FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

ProFAL: Protein Functional Annotation through Literature

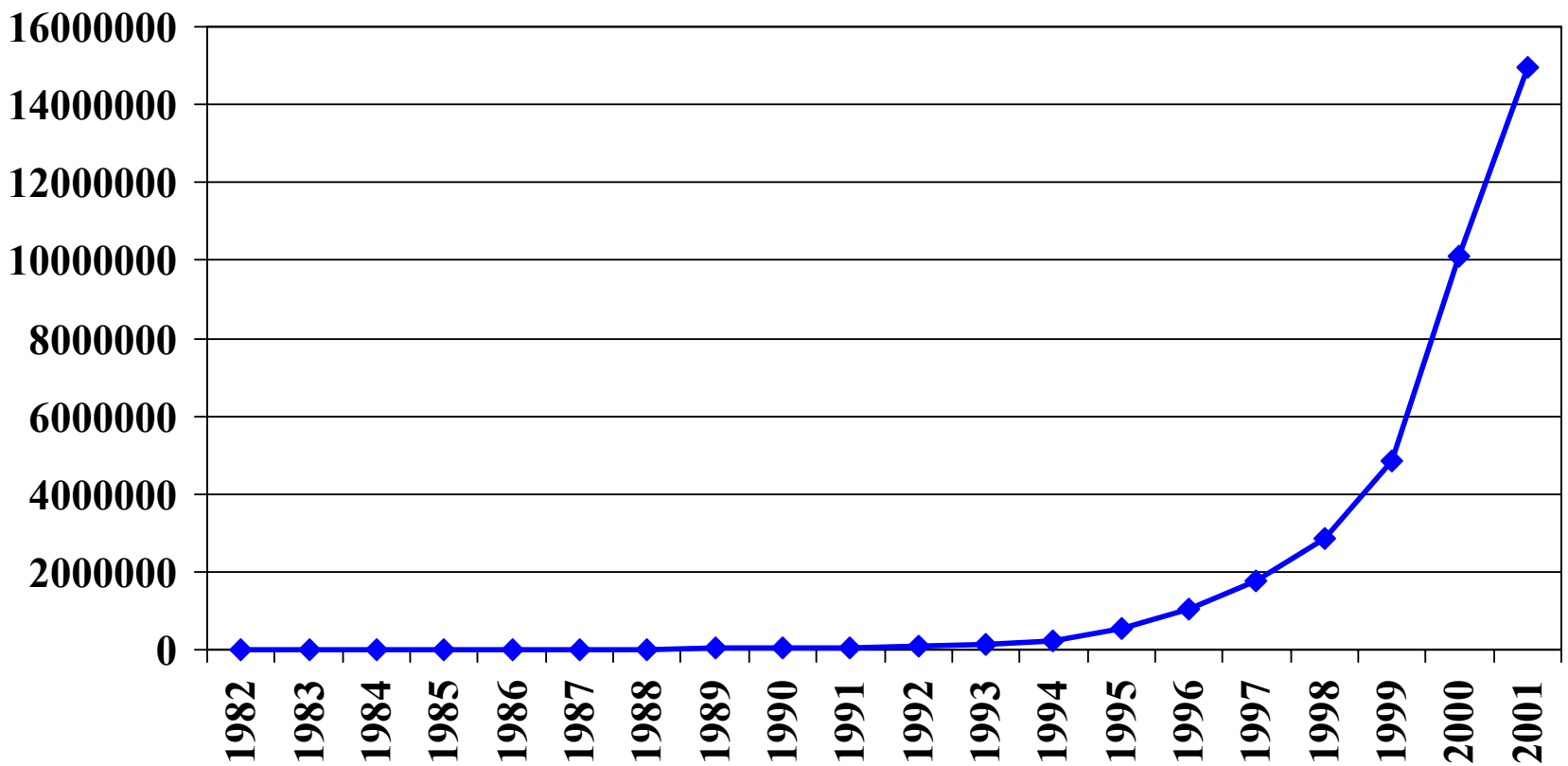
Francisco M. Couto, Mário J. Silva
{fjmc,mjs}@di.fc.ul.pt
Universidade de Lisboa

Pedro.Coutinho@afmb.cnrs-mrs.fr
(Université de Provence/CNRS, France)

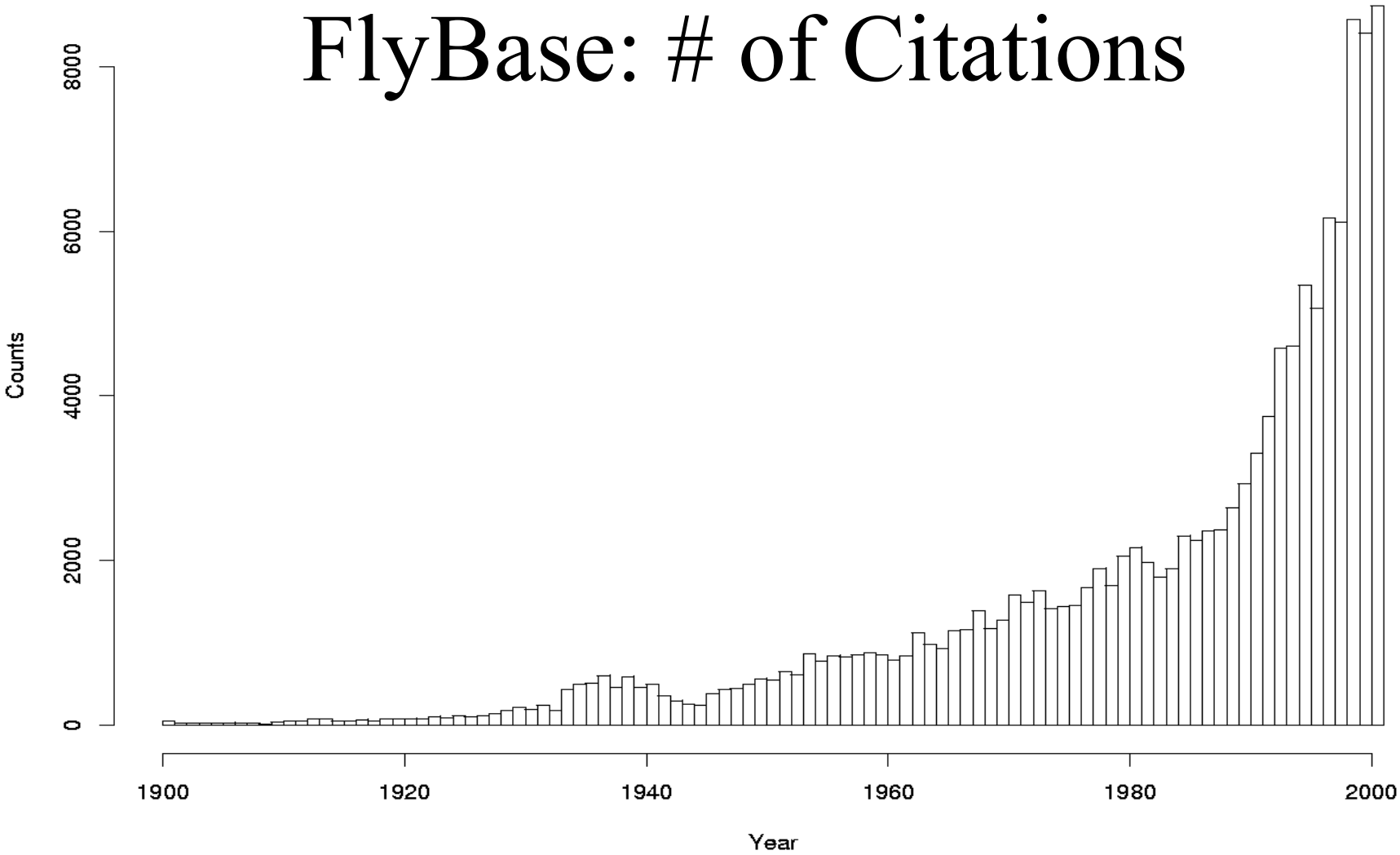
Outline

- Motivation
- ProFAL
- Case-Study
- Results
- Conclusions

Genbank


# of sequences
16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001

FlyBase: # of Citations



Biologic Literature

- MEDLINE
  - "Beginning in 2002 over **2,000** completed references are added daily each Tuesday through Saturday, January through October"
  - "over **460,000** added last year"
  - Reading 10 articles per day, takes **112** years to read those articles

Motivation

Common Knowledge

Common Knowledge

Two heads (or more) are better than one, except when they don't share information. That's where knowledge management comes in.

