#BLAHmuc
Biomedical Linked Annotation
Hackathon

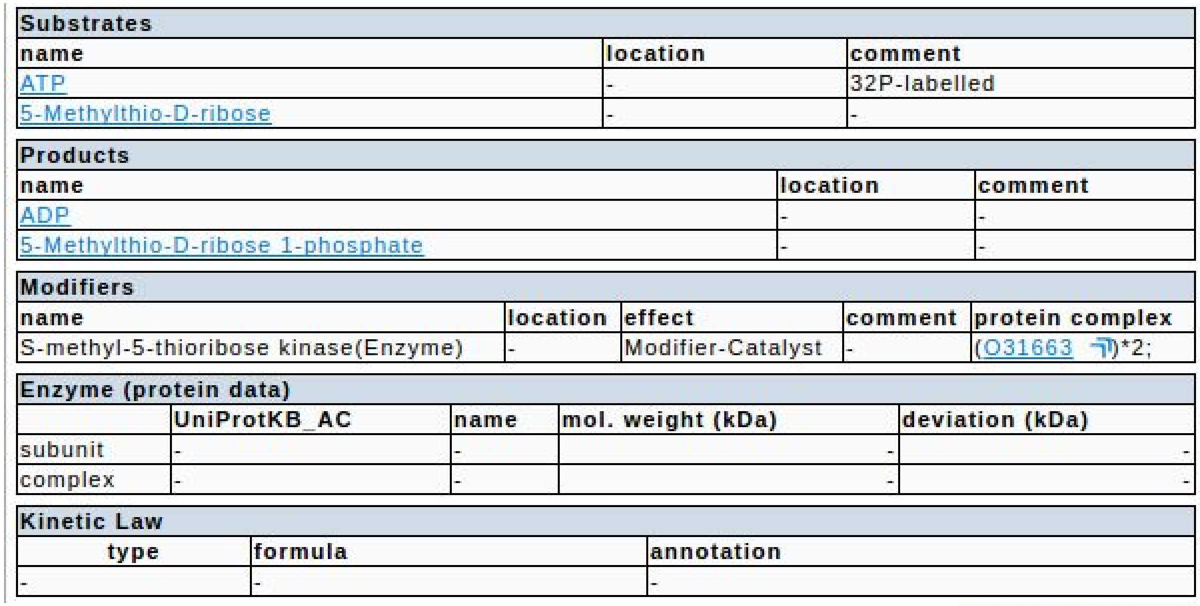
Text Mining to Support Data Curation for SABIO-RK

Mariana Neves, Maja Rey, Ulrike Wittig

September 19th, 2016

SABIO-RK database

- It contains structured information about biochemical reactions and their corresponding kinetics
- It describes participants and modifiers of the reactions, as well as measured kinetic data embedded in their experimental and environmental context



Substrates

name	location	comment
ATP	-	32P-labelled
5-Methylthio-D-ribose	-	-

Products

name	location	comment
ADP	-	-
5-Methylthio-D-ribose 1-phosphate	-	-

Modifiers

name	location	effect	comment	protein complex
S-methyl-5-thioribose kinase(Enzyme)	-	Modifier-Catalyst	-	(O31663)*2;

Enzyme (protein data)

	UniProtKB_AC	name	mol. weight (kDa)	deviation (kDa)
subunit	-	-	-	-
complex	-	-	-	-

