratic) at a flow rate of 0.6 ml/min to analyze the degradation products. Product structures were analyzed by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry and 13 C NMR spectroscopy. MALDI-TOF mass spectra were recorded using a Kompact MALDI IV instrument (Shimadzu Co., Kyoto, Japan) according to previously described methods (13). NMR spectra were obtained with a JNM-GX270 apparatus (Jeol Ltd., Tokyo) at 68.5 MHz in the SGBCM mode, with complete proton decoupling at 24.5 °C, using D 2 O as a solvent in a 0.5-cm inner diameter tube. Chemical shifts were reported as δ (ppm) relative to the resonance of sodium 4,4-dimethyl-4-silopentano-1-sulfonate, which was used as an internal standard (δ = 0). The spectral width was 200 ppm.
Purification of OXG-RCBH from Geotrichum sp. M128

Geotrichum sp. M128 was cultivated in 4-liters cultures consisting of 0.8% xyloglucan from Tamarindus indica (Glyloid, Dainippon Pharmaceutical, Osaka, Japan), 0.8% Bactopeptone, 0.2% KH 2 PO 4 , 0.05% MgSO 4 , and 0.05% yeast extract (pH 6.0) at 30 °C for 6 days. Cells were removed by centrifugation, and the broth supernatant was concentrated by ultrafiltration. The concentrated supernatant was diluted with 25 m m imidazole HCl buffer (pH 7.4), loaded onto a chromatofocusing column containing Polybuffer Exchanger PBE media, PBE 94 (Amersham Biosciences), equilibrated with 25 m m imidazole HCl buffer (pH 7.4), and then eluted with a linear gradient of NaCl (0-0.5 m). Each separated fraction was assayed for XXXG hydrolytic activity by the method of Nelson (14) and Somogyi (15 , 16). The active fractions were pooled and then applied to the same column with a pH gradient using Polybuffer 74 (Amersham Biosciences). The active fractions were gel-filtrated twice with an HW55F column (Tosoh Co.). The final active fraction was subjected to SDS-PAGE.
Identification of Optimum pH, Temperature, and Stability

Mclivaine buffer solutions (pH 2.5-9.0), prepared from 0.2 m disodium hydrogen phosphate and 0.1 m citric acid, were used to identify the optimum pH and to determine pH stability. The optimum pH was examined by

☐ ActivityAssayConditions
☑ Assay
☐ Buffer
☑ Organism
☑ Protein
☐ SpecificActivity
☐ Substrate
☐ Temperature
☐ genus
☑ pH
☐ species
☐ strain

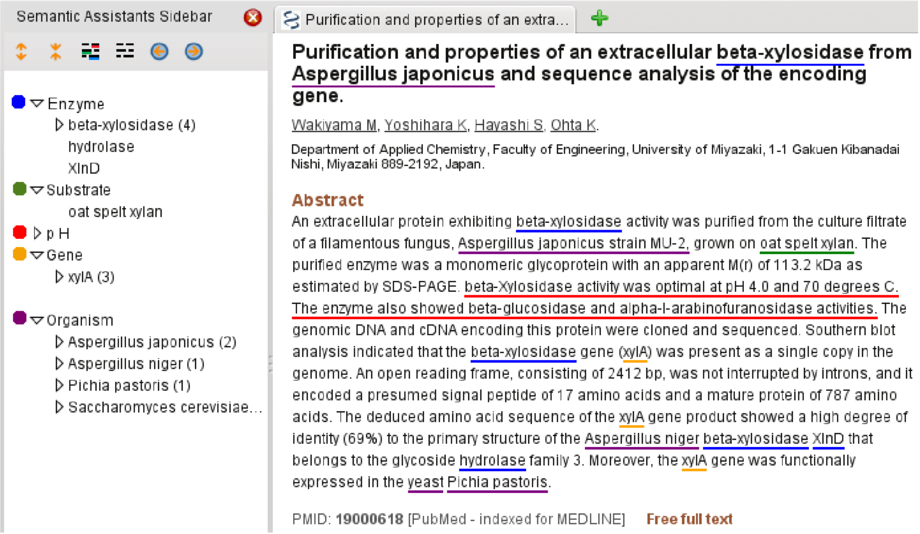
Organism

Genus	Geotrichum
Species	sp.
Strain	M128
class	Organism
ncbiId	203496
organismName	Geotrichum sp. M128

Type	Set	Start	End	Id	
Organism		6398	6417	130178	{Genus=Geotrichum, Species=sp., Strain=M128, class=Organism, ncbiId=203496, organismName=Geotrichum sp. M128}
pH		6398	6662	130831	{docName=Geotrichum sp. M128 was cultivated in 4-liters cultures consisting of 0.8% xyloglucan from Tamarindus indica (Glyloid, Dainippon Pharmaceutical, Osaka, Japan),
Organism		6489	6506	130182	{Genus=Tamarindus, Species=indica, class=Organism, ncbiId=58860, organismName=Tamarindus indica}

183 Annotations (1 selected) Select:

System output — Literature Curation: results displayed in Firefox



Semantic Assistants Sidebar
Purification and properties of an extra...

