Large-scale Information Extraction for Biomedical Literature

1st Swiss Text Analytics Conference (Swisstext 2016)

Fabio Rinaldi, Lenz Furrer
www.ontogene.org

June 8, 2016

the cogito foundation

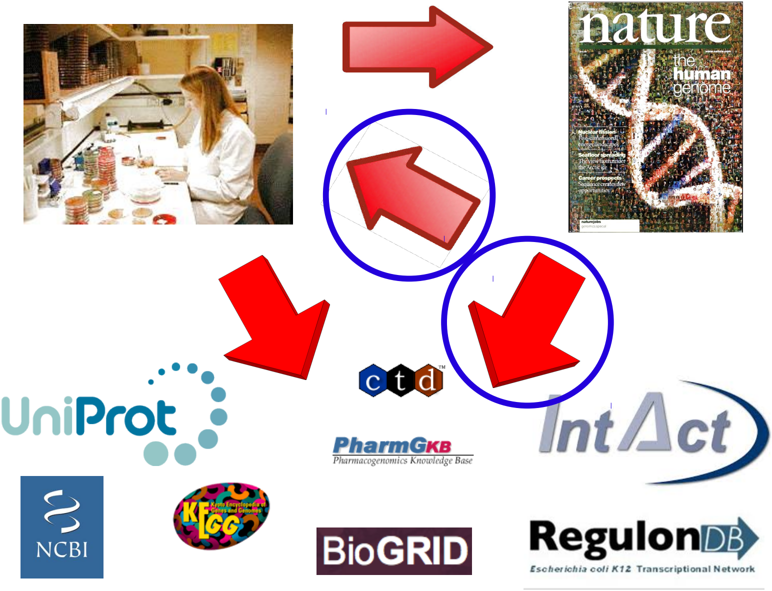
FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Roche

Outline

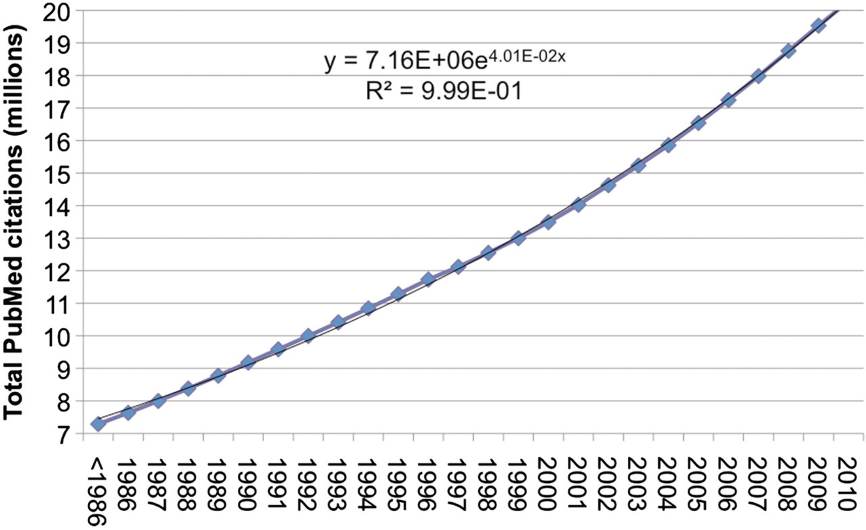
1. Motivation
2. Text mining in a curation workflow
3. Large-scale detection of protein interactions
4. Biomedical text mining: competitive evaluations
5. The OntoGene approach and highlights
6. Preview of recent work

Flow of Information – Role of Curation


nature
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
UniProt
ctd
PharmGKB
Pharmacogenomics Knowledge Base
IntAct
NCBI
KEGG
Kyoto Encyclopedia of Genes and Genomes
BioGRID
RegulonDB
Escherichia coli K12 Transcriptional Network

Growth of Information

Pubmed citations 1986–2010

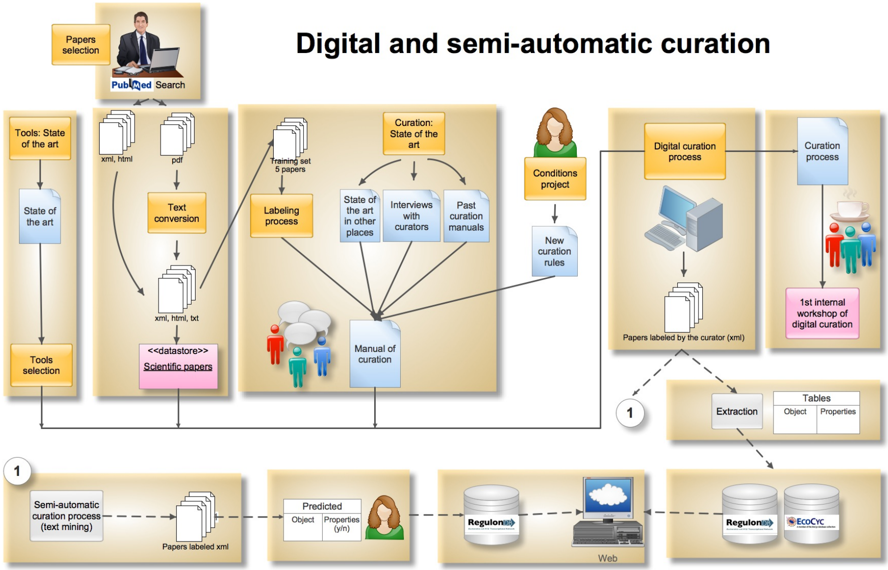


Zhiyong Lu: PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011:baq036

Outline

1. Motivation
2. Text mining in a curation workflow
3. Large-scale detection of protein interactions
4. Biomedical text mining: competitive evaluations
5. The OntoGene approach and highlights
6. Preview of recent work