Kinetic Law

type	formula	annotation
-	-	-

Heidelberg Institute for Theoretical Studies

HITS

SABIO-RK
Biochemical Reaction Kinetics Database

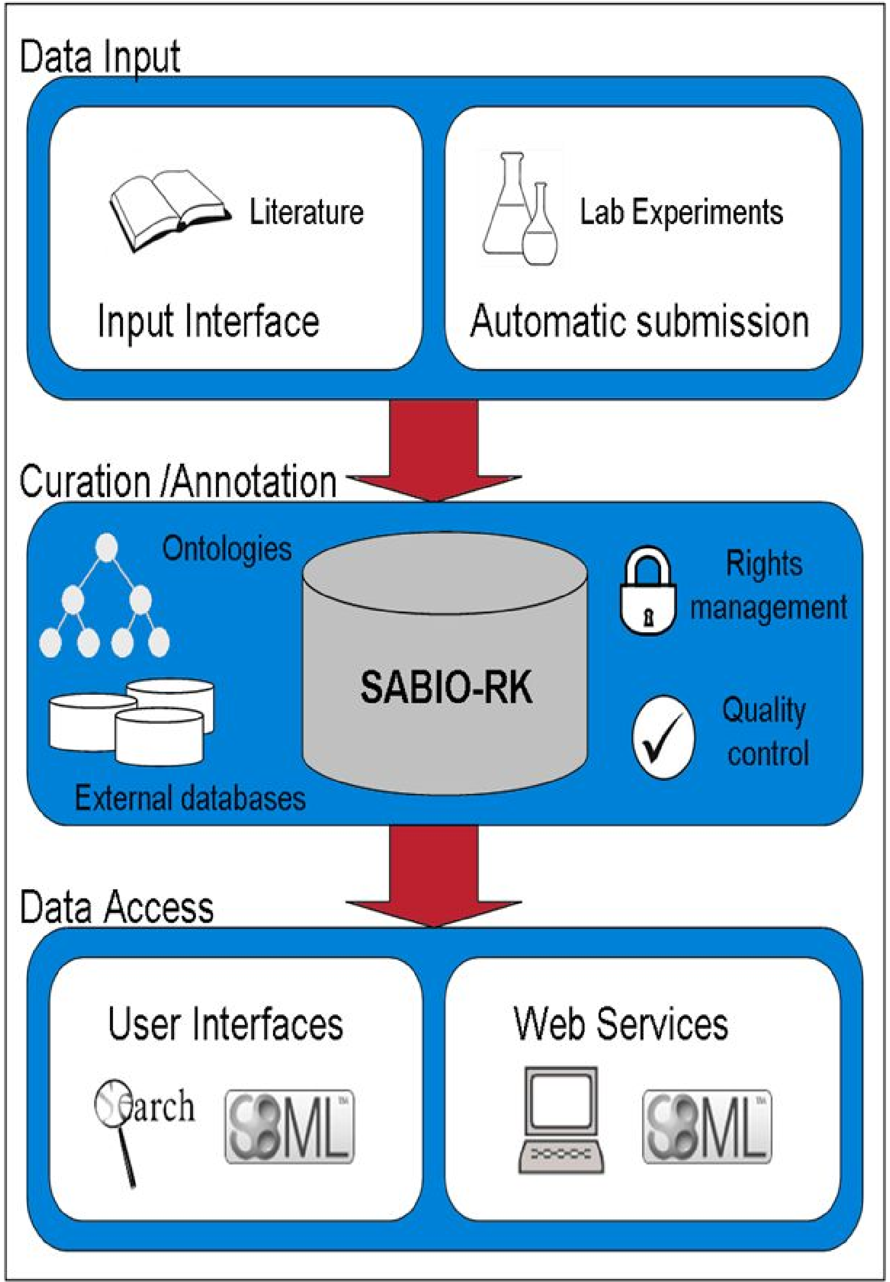
Workflow in SABIO-RK

99.8 % from publications (ca. 5300), by students

100 % Rereading and curating by experts

http://sabiork.h-its.org

Heidelberg Institute for Theoretical Studies
HITS

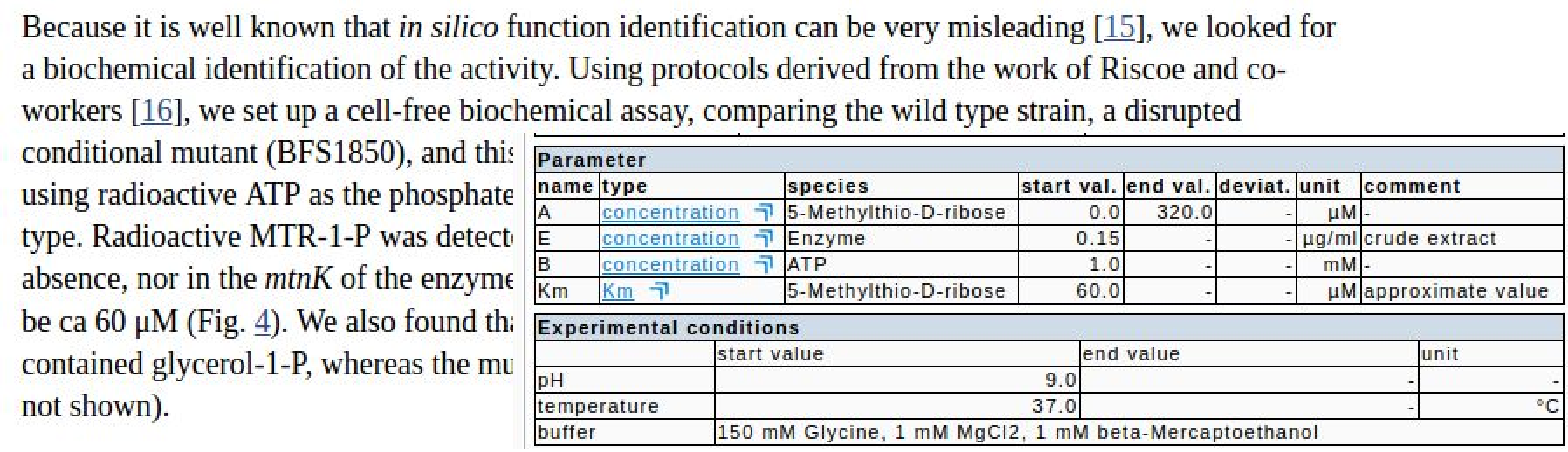


0.2 % from direct submissions, experiments or models

Data are:
- complete
- structured
- standardized
- linked
- annotated

SABIO-RK database

- Only scientific publications containing kinetic data such as Km, kcat or
  Vmax values are included in the database



Because it is well known that *in silico* function identification can be very misleading [15], we looked for a biochemical identification of the activity. Using protocols derived from the work of Riscoe and co-workers [16], we set up a cell-free biochemical assay, comparing the wild type strain, a disrupted
conditional mutant (BFS1850), and thi
using radioactive ATP as the phosphate
type. Radioactive MTR-1-P was detect
absence, nor in the *mtnK* of the enzyme
be ca 60 µM (Fig. 4). We also found th
contained glycerol-1-P, whereas the mu
not shown).

Parameter							
name	**type**	**species**	**start val.**	**end val.**	**deviat.**	**unit**	**comment**
A	concentration	5-Methylthio-D-ribose	0.0	320.0	-	µM	-
E	concentration	Enzyme	0.15	-	-	µg/ml	crude extract
B	concentration	ATP	1.0	-	-	mM	-
Km	Km	5-Methylthio-D-ribose	60.0	-	-	µM	approximate value

Experimental conditions			
	start value	end value	unit
pH	9.0	-	-
temperature	37.0	-	°C
buffer	150 mM Glycine, 1 mM MgCl2, 1 mM beta-Mercaptoethanol		