- ● Enzyme
  - beta-xylosidase (4)
  - hydrolase
  - XlnD
- ● Substrate
  - oat spelt xylan
- ● p H
- ● Gene
  - xylA (3)
- ● Organism
  - Aspergillus japonicus (2)
  - Aspergillus niger (1)
  - Pichia pastoris (1)
  - Saccharomyces cerevisiae...

Purification and properties of an extracellular beta-xylosidase from Aspergillus japonicus and sequence analysis of the encoding gene.

Wakiyama M, Yoshihara K, Hayashi S, Ohta K.

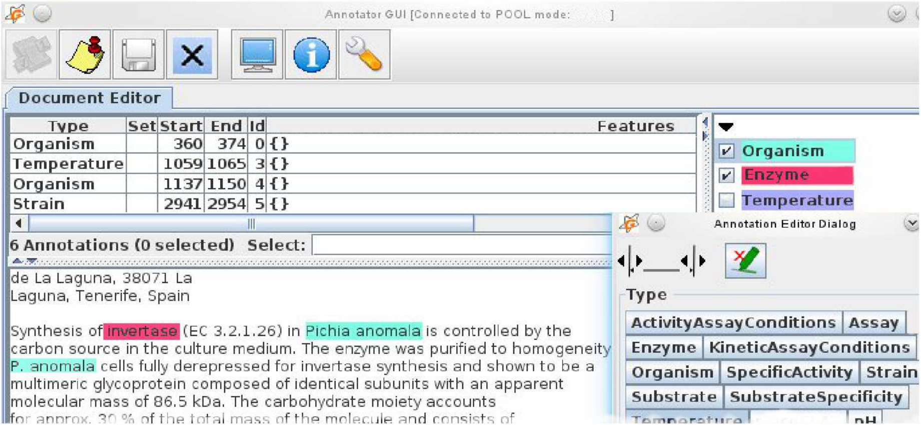
Department of Applied Chemistry, Faculty of Engineering, University of Miyazaki, 1-1 Gakuen Kibanadai Nishi, Miyazaki 889-2192, Japan.

Abstract

An extracellular protein exhibiting beta-xylosidase activity was purified from the culture filtrate of a filamentous fungus, Aspergillus japonicus strain MU-2, grown on oat spelt xylan. The purified enzyme was a monomeric glycoprotein with an apparent M(r) of 113.2 kDa as estimated by SDS-PAGE. beta-Xylosidase activity was optimal at pH 4.0 and 70 degrees C. The enzyme also showed beta-glucosidase and alpha-l-arabinofuranosidase activities. The genomic DNA and cDNA encoding this protein were cloned and sequenced. Southern blot analysis indicated that the beta-xylosidase gene (xylA) was present as a single copy in the genome. An open reading frame, consisting of 2412 bp, was not interrupted by introns, and it encoded a presumed signal peptide of 17 amino acids and a mature protein of 787 amino acids. The deduced amino acid sequence of the xylA gene product showed a high degree of identity (69%) to the primary structure of the Aspergillus niger beta-xylosidase XlnD that belongs to the glycoside hydrolase family 3. Moreover, the xylA gene was functionally expressed in the yeast Pichia pastoris.