BY SALVATORE SALAMONE

What's more difficult than finding a needle in a haystack? Defining knowledge management.

What's in a Name?

A somewhat different approach to finding answers in numerous databases of unstructured data is to use more intelligent search tools. More intelligent tools are needed because common text search tools that find things so well on the Web do not work as well with life science databases, largely because there is no universal, consistent way to denote genes, disease names, or other entities in the scientific literature.

Motivation

nature genetics
volume 28 no. 1 may 2001

Community watch

Microarray analysis yields a wealth of information that presents the geneticist with a vexatious challenge: where does one go with a cluster of genes defined by similar or identical expression levels in different tissues, or during adaptation to a new environment? How does one single out genes that warrant further analysis? Three

PubGene, the GO initiative and other efforts to establish controlled vocabularies are testimony to the power of *in silico* analysis and cast a new light on the commodity that is published text. The extent to which researchers are able to use published text—as a tool—depends on publishing strategy, and poses interesting questions for the publishers of on-line scientific literature. It is one that will become more interesting as literature-mining tools evolve.

Motivation

Economist.com **PREMIUM CONTENT**

denotes premium content | Log in | Free registration | Help

The better use of scientific literature

A genetic paperchase

May 3rd 2001
From The Economist print edition

THERE is a difference between information and knowledge. With the completion of the Human Genome Project, biologists have a lot of information about what human genes are. Knowledge of how they work—and in particular of how their products interact to form the vast network of biochemical pathways referred to as "life"—is harder to assemble. But a group of biologists led by Eivind Hovig of the Norwegian Radium Hospital in Oslo has managed to find a way to speed the process up, by linking together the vast quantities of disconnected biochemical information that have already been published.

Select an option below to continue:

Subscribe

Get unlimited site access. Sign up for:

- ○ a month, $19.95 (automatic renewal)
- ◉ a year, $69 (automatic renewal; save 70%)
- ○ a print subscription (complimentary web access)

Subscribe

Pay Per View

Buy credits to view individual articles:

- ○ 1 article-credit, $2.95
- ○ 5 article-credits, $9.95
- ◉ Use credits already on my account

Go

Current Subscribers

Log in to view the article.

E-mail address

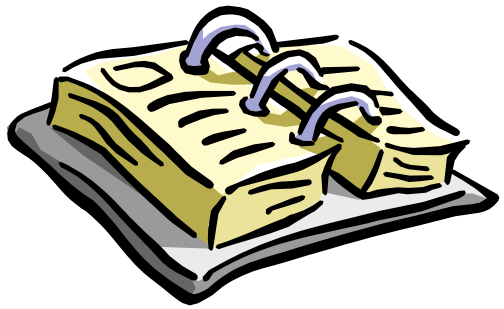
Password

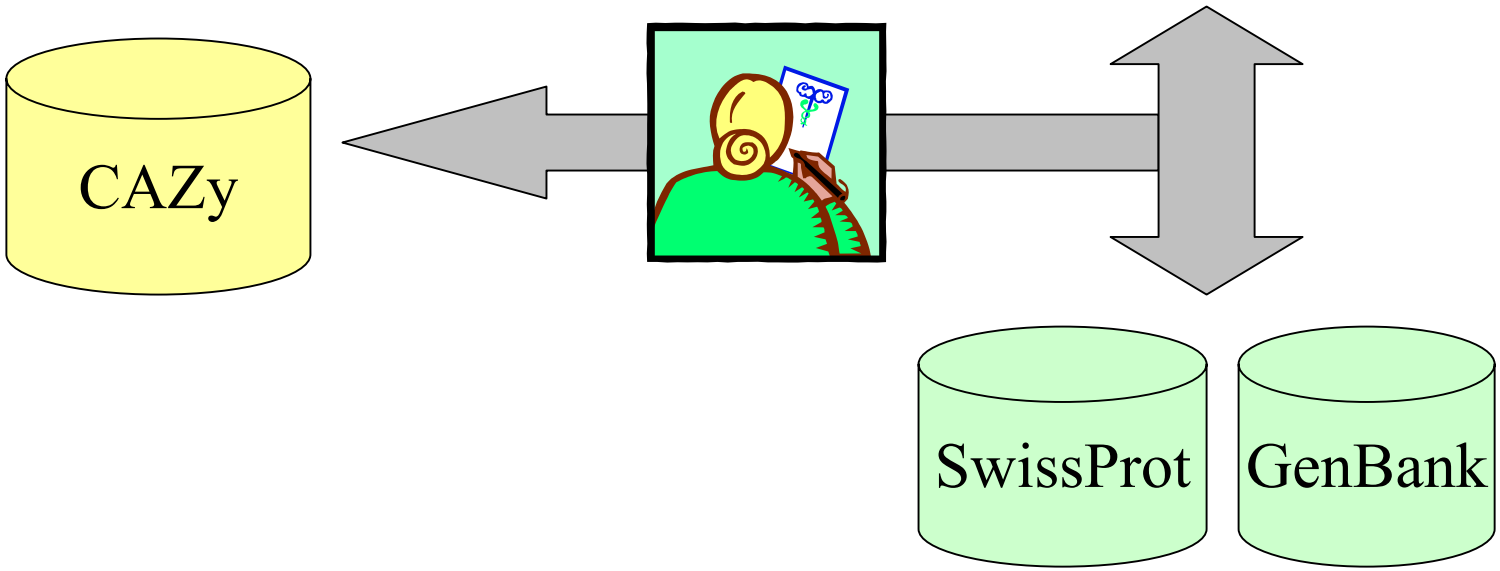
Outline

- Motivation
- ProFAL
- Case-Study
- Results
- Conclusions

Problem

- Human curators read the literature to identify annotations and update their databases