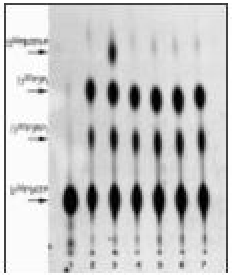
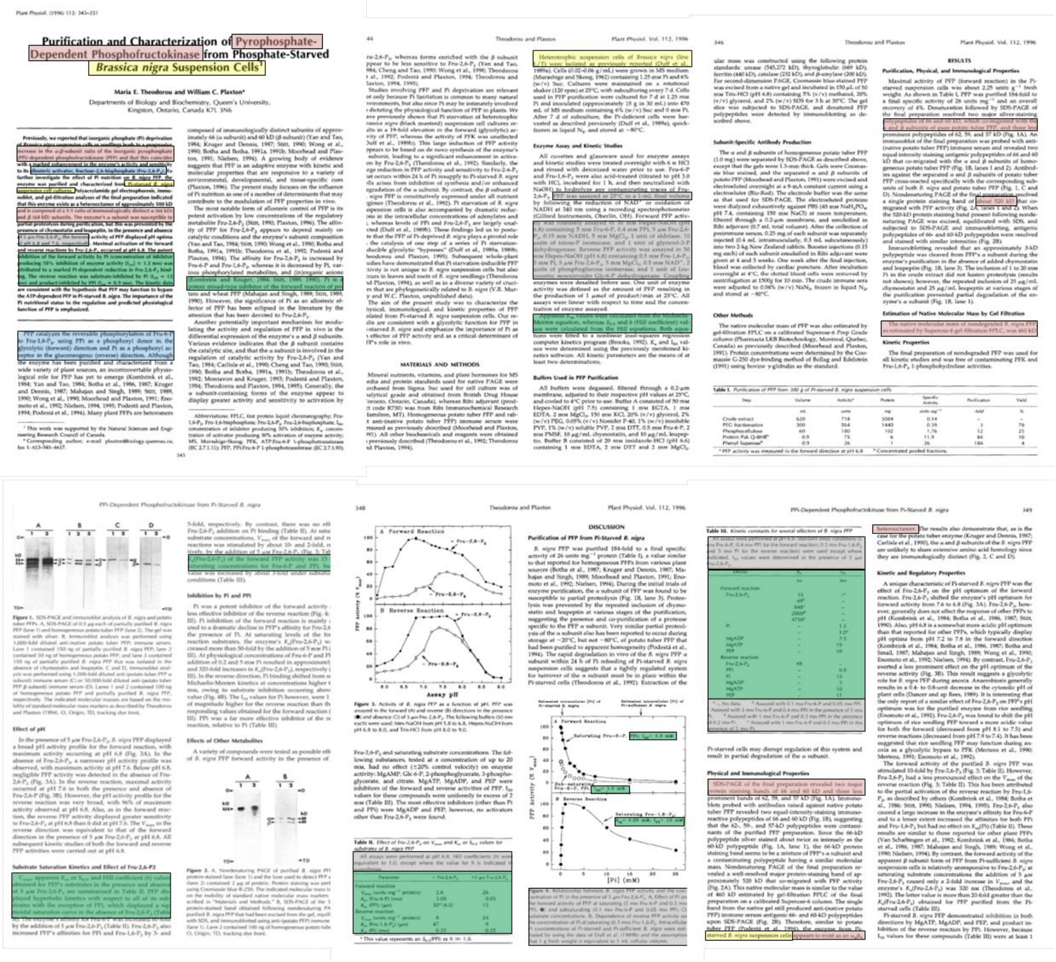


Figure 3

Autoradiograph showing the outcome of the MTR kinase assay. The assay was carried out for 90 min at 37°C (see Materials and Methods). Lane 1 corresponds to [•^{32}P]ATP as standard; lanes 2, 4 and 6: no MTR was added for the reactions with ...

Distribution of relevant data within a publication



Difficulties in data extraction for SABIO-RK:

- Relevant data is scattered throughout the publication within free text, tables, figures
- Identifying data relations
- No defined data format
- Missing information (EC, Temp...)
- Inconsistencies
- Referenced data must be excluded

🟥 Protein and enzyme data

🟨 Biological source (organism, tissue, cell type)