Assisted Curation



Digital and semi-automatic curation

Papers selection

PubMed Search

Tools: State of the art

State of the art

Tools selection

xml, html

pdf

Text conversion

xml, html, txt

<<datastore>>
Scientific papers

Training set
5 papers

Labeling process

Curation: State of the art

State of the art in other places

Interviews with curators

Past curation manuals

Manual of curation

Conditions project

New curation rules

Digital curation process

Papers labeled by the curator (xml)

Curation process

1st internal workshop of digital curation

1

Extraction

Tables	
Object	Properties

1

Semi-automatic curation process (text mining)

Papers labeled xml

Predicted	
Object	Properties (y/n)

RegulonDB

Web

RegulonDB

EcoCyc

RegulonDB

RegulonDB

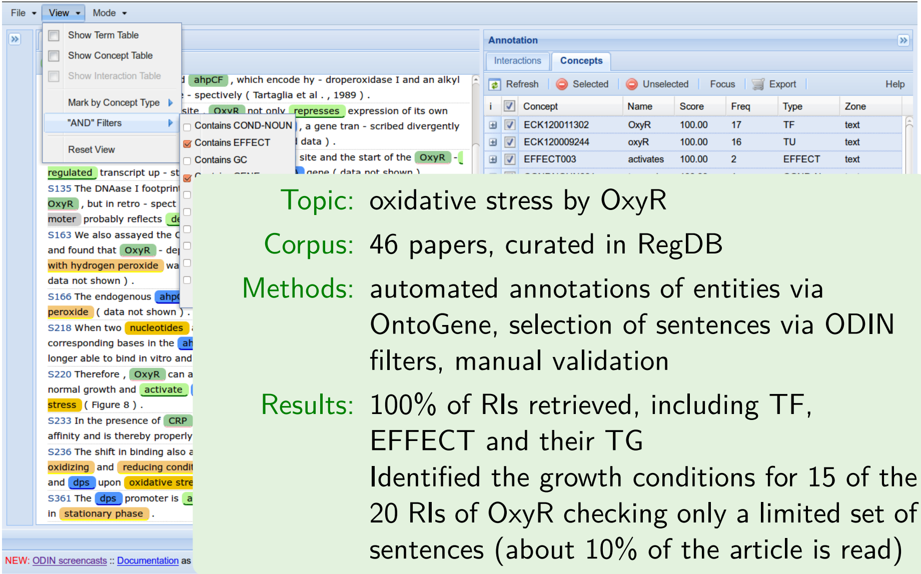
NIH National Institutes of Health
Turning Discovery Into Health

***Escherichia coli* K-12**
Transcriptional Regulatory Network

High-throughput literature curation of genetic regulation in bacterial models

- Funded by the NIH
- Grant ID: GM110597 (NIGMS-NIH)
- Funding: $1.6 million
- Duration: 4 years (Jan 2015 – Dec 2018)
- PI: Dr. Julio Collado-Vides (UNAM)
- Collaborators: Dr. Michael Savageau (UCDavis), Dr. Stephen Busby (Univ. of Birmingham), Dr. Fabio Rinaldi (Univ. Zurich)

Assisted Curation: Example



i	☑	Concept	Name	Score	Freq	Type	Zone
⊞	☑	ECK120011302	OxyR	100.00	17	TF	text
⊞	☑	ECK120009244	oxyR	100.00	16	TU	text
⊞	☑	EFFECT003	activates	100.00	2	EFFECT	text

d ahpCF , which encode hy - droperoxidase I and an alkyl
e - spectively (Tartaglia et al . , 1989) .
site . OxyR not only represses expression of its own
, a gene tran - scribed divergently
l data) .
site and the start of the OxyR -
gene (data not shown)

regulated transcript up - st
S135 The DNAase I footprin
OxyR , but in retro - spect
moter probably reflects de
S163 We also assayed the
and found that OxyR - de
with hydrogen peroxide wa
data not shown) .
S166 The endogenous ahp
peroxide (data not shown) .
S218 When two nucleotides
corresponding bases in the ah
longer able to bind in vitro and
S220 Therefore , OxyR can a
normal growth and activate
stress (Figure 8) .
S233 In the presence of CRP
affinity and is thereby properly
S236 The shift in binding also a
oxidizing and reducing condi
and dps upon oxidative stre
S361 The dps promoter is a
in stationary phase .

NEW: ODIN screencasts :: Documentation as

Topic: oxidative stress by OxyR

Corpus: 46 papers, curated in RegDB

Methods: automated annotations of entities via OntoGene, selection of sentences via ODIN filters, manual validation

Results: 100% of RIs retrieved, including TF, EFFECT and their TG

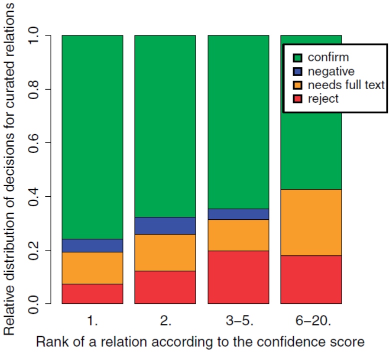
Identified the growth conditions for 15 of the 20 RIs of OxyR checking only a limited set of sentences (about 10% of the article is read)