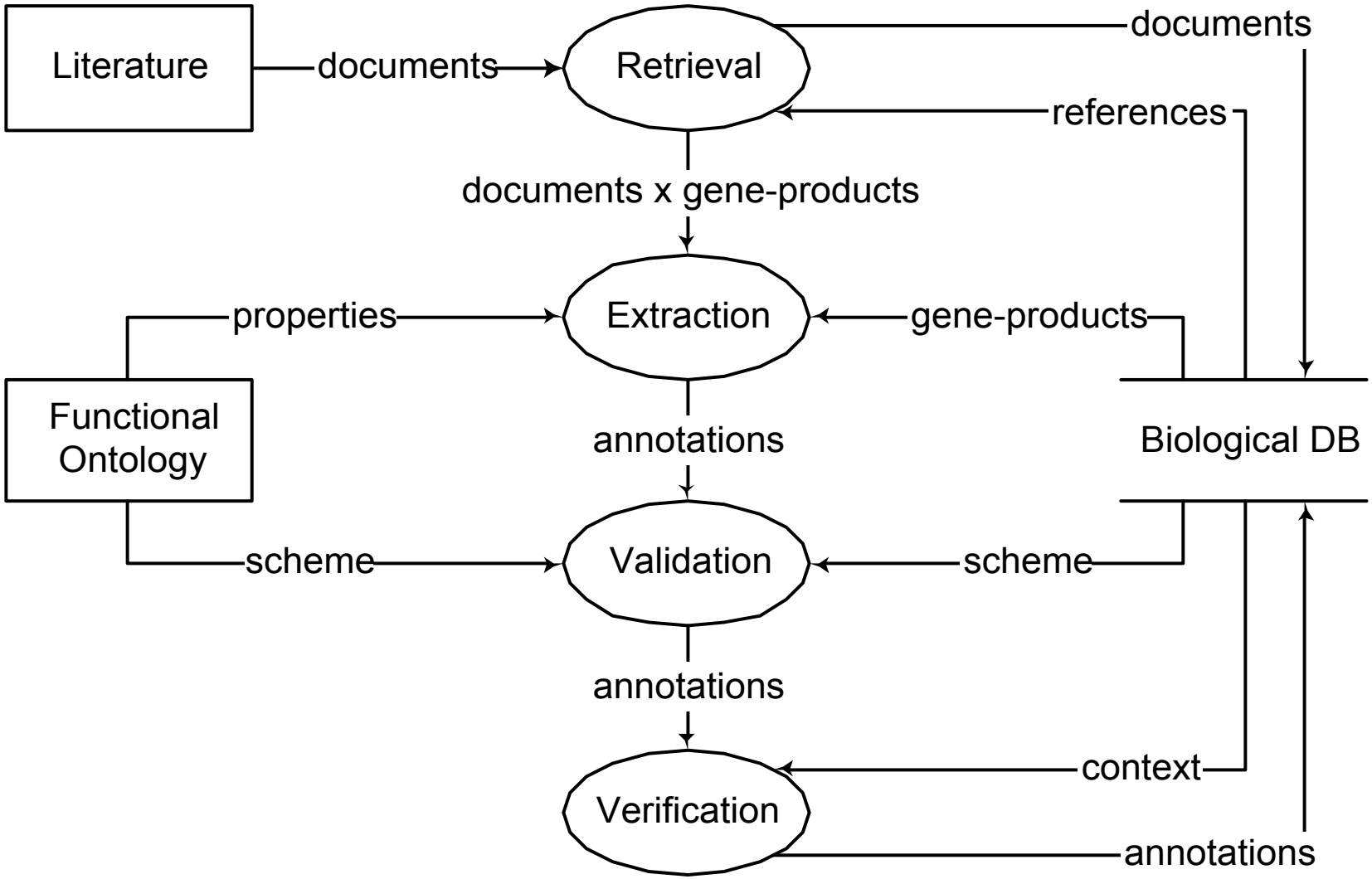




ProFAL - Protein Functional Annotation Through Literature

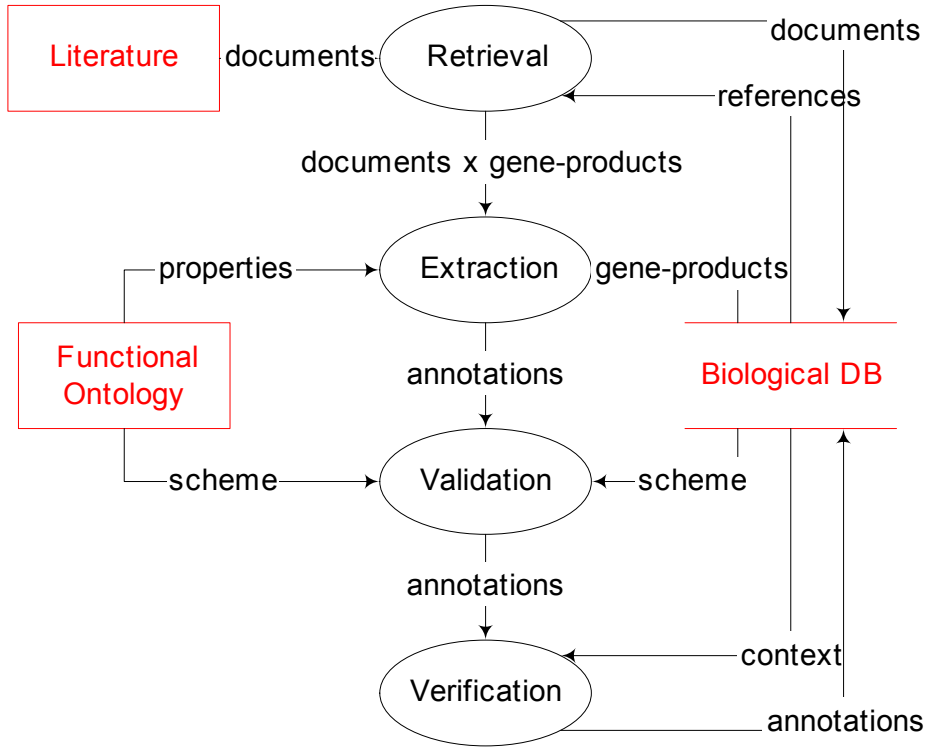
- Software tool for biologic database annotatation and verification
  - Automatic retrieval and extraction of annotations from literature
  - Annotation validation method based on the correlation between strucuture and biologic function

ProFAL

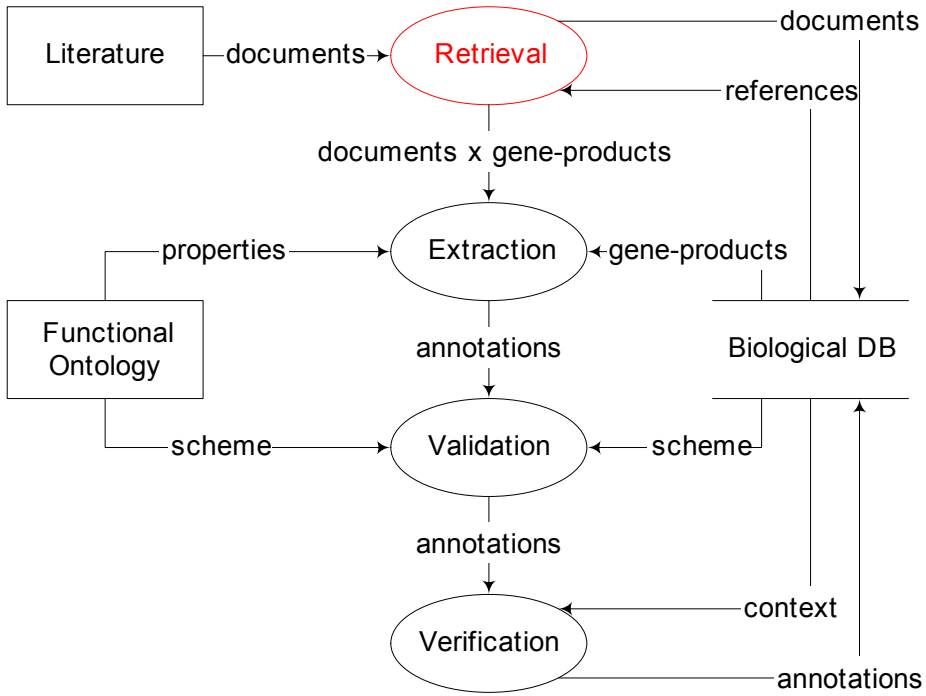

Literature
documents
Retrieval
documents
references
documents x gene-products
properties
Extraction
gene-products
Functional Ontology
annotations
Biological DB
scheme
Validation
scheme
annotations
context
Verification
annotations

Information Sources

- Biological DB
  - Stores gene-products
  - Classified in families according to their structure
- Functional Ontology
  - A set of functional terms
  - Classified in a hierarchical taxonomy
- Literature
  - A collection of documents