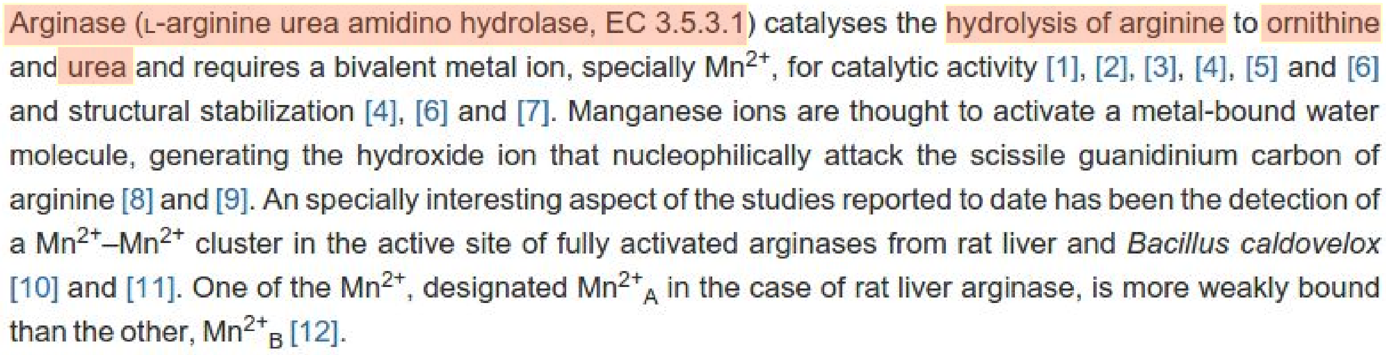
🟦 Reactions and chemical compounds

🟩 Kinetic data

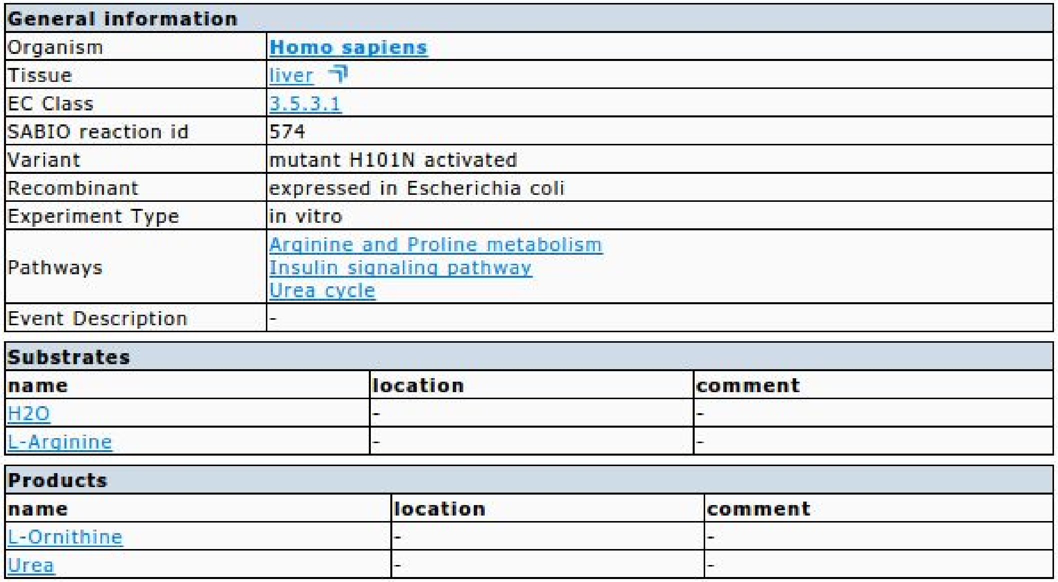
⬜ Experimental conditions

[Wittig U, Kania R, Bittkowski M, Wetsch E, Shi L, Jong L, Golebiewski M, Rey M, Weidemann A, Rojas I, Müller W. Data extraction for the reaction kinetics database SABIO-RK, Perspectives in Science, (2014) 1, 33–40.]

Data curation in SABIO-RK



Arginase (L-arginine urea amidino hydrolase, EC 3.5.3.1) catalyses the hydrolysis of arginine to ornithine and urea and requires a bivalent metal ion, specially Mn^{2+}, for catalytic activity [1], [2], [3], [4], [5] and [6] and structural stabilization [4], [6] and [7]. Manganese ions are thought to activate a metal-bound water molecule, generating the hydroxide ion that nucleophilically attack the scissile guanidinium carbon of arginine [8] and [9]. An specially interesting aspect of the studies reported to date has been the detection of a Mn^{2+}–Mn^{2+} cluster in the active site of fully activated arginases from rat liver and *Bacillus caldovelox* [10] and [11]. One of the Mn^{2+}, designated Mn$^{2+}{}_{A}$ in the case of rat liver arginase, is more weakly bound than the other, Mn$^{2+}{}_{B}$ [12].