Assisted Curation: ODIN Interface

- Lightweight browser-based graphical interface
- Purpose: literature-based curation tasks
- Coupled with OntoGene pipeline
- Easily customizable

Applications

- Novartis (2008–2012)
- PharmGKB (2011)
- CTD (2012)
- RegulonDB (2013)



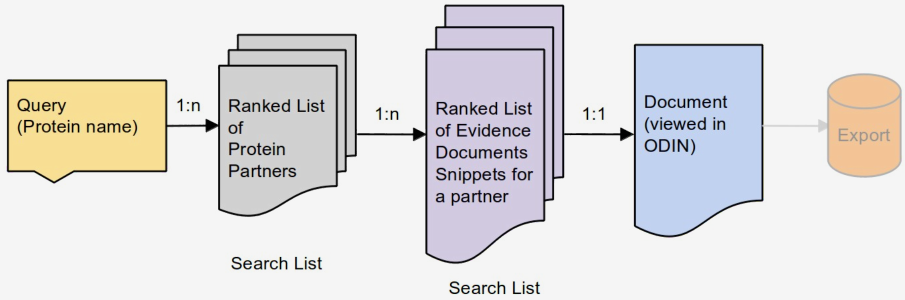
Outline

1. Motivation
2. Text mining in a curation workflow
3. Large-scale detection of protein interactions
4. Biomedical text mining: competitive evaluations
5. The OntoGene approach and highlights
6. Preview of recent work

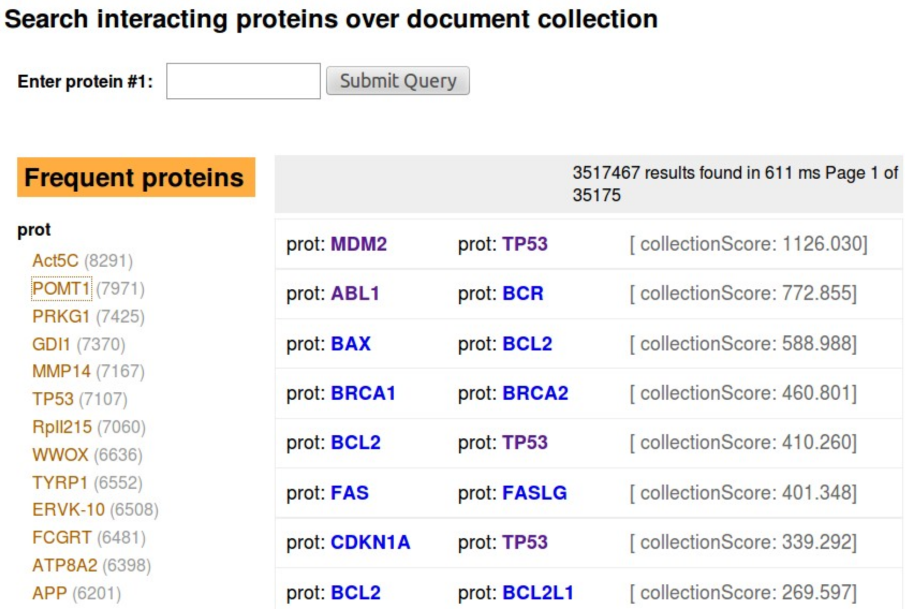
Large scale mining of protein interactions

- Text mining in an industrial context
- Concept filtering and relation ranking
- Collection-based ranking

Search interface built on Apache Solr


Query
(Protein name)
1:n
Ranked List of Protein Partners
Search List
1:n
Ranked List of Evidence Documents Snippets for a partner
Search List
1:1
Document (viewed in ODIN)
Export

Search Interface Apache Solr



Search interacting proteins over document collection

Frequent proteins

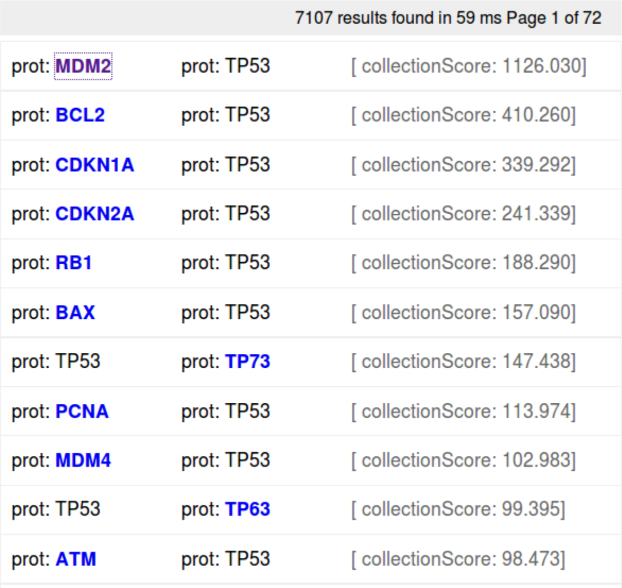
prot

Act5C	(8291)
POMT1	(7971)
PRKG1	(7425)
GDI1	(7370)
MMP14	(7167)
TP53	(7107)
Rpll215	(7060)
WWOX	(6636)
TYRP1	(6552)
ERVK-10	(6508)
FCGRT	(6481)
ATP8A2	(6398)
APP	(6201)