PMID: **19000618** [PubMed - indexed for MEDLINE] **Free full text**

System output — Manual Annotation: pre-annotation in Teamware



Annotator GUI [Connected to POOL mode:]

Document Editor

Type	Set	Start	End	Id	Features
Organism		360	374	0	{}
Temperature		1059	1065	3	{}
Organism		1137	1150	4	{}
Strain		2941	2954	5	{}

6 Annotations (0 selected) Select:

de La Laguna, 38071 La
Laguna, Tenerife, Spain

Synthesis of invertase (EC 3.2.1.26) in Pichia anomala is controlled by the carbon source in the culture medium. The enzyme was purified to homogeneity P. anomala cells fully derepressed for invertase synthesis and shown to be a multimeric glycoprotein composed of identical subunits with an apparent molecular mass of 86.5 kDa. The carbohydrate moiety accounts for approx. 30 % of the total mass of the molecule and consists of

☑ Organism
☑ Enzyme
□ Temperature

Annotation Editor Dialog

Introduction

Project Context and System Architecture

Text Mining Pipelines

Evaluation

- The Manual Annotation Process
- Precision, Recall and F-measure

Conclusion

The Manual Annotation Process (1)

The annotation team:

- four biology researchers
- the researcher in charge of the curation task
- inter-annotator agreement over 80%

Table: Inter-Annotator Agreement and characteristics of the annotation tasks

Task	T1	T2	T3	T4	T5	T6	T7	Corpus
#papers	1	3	4	4	4	5	5	26
#entities	321	1158	1673	2072	2248	2288	2133	11893
IAA (%)	80	81	81	82	87	84	88	85

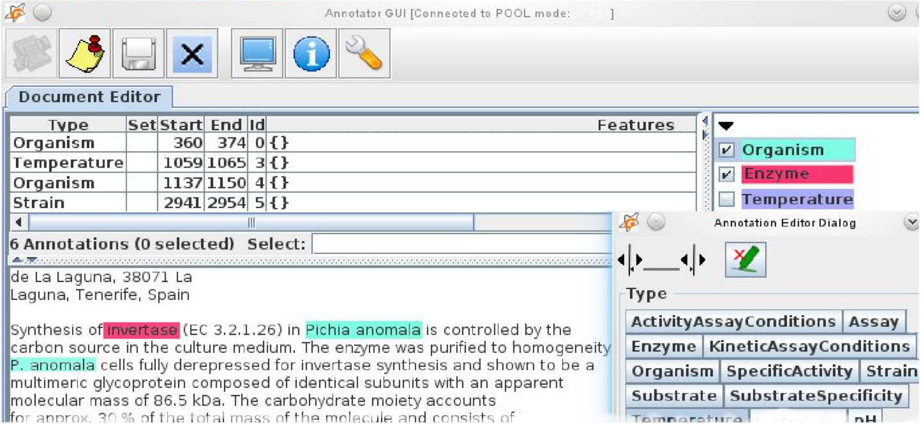
The Manual Annotation Process (2)

The gold standard corpus:

- 26 full text articles [21 freely accessible]
- manually annotated using GATE Teamware
- enzyme categories:
  - Glycoside Hydrolase = 69%
  - Lipase = 12%
  - Peroxidase = 19%
- papers published between 1996 and 2011
- 23 articles available on PubMed http://www.ncbi.nlm.nih.gov/pubmed/
- 3 articles available on ScienceDirect http://www.sciencedirect.com/
- adjudication task achieved on 11 papers

The Manual Annotation Process (3)

The GATE Teamware:



Type	Set	Start	End	Id	Features
Organism		360	374	0	{}
Temperature		1059	1065	3	{}
Organism		1137	1150	4	{}
Strain		2941	2954	5	{}

6 Annotations (0 selected) Select:

☑ Organism
☑ Enzyme
☐ Temperature

de La Laguna, 38071 La
Laguna, Tenerife, Spain

Synthesis of invertase (EC 3.2.1.26) in Pichia anomala is controlled by the carbon source in the culture medium. The enzyme was purified to homogeneity P. anomala cells fully derepressed for invertase synthesis and shown to be a multimeric glycoprotein composed of identical subunits with an apparent molecular mass of 86.5 kDa. The carbohydrate moiety accounts for approx. 30 % of the total mass of the molecule and consists of

Evaluation

- Correctness evaluated in terms of precision, recall and F-measure
- Reference = 11 adjudicated papers

Results on the four most common entities:

	Strict (overlaps discarded)			Lenient (overlaps included)		
	Recall	**Precision**	**F-m**	**Recall**	**Precision**	**F-m**
Enzyme	0.79	0.64	0.71	0.91	0.75	0.82
Organism	0.87	0.86	0.87	0.91	0.91	0.91
pH	0.79	0.81	0.80	0.96	0.99	0.98
Temperature	0.70	0.66	0.68	0.93	0.88	0.91

Conclusions

Contributions:

- mycoMINE
- text mining pipelines combined with ontological resources

Results:

- state-of-the-art results
- available gold standard corpus and system

→ future work:

- user-system interaction for data validation
- quality assessment of the curated data
- impact on the Genozymes research workflow