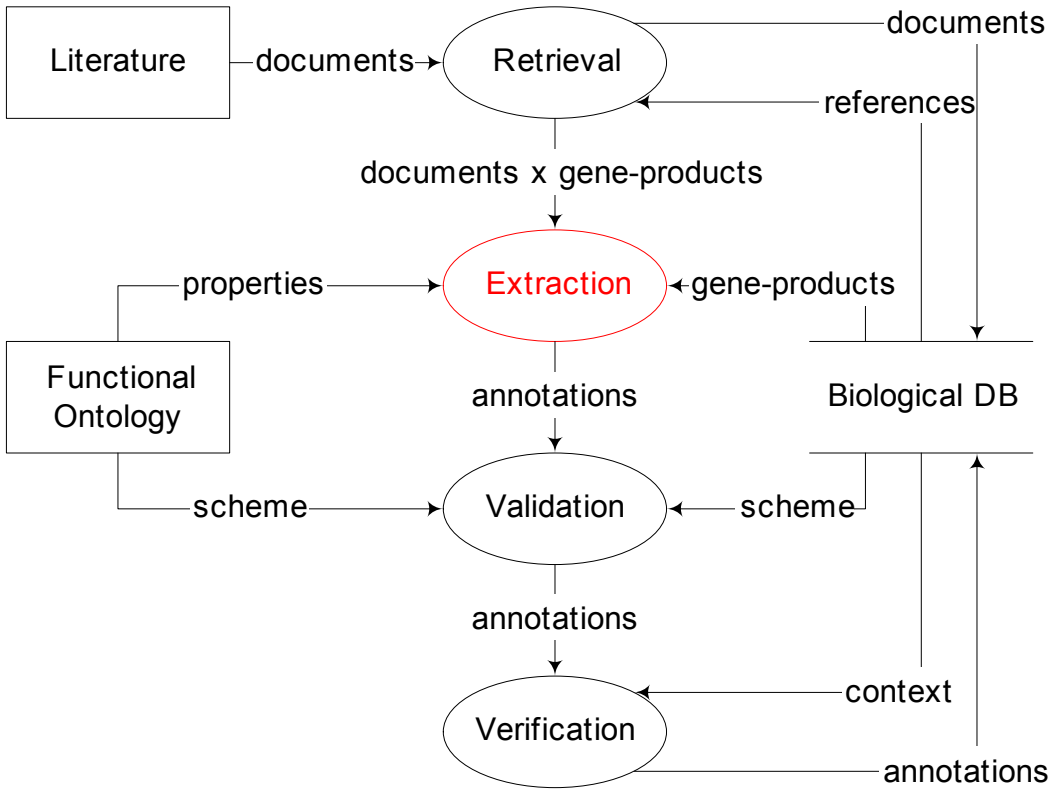


Processes: Retrieval



- Assign each gene-product to a set of documents
- Retrieve the document information

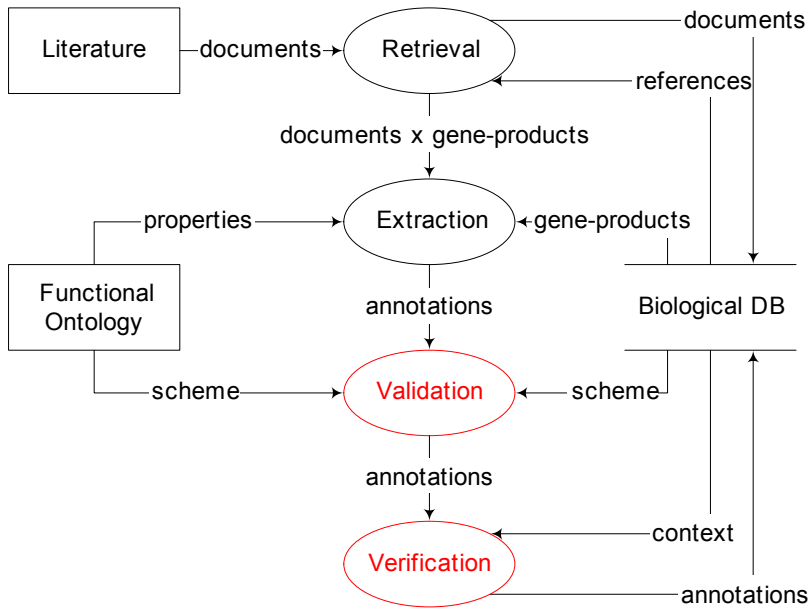
Extraction



- Annotate each gene-product with functional properties found in its assigned documents.

Validation & Verification

- Validation (automatic)
  - Use of heuristics about the domain
  - Quantitative classification of each annotation according to its realibility.
- Verification (computer assisted)
  - Analysis of annotations
  - Discarding of misannotations



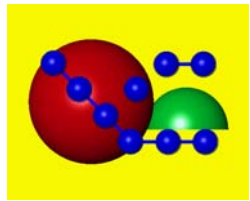
Outline

- Motivation
- ProFAL
- Case-Study
- Results
- Conclusions

Goals

- Validate the use of our annotation extraction and validation methods in a real environment
- Automate the annotation process on an existing biologic database

CAZy



- CAZy - Carbohydrate Active enZYmes
- Enzymes Classified in families
  - annotated with EC (Enzyme Comission) numbers
- Enzyme Sequences structure in modules (functional/structural segments of sequence)
  - Catalytic
  - carbohydrate Families

(http://afmb.cnrs-mrs.fr/CAZY)

Cazy Information Sources

- GenBank/GenPept,
- SwissProt,
- PDB
- Literature
- …

Case-Study

<< phpCAZy >> CAZyModO

CAZyModO - Glycoside Hydrolase Classification

Family GH48 | Modular Organization | Go

CAZy Family	Glycoside Hydrolase Family 48
Known Activities	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91).
Mechanism	Inverting
Catalytic Nucleophile/Base	Not known
Catalytic Proton Donor	Not known
3D Structure Status	Available (see PDB)
Note	formerly known as cellulase family L.
Relevant Links	InterPro; PRINTS
Statistics	CAZyModO(12)

#ac	Protein		Organism	Modular Organisation	Status
154	cellulase CelA		*Anaerocellum thermophilum*	1 GH9 443 445CBM3612 613 681 682CBM3834 835 880 881CBM31034 1035 1080 1085 GH48 1371	
1497	CelA		*Caldicellulosiruptor saccharolyticus*	1 23 24 GH9 466 467CBM3642 643 700 701CBM3857 858 903 904CBM31060 1061 1112 1113 GH48 1748	
1591	cellobiohydrolase B	*Cel48A*	*Cellulomonas fimi*	1 33 34 53 54 GH48 699 700FN3785 794FN3884 891FN3978 989CBM21090	
10828	ORF CAC0911		*Clostridium acetobutylicum ATCC 824*	40 GH48 660 661DOC1726	
1716	CelF	*Cel48A*	*Clostridium cellulolyticum*	1 29 30 GH48 660 665DOC1722	3D
1730	exoglucanase S		*Clostridium cellulovorans*	1 32 33 GH48 643 644DOC1703	
1733	CelD		*Clostridium josui*	1 29 30 GH48 650 651DOC1719	

Case-Study

>> CAZyModO

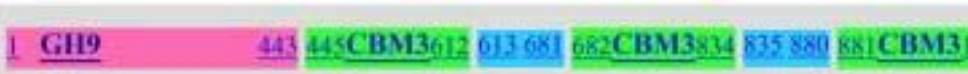
on

Family GH48 | Modular Organization | Go

CAZy Family	Glycoside Hydrolase Family 48
Known Activities	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91).
Mechanism	Inverting
Catalytic Nucleophile/Base	Not known
Catalytic Proton Donor	Not known
3D Structure Status	Available (see PDB)
Note	formerly known as cellulase family L.
Relevant Links	InterPro; PRINTS
Statistics	CAZyModO(12)