3517467 results found in 611 ms Page 1 of 35175

prot: **MDM2**	prot: **TP53**	[collectionScore: 1126.030]
prot: **ABL1**	prot: **BCR**	[collectionScore: 772.855]
prot: **BAX**	prot: **BCL2**	[collectionScore: 588.988]
prot: **BRCA1**	prot: **BRCA2**	[collectionScore: 460.801]
prot: **BCL2**	prot: **TP53**	[collectionScore: 410.260]
prot: **FAS**	prot: **FASLG**	[collectionScore: 401.348]
prot: **CDKN1A**	prot: **TP53**	[collectionScore: 339.292]
prot: **BCL2**	prot: **BCL2L1**	[collectionScore: 269.597]

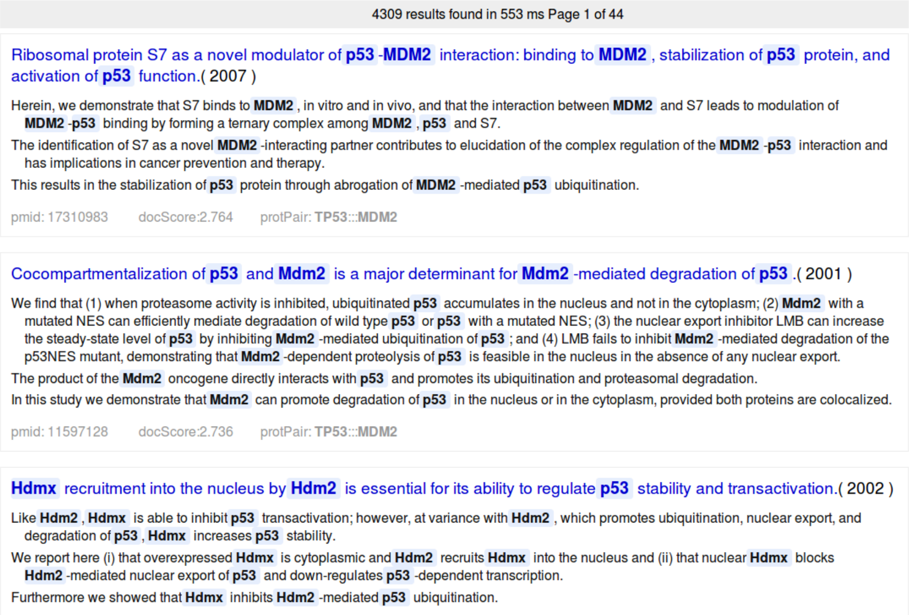
Search Interface Apache Solr



7107 results found in 59 ms Page 1 of 72

prot: **MDM2**	prot: TP53	[collectionScore: 1126.030]
prot: **BCL2**	prot: TP53	[collectionScore: 410.260]
prot: **CDKN1A**	prot: TP53	[collectionScore: 339.292]
prot: **CDKN2A**	prot: TP53	[collectionScore: 241.339]
prot: **RB1**	prot: TP53	[collectionScore: 188.290]
prot: **BAX**	prot: TP53	[collectionScore: 157.090]
prot: TP53	prot: **TP73**	[collectionScore: 147.438]
prot: **PCNA**	prot: TP53	[collectionScore: 113.974]
prot: **MDM4**	prot: TP53	[collectionScore: 102.983]
prot: TP53	prot: **TP63**	[collectionScore: 99.395]
prot: **ATM**	prot: TP53	[collectionScore: 98.473]

Search Interface Apache Solr



4309 results found in 553 ms Page 1 of 44

Ribosomal protein S7 as a novel modulator of **p53** -**MDM2** interaction: binding to **MDM2** , stabilization of **p53** protein, and activation of **p53** function.(2007)

Herein, we demonstrate that S7 binds to **MDM2** , in vitro and in vivo, and that the interaction between **MDM2** and S7 leads to modulation of **MDM2** -**p53** binding by forming a ternary complex among **MDM2** , **p53** and S7.

The identification of S7 as a novel **MDM2** -interacting partner contributes to elucidation of the complex regulation of the **MDM2** -**p53** interaction and has implications in cancer prevention and therapy.

This results in the stabilization of **p53** protein through abrogation of **MDM2** -mediated **p53** ubiquitination.

pmid: 17310983 docScore:2.764 protPair: **TP53:::MDM2**

Cocompartmentalization of **p53** and **Mdm2** is a major determinant for **Mdm2** -mediated degradation of **p53** .(2001)

We find that (1) when proteasome activity is inhibited, ubiquitinated **p53** accumulates in the nucleus and not in the cytoplasm; (2) **Mdm2** with a mutated NES can efficiently mediate degradation of wild type **p53** or **p53** with a mutated NES; (3) the nuclear export inhibitor LMB can increase the steady-state level of **p53** by inhibiting **Mdm2** -mediated ubiquitination of **p53** ; and (4) LMB fails to inhibit **Mdm2** -mediated degradation of the p53NES mutant, demonstrating that **Mdm2** -dependent proteolysis of **p53** is feasible in the nucleus in the absence of any nuclear export.

The product of the **Mdm2** oncogene directly interacts with **p53** and promotes its ubiquitination and proteasomal degradation.

In this study we demonstrate that **Mdm2** can promote degradation of **p53** in the nucleus or in the cytoplasm, provided both proteins are colocalized.

pmid: 11597128 docScore:2.736 protPair: **TP53:::MDM2**

Hdmx recruitment into the nucleus by **Hdm2** is essential for its ability to regulate **p53** stability and transactivation.(2002)

Like **Hdm2** , **Hdmx** is able to inhibit **p53** transactivation; however, at variance with **Hdm2** , which promotes ubiquitination, nuclear export, and degradation of **p53** , **Hdmx** increases **p53** stability.