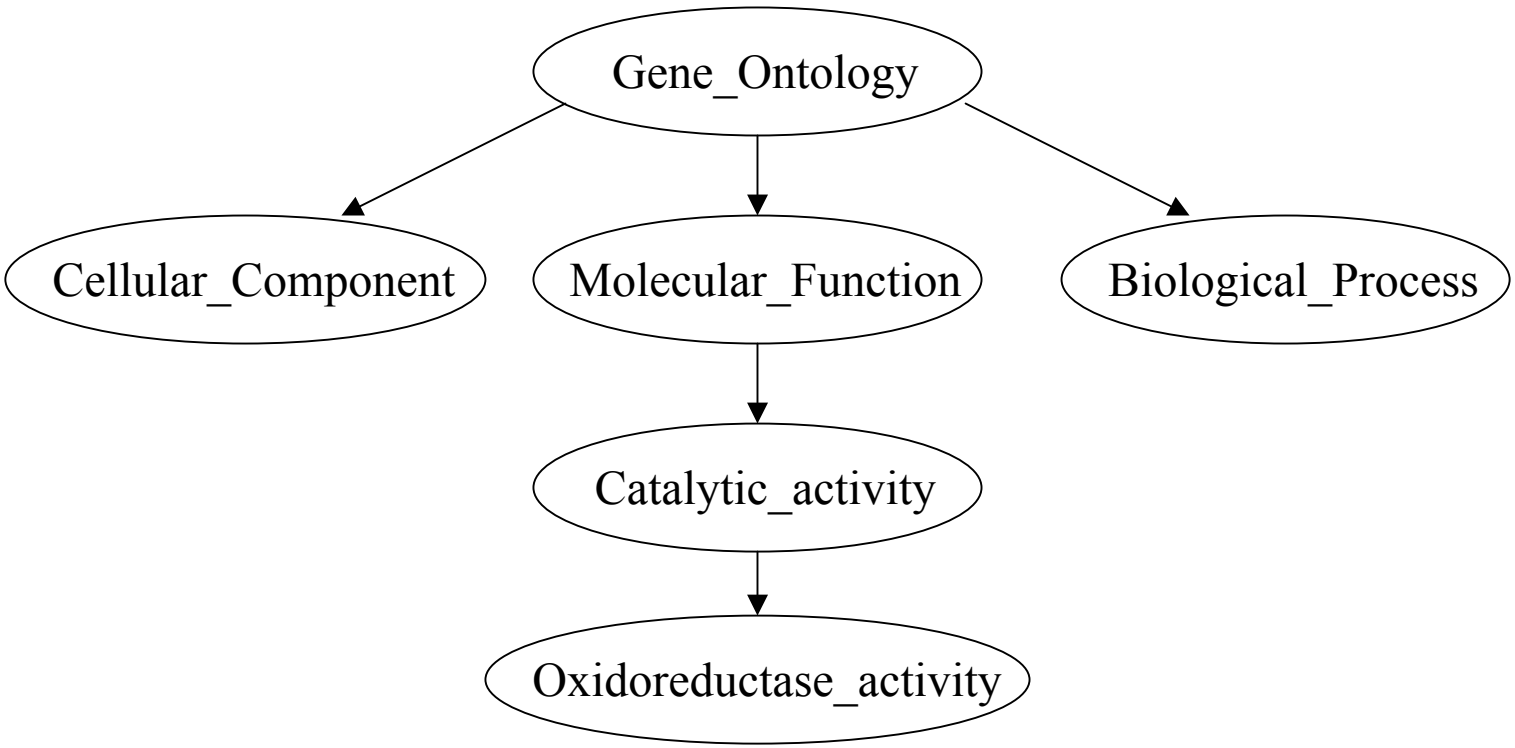
Modular Organisation **Status**

Mechanism	Inverting
Catalytic Nucleophile/Base	Not known
Catalytic Proton Donor	Not known
3D Structure Status	Available (see
Note	formerly know
Relevant Links	InterPro; PRIN
Statistics	CAZyModO(1

#ac	Protein		Organism	Modular Organisation
154	cellulase CelA		*Anaerocellum thermophilum*	
1497	CelA		*Caldicellulosiruptor saccharolyticus*	
1591	cellobiohydrolase B	*Cel48A*	*Cellulomonas fimi*	
10828	ORF CAC0911		*Clostridium acetobutylicum ATCC 824*	
1716	CelF	*Cel48A*	*Clostridium cellulolyticum*	
1730	exoglucanase S		*Clostridium cellulovorans*	
1733	CelD		*Clostridium josui*	

Gene Ontology

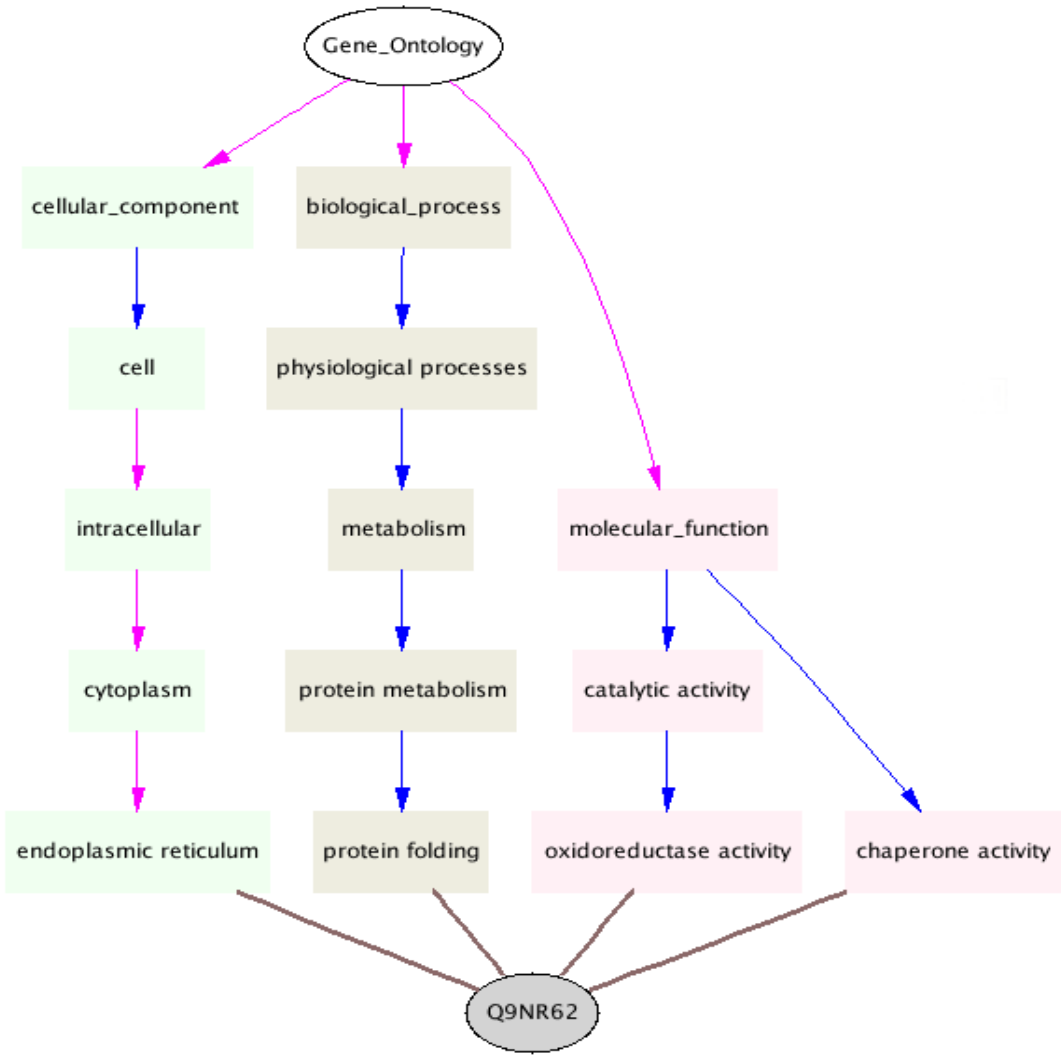
- Structure concepts and their relations:



Case-Study

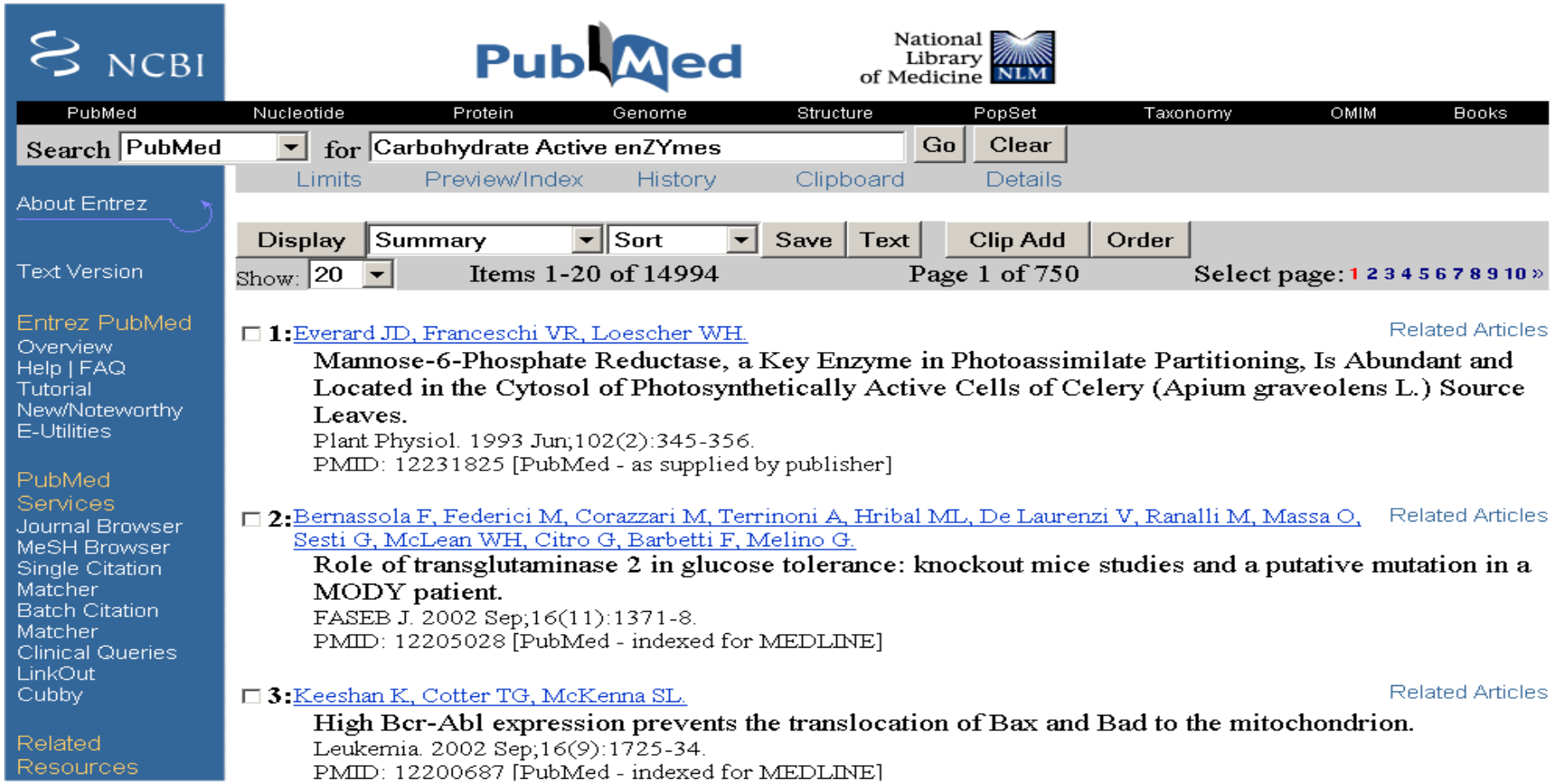
Annotation

- Annotate objects related to the concepts:
- Example:
  Protein name
  Endoplasmic reticulum oxidoreductin 1-Lbeta

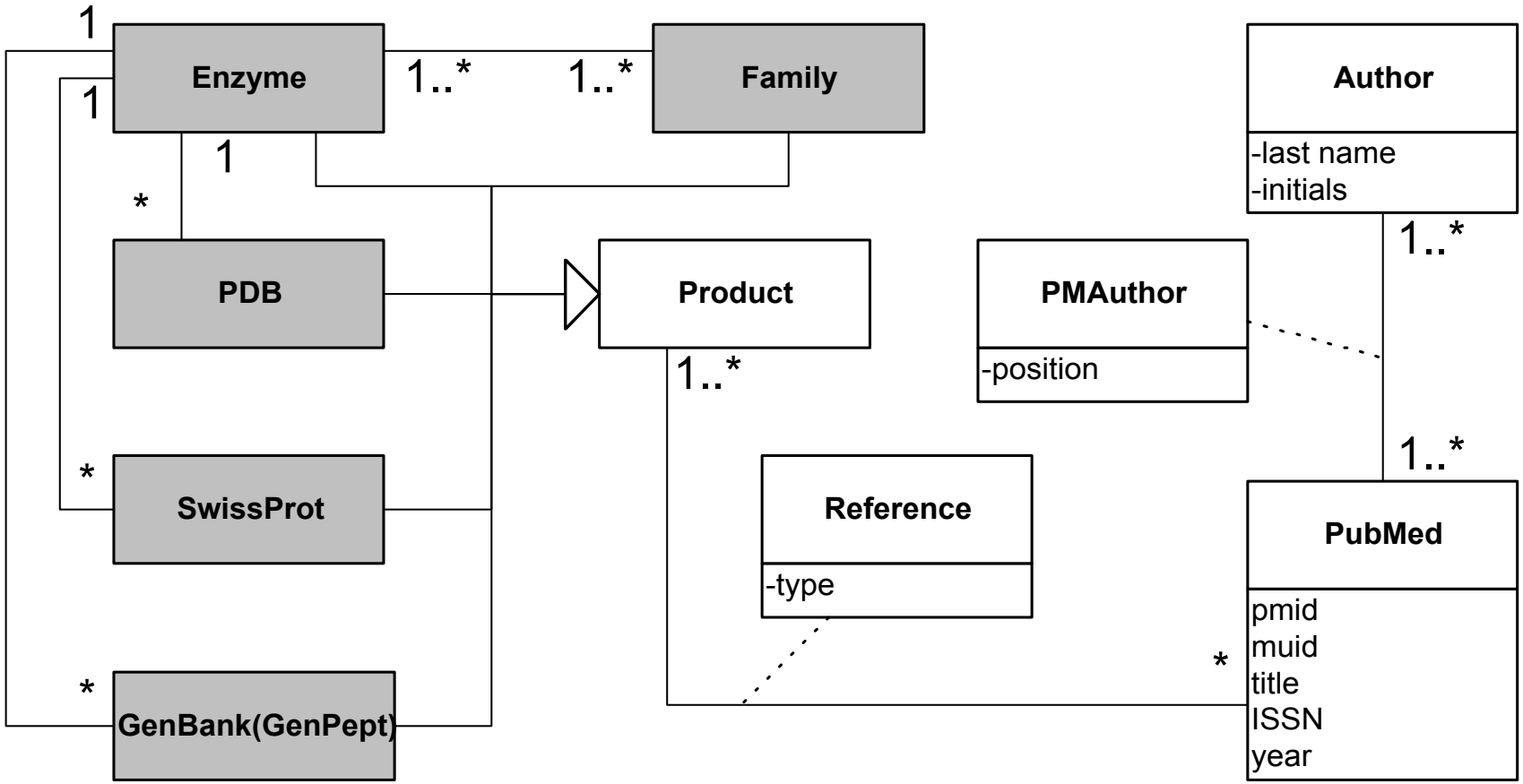


Case-Study

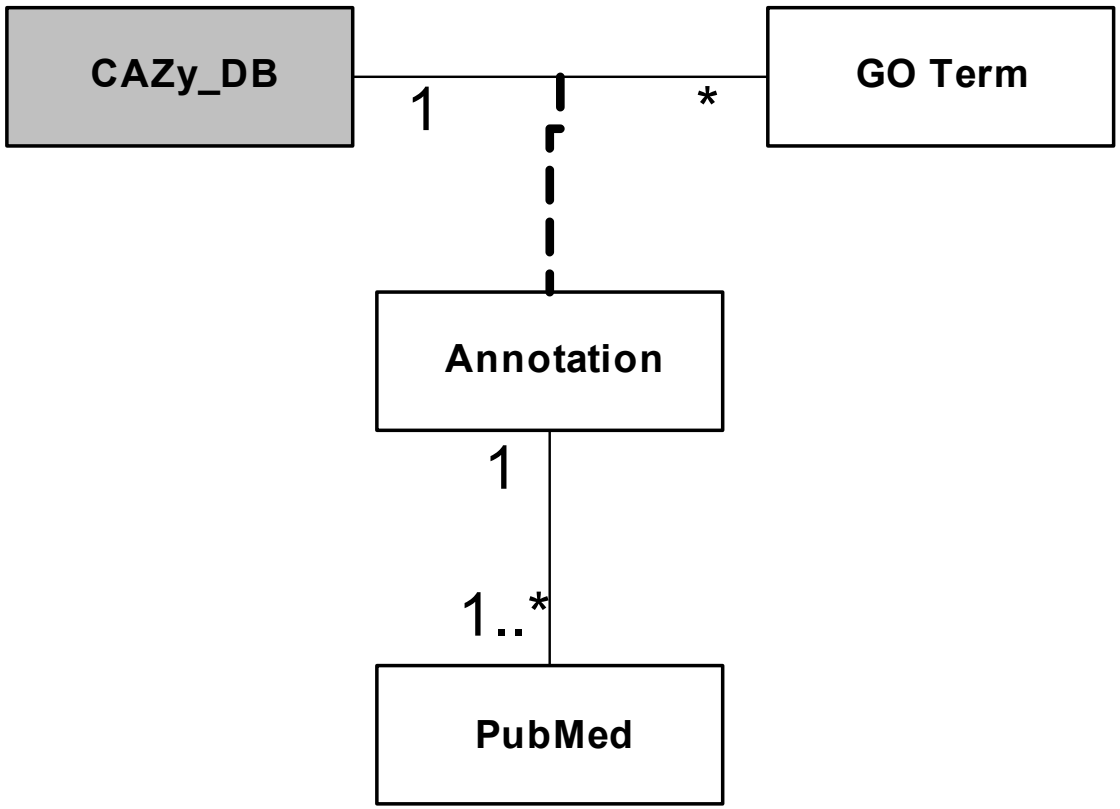
PubMed Abstracts



Retrieval

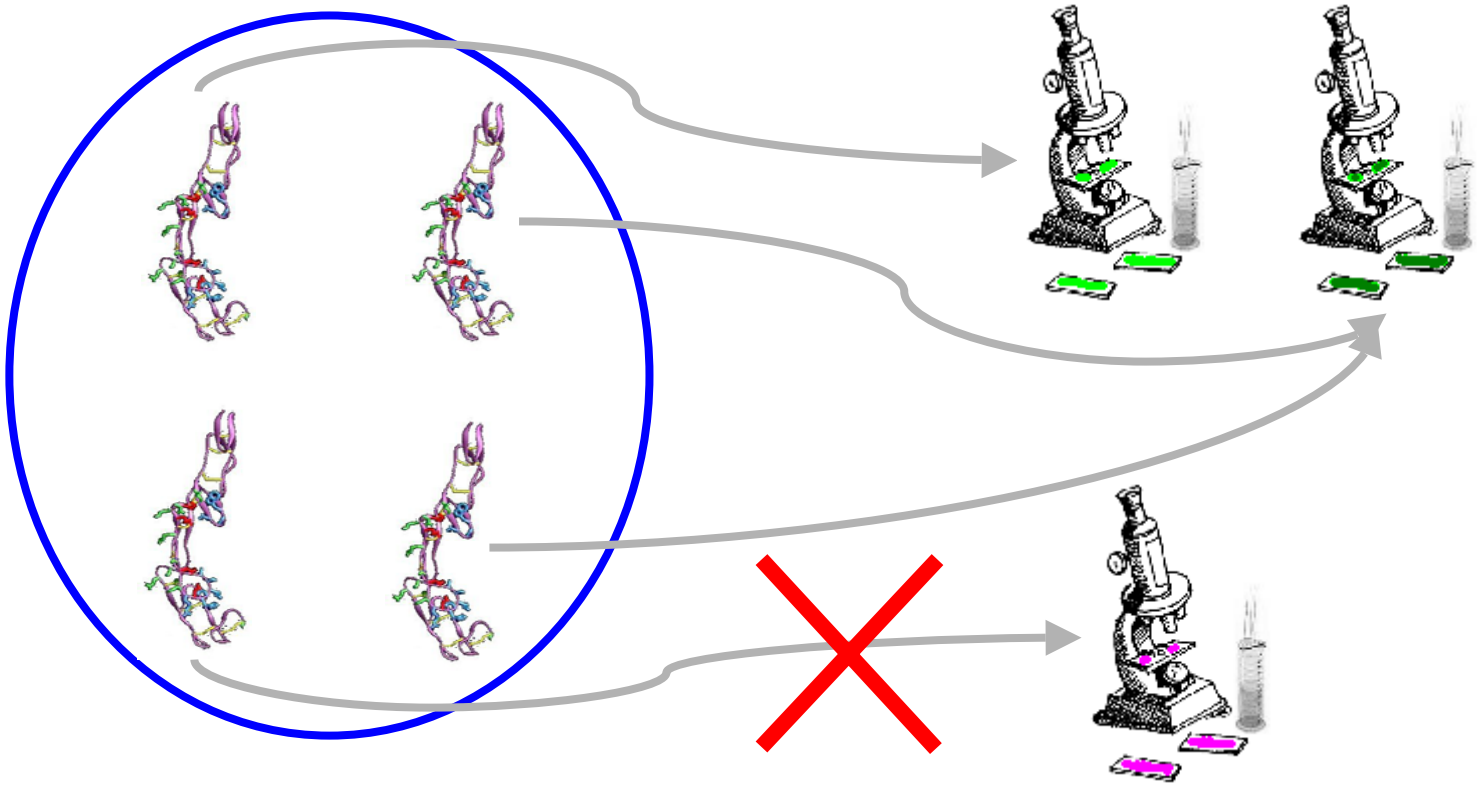

1
Enzyme
1
1..*
1..*
Family
1
*
PDB
Product
1..*
*
SwissProt
*
GenBank(GenPept)
Reference
-type
Author
-last name
-initials
1..*
PMAuthor
-position
1..*
PubMed
pmid
muid
title
ISSN
year
*

Extraction



Validation Method

- Corelation between function and structure



Case-Study

Verification

Publications

PubMedID	MedlineID	Title	ISSN	Year	Classification	Note	#Authors	DB_ac
11435116	21345419	The kappa-carrageenase of P. carrageenovora features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases.	0969-2126	2001	1	-3D-	7	
8112578	94156170	The gene encoding the kappa-carrageenase of Alteromonas carrageenovora is related to β-1,3-1,4-glucanases.	0378-1119	1994	1		3	

Options | Search All | Insert PMID | 11435116 | 9 | Alter Classification

Annotations

TermsID	TermsName	Classification	Note	PubMedIDs
GO:0016787	hydrolase	1		8112578
GO:0008810	cellulase	1		11435116