We report here (i) that overexpressed **Hdmx** is cytoplasmic and **Hdm2** recruits **Hdmx** into the nucleus and (ii) that nuclear **Hdmx** blocks **Hdm2** -mediated nuclear export of **p53** and down-regulates **p53** -dependent transcription.

Furthermore we showed that **Hdmx** inhibits **Hdm2** -mediated **p53** ubiquitination.

Outline

1. Motivation

2. Text mining in a curation workflow

3. Large-scale detection of protein interactions

4. Biomedical text mining: competitive evaluations

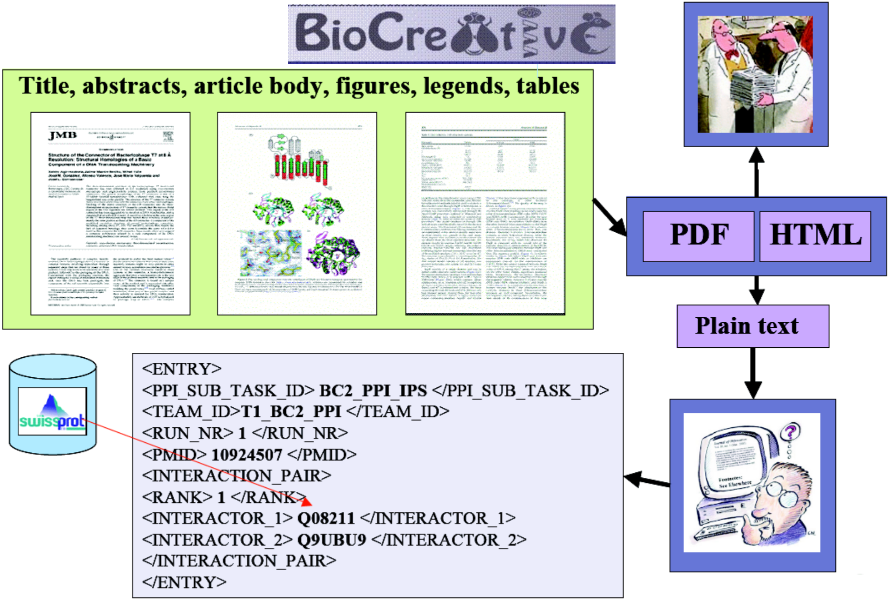
5. The OntoGene approach and highlights

6. Preview of recent work

Competitive Evaluations

- BioCreative
- BioNLP
- CALBC
- CLEF-ER
- QA4MRE
- DDI @ Semeval
- BioASQ
- I2B2

BioCreative Shared Task



```
<ENTRY>
<PPI_SUB_TASK_ID> BC2_PPI_IPS </PPI_SUB_TASK_ID>
<TEAM_ID>T1_BC2_PPI </TEAM_ID>
<RUN_NR> 1 </RUN_NR>
<PMID> 10924507 </PMID>
<INTERACTION_PAIR>
<RANK> 1 </RANK>
<INTERACTOR_1> Q08211 </INTERACTOR_1>
<INTERACTOR_2> Q9UBU9 </INTERACTOR_2>
</INTERACTION_PAIR>
</ENTRY>
```

Martin Krallinger/Florian Leitner

BioCreative Shared Task

- 2004 (I) gene mentions, GO annotations
- 2006 (II) GM, GN, PPI
- 2009 (II.5) PPI
- 2010 (III) GN, PPI-ACT, PPI-IMT, IAT
- 2012 CTD-triage, curation workflow, IAT
- 2013 (IV) BioC, CHEMDNER, CTD, GO, IAT
- 2015 (V) BioC, CHEMDNER, Chem/Dis, BEL, IAT

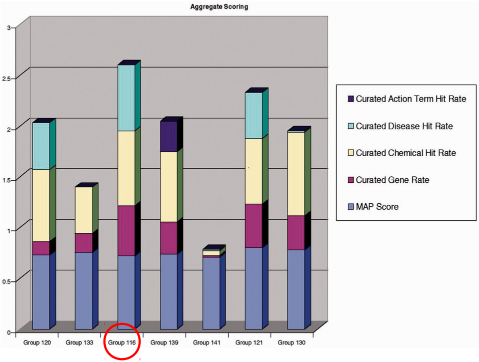
CTD-triage

Purpose

“promotes understanding about the effects of environmental chemicals on human health by integrating data from curated scientific literature”