Options | Insert Term | GO:0008810 | 9 | Alter Classification

Outline

- Motivation
- ProFAL
- Case-Study
- Results
- Conclusions

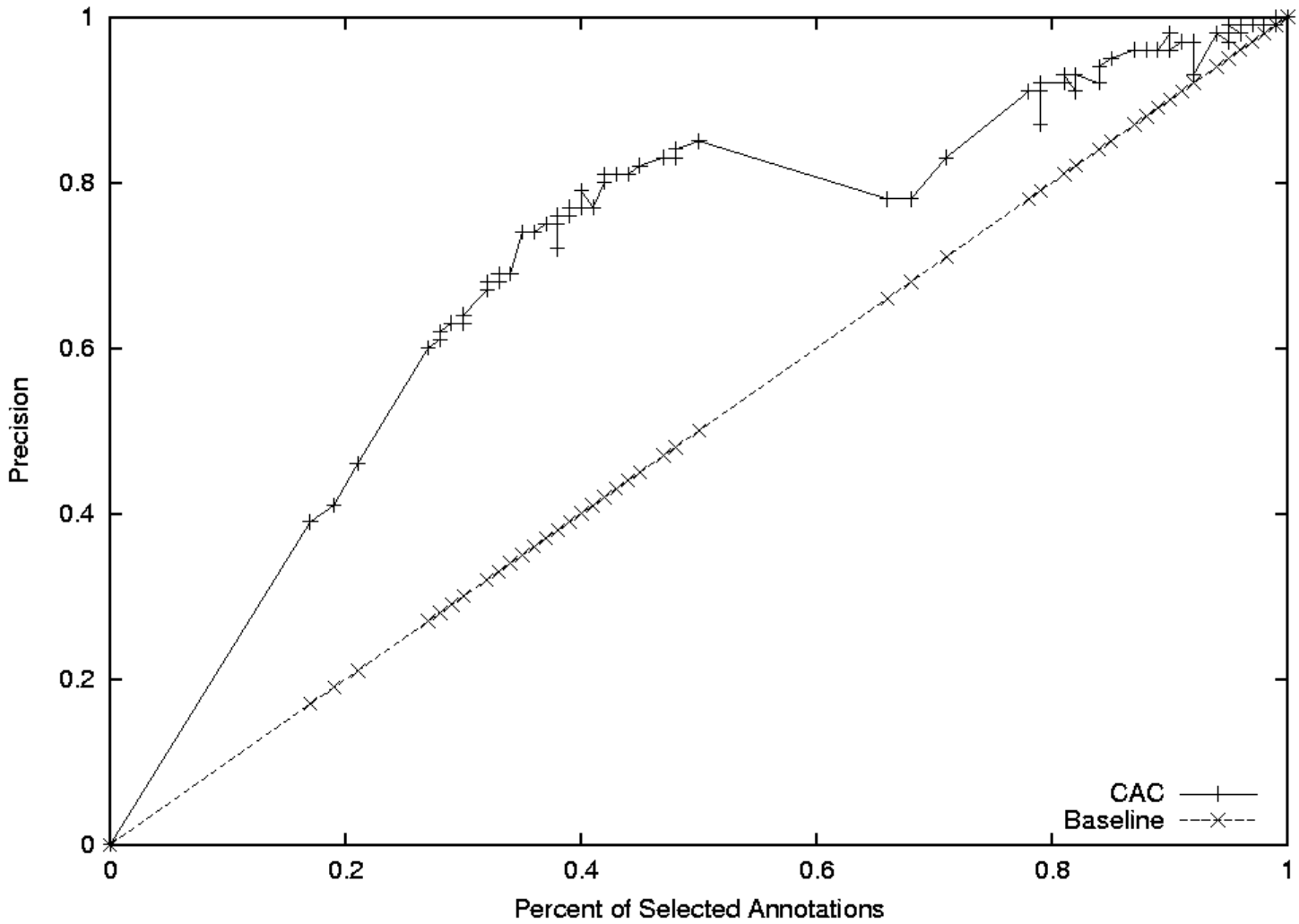
Retrieval

	References /Citations	Documents
GenBank(GenPept)	22849	4575
SwissProt	8998	4006
PDB	3561	785
Total	35408	6377

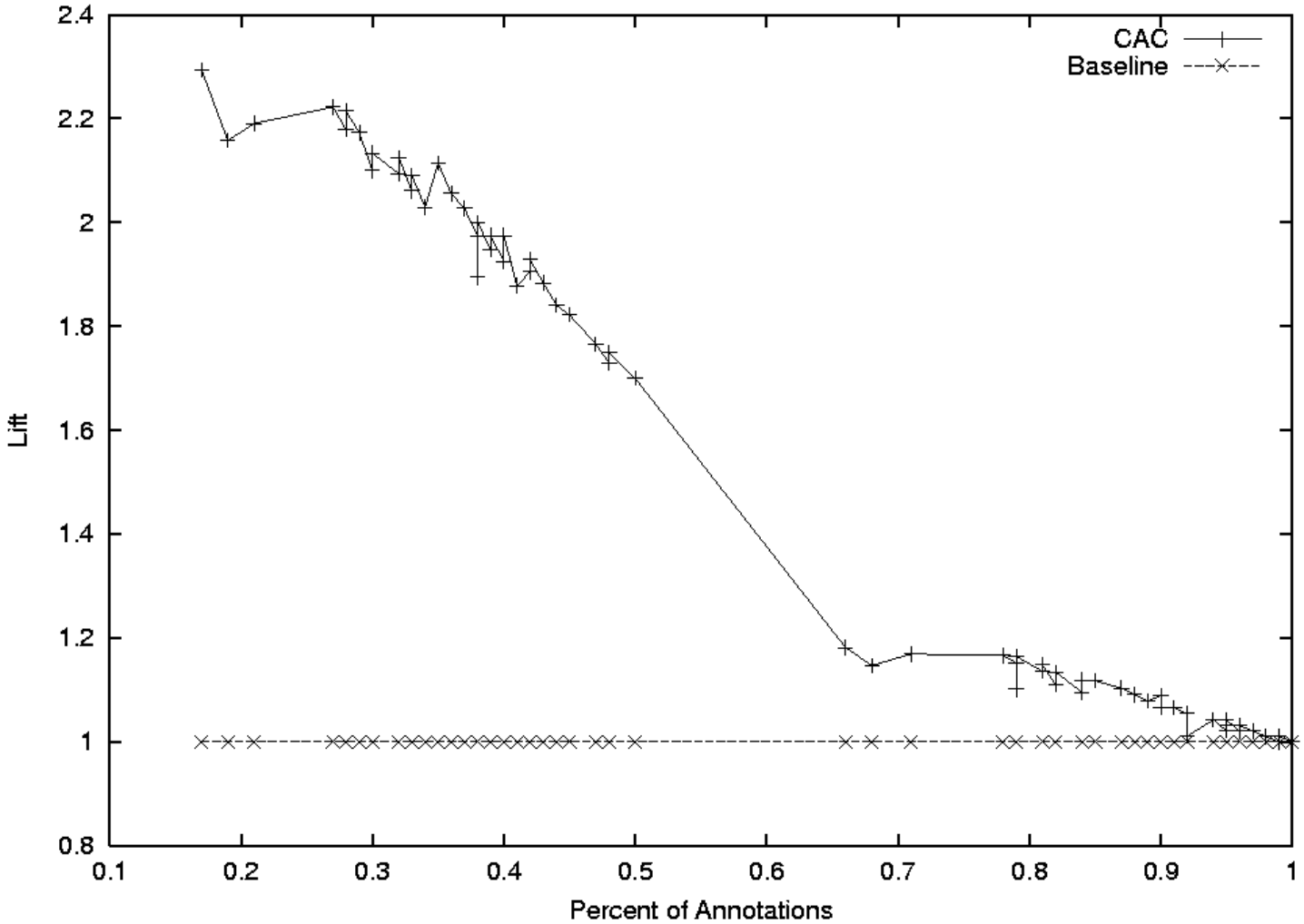
Extraction

- 13869 annotations
- 6918 enzymes
- 1342 terms
- Only 40% of enzymes were annotated
  - Lack of bibliographic references for 60% of enzymes
  - Few references in Cazy
  - Abstracts only
- Average of 2.2 annotations per document

Validation: Gain Chart



Validation: Lift Chart



Verification

- 173 annotations from 5 familes
- Human verification
  - 95 correct and 78 incorrect annotations
  - Precision of 55%
  - Task completed in one hour
- Assuming a correct annotation per document:
  - The curator had to read 95 documents to identify 95 correct annotations
  - Assuming that reading a document takes 10 minutes, this would require about 16 hours without ProFAL.

Outline

- Motivation
- ProFAL
- Case-Study
- Results
- Conclusions

Conclusions

- ProFAL
  - Retrieves related Literature
  - Extracts annotations between gene-products and functional properties
  - Validates the annotations
  - Verifies the annotations
- Applied to CAZy
  - Improves the curation process done by humans
- Application to Arabidopsis gene expression data under development

More Information:

http://xldb.di.fc.ul.pt/rebil/