Task

entity extraction and triage

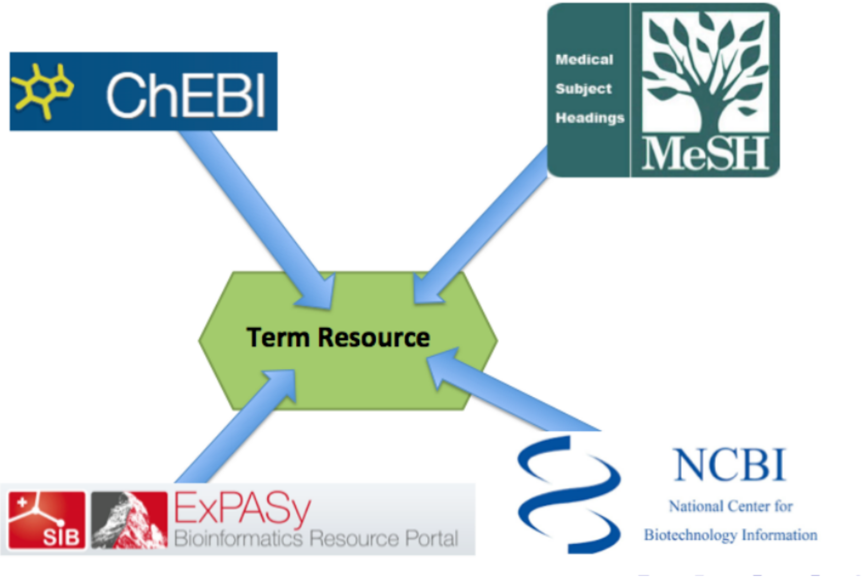


Best overall results,
best detection of genes and diseases

Outline

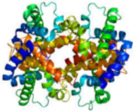
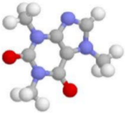
1. Motivation
2. Text mining in a curation workflow
3. Large-scale detection of protein interactions
4. Biomedical text mining: competitive evaluations
5. The OntoGene approach and highlights
6. Preview of recent work

Single Term Resource



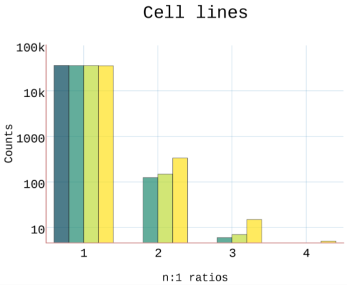
Single Term Resource: Entities

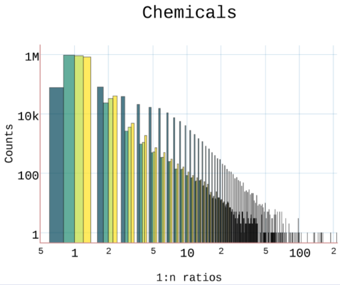
- Genes and proteins (NCBI gene, UniProt)
- Chemicals (MeSH, ChEBI, CTD)
- Diseases (MeSH, CTD)
- Organism and species (MeSH, NCBI taxonomy)
- Cell lines (Cellosaurus)

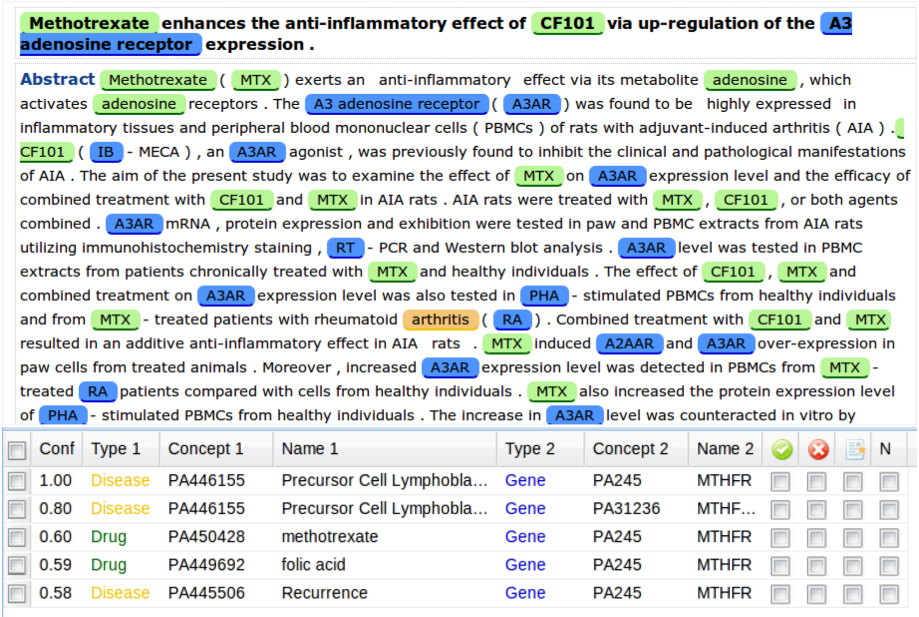
Term Resource: Some Figures

	genes/ proteins	chemicals	diseases	species	cell lines	total
count	**10.4 M**	979 k	67 k	1.3 M	36 k	12.8 M
avg. length	11.73	**37.49**	26.98	22.87	**7.611**	14.92
terms/ID	1.1455	**3.545**	**6.018**	1.326	**1.000**	1.328
IDs/term	**1.371**	1.049	**1.000**	1.003	1.004	1.306





Term Resource: Annotated Document in ODIN

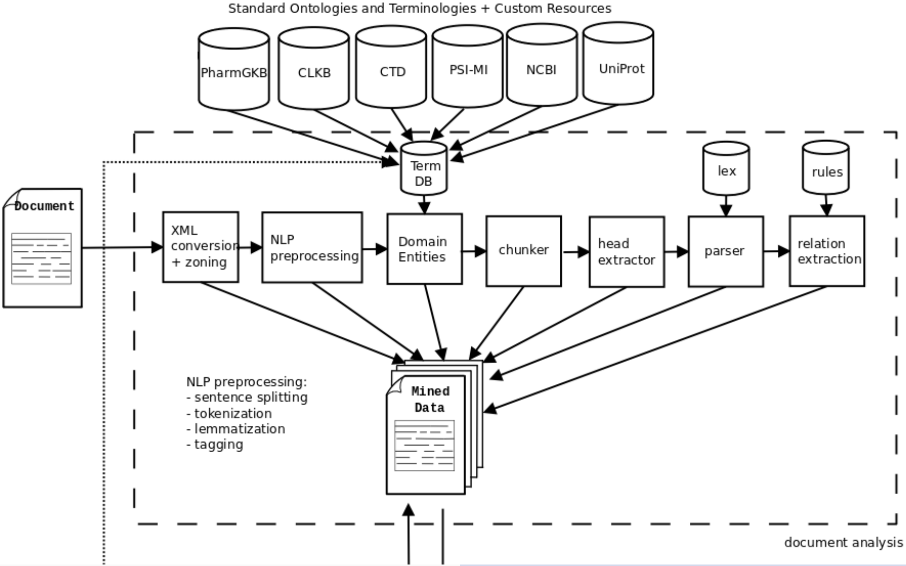


Methotrexate enhances the anti-inflammatory effect of CF101 via up-regulation of the A3 adenosine receptor expression .

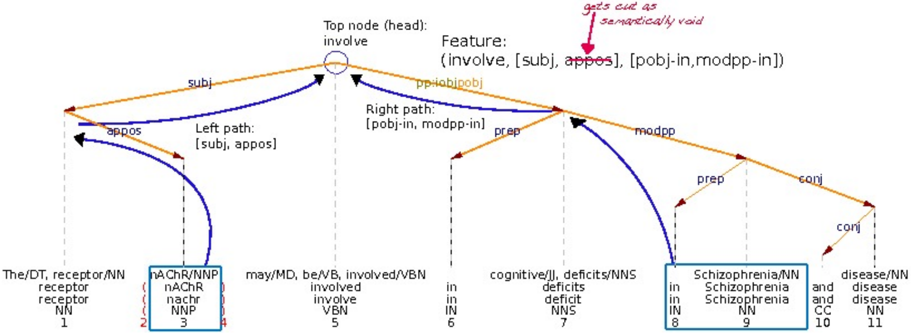
Abstract Methotrexate (MTX) exerts an anti-inflammatory effect via its metabolite adenosine , which activates adenosine receptors . The A3 adenosine receptor (A3AR) was found to be highly expressed in inflammatory tissues and peripheral blood mononuclear cells (PBMCs) of rats with adjuvant-induced arthritis (AIA) . CF101 (IB - MECA) , an A3AR agonist , was previously found to inhibit the clinical and pathological manifestations of AIA . The aim of the present study was to examine the effect of MTX on A3AR expression level and the efficacy of combined treatment with CF101 and MTX in AIA rats . AIA rats were treated with MTX , CF101 , or both agents combined . A3AR mRNA , protein expression and exhibition were tested in paw and PBMC extracts from AIA rats utilizing immunohistochemistry staining , RT - PCR and Western blot analysis . A3AR level was tested in PBMC extracts from patients chronically treated with MTX and healthy individuals . The effect of CF101 , MTX and combined treatment on A3AR expression level was also tested in PHA - stimulated PBMCs from healthy individuals and from MTX - treated patients with rheumatoid arthritis (RA) . Combined treatment with CF101 and MTX resulted in an additive anti-inflammatory effect in AIA rats . MTX induced A2AAR and A3AR over-expression in paw cells from treated animals . Moreover , increased A3AR expression level was detected in PBMCs from MTX - treated RA patients compared with cells from healthy individuals . MTX also increased the protein expression level of PHA - stimulated PBMCs from healthy individuals . The increase in A3AR level was counteracted in vitro by

□	Conf	Type 1	Concept 1	Name 1	Type 2	Concept 2	Name 2	✓	✗	📝	N
□	1.00	Disease	PA446155	Precursor Cell Lymphobla...	Gene	PA245	MTHFR	□	□	□	□
□	0.80	Disease	PA446155	Precursor Cell Lymphobla...	Gene	PA31236	MTHF...	□	□	□	□
□	0.60	Drug	PA450428	methotrexate	Gene	PA245	MTHFR	□	□	□	□
□	0.59	Drug	PA449692	folic acid	Gene	PA245	MTHFR	□	□	□	□
□	0.58	Disease	PA445506	Recurrence	Gene	PA245	MTHFR	□	□	□	□

Relation Extraction Pipeline



Syntax-based Approach



gets cut as semantically void

Feature:
(involve, [subj, ~~appos~~], [pobj-in,modpp-in])

“The neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7) may be involved in cognitive deficits in Schizophrenia and Alzheimer’s disease.” [PMID 15695160]

Highlights

- [2006] BioCreative II: PPI (3rd), IMT (best)
- [2009] BioCreative II.5 PPI (best results); BioNLP
- [2010] BioCreative III: ACT, IMT, IAT
- [2011] CALBC (large scale entity extraction), BioNLP
- [2012] CTD task at BioCreative 2012
- [2013] BioCreative IV: BioC, CTD, IAT
- 80+ publications, 20+ journal articles

Outline

1. Motivation
2. Text mining in a curation workflow
3. Large-scale detection of protein interactions
4. Biomedical text mining: competitive evaluations
5. The OntoGene approach and highlights
6. Preview of recent work

Veterinary Pathology Text Mining

University of Zurich^UZH

Collaboration with the veterinary faculty of the University of Bern

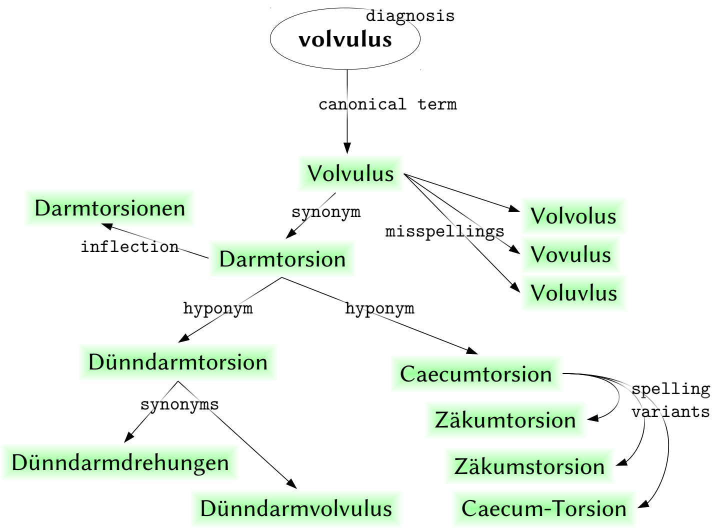
VETERINARY • PUBLIC • HEALTH • INSTITUTE

Task

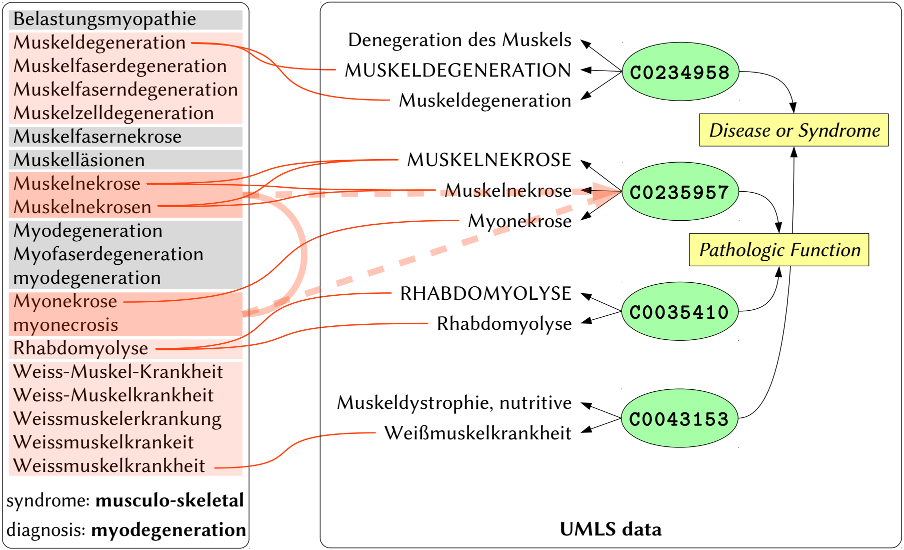
Development and evaluation of an automated text-mining and syndrome-classifying tool:

- extract relevant information from pathology reports with minimal expert intervention
- classify pathology findings into syndromic groups to enhance the efficiency of health event detection

Veterinary TM: Term Variation



Veterinary TM: Enhancing Terminology



Belastungsmyopathie

Muskeldegeneration
Muskelfaserdegeneration
Muskelfaserndegeneration
Muskelzelldegeneration

Muskelfasernekrose

Muskelläsionen

Muskelnekrose
Muskelnekrosen

Myodegeneration
Myofaserdegeneration
myodegeneration

Myonekrose
myonecrosis

Rhabdomyolyse

Weiss-Muskel-Krankheit
Weiss-Muskelkrankheit
Weissmuskelerkrankung
Weissmuskelkrankeit
Weissmuskelkrankheit

syndrome: **musculo-skeletal**

diagnosis: **myodegeneration**

Denegeration des Muskels
MUSKELDEGENERATION
Muskeldegeneration
C0234958

Disease or Syndrome

MUSKELNEKROSE
Muskelnekrose
Myonekrose
C0235957

Pathologic Function

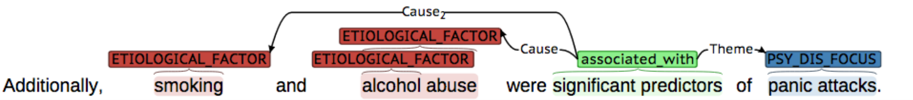
RHABDOMYOLYSE
Rhabdomyolyse
C0035410

Muskeldystrophie, nutritive
Weißmuskelkrankheit
C0043153

UMLS data

PsyMine

the cogito foundation



Additionally, smoking and alcohol abuse were significant predictors of panic attacks.

Text mining in support of psychiatric research: overcoming fragmented knowledge

Collaboration with the Compentence Center for Mental Health at the Epidemiology, Biostatistics and Prevention Institute

- Goal: identify potential causes of mental diseases
- Methods: analyse the whole biomedical literature, identify causes of mental disorders (genetic/disease/social), rank and correlate
- Vision: “global overview” of knowledge in literature

MelanoBase

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Large-scale automatic extraction of actionable information from the biomedical literature

- integration with existing structured knowledge
- use-case scenario: melanoma
- results to be integrated within the Melanoma Molecular Map repository (S. Mocellin, Padua)
- collaborations with clinical researchers (Marisol Soengas, CNIO, Spain).

Summary

- Text mining technologies can provide an effective support in biomedical curation
- ODIN is a user-friendly tool for text-mining supporting interactive (collaborative) curation of the biomedical literature.
- OntoGene provides competitive text mining technologies (BioCreative, CALBC prove quality)
- New projects and applications: VetSuisse, PsyMine, MelanoBase