General information	
Organism	**Homo sapiens**
Tissue	liver
EC Class	3.5.3.1
SABIO reaction id	574
Variant	mutant H101N activated
Recombinant	expressed in Escherichia coli
Experiment Type	in vitro
Pathways	Arginine and Proline metabolism Insulin signaling pathway Urea cycle
Event Description	-

Substrates

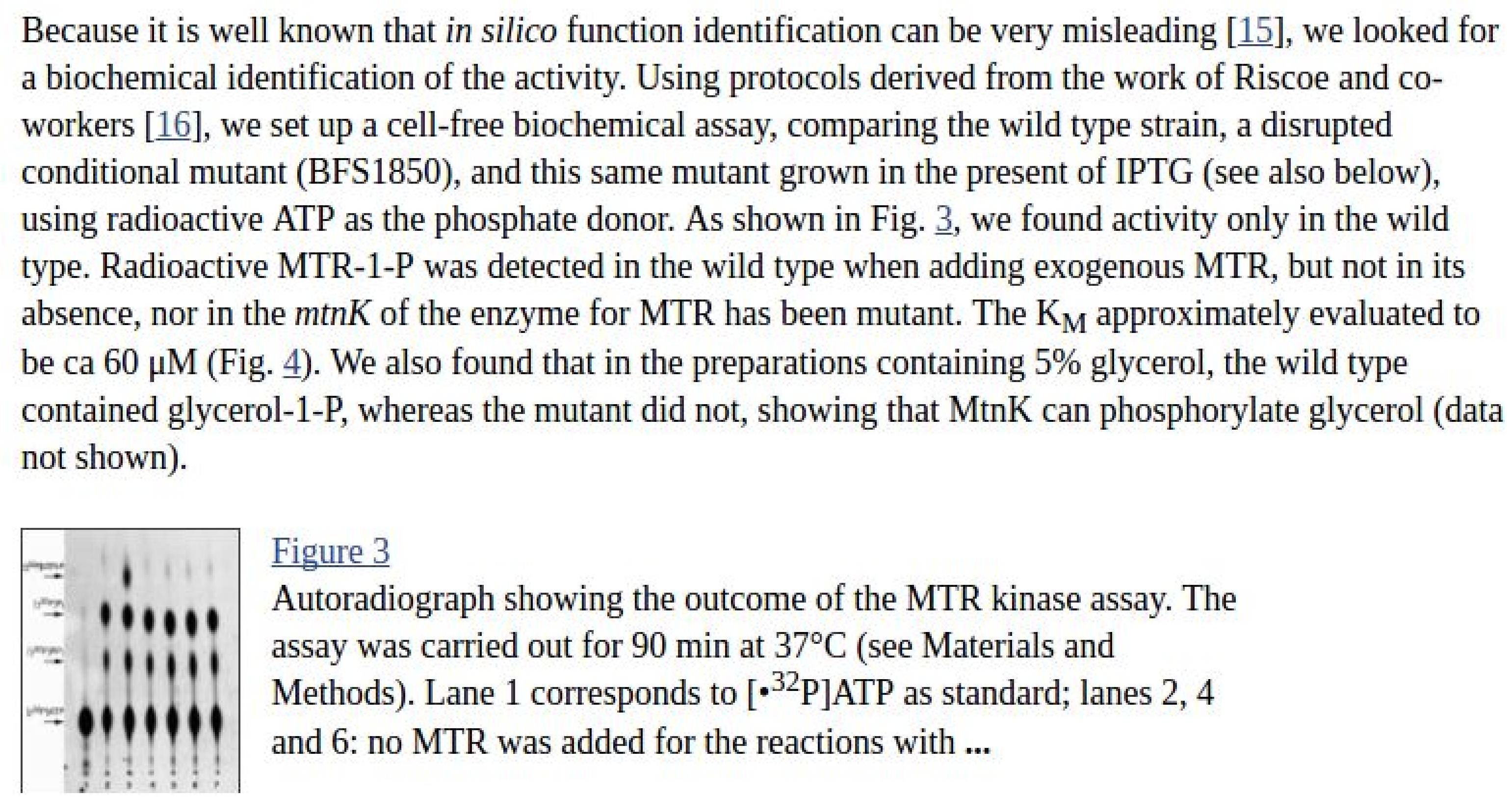
name	location	comment
H2O	-	-
L-Arginine	-	-

Products

name	location	comment
L-Ornithine	-	-
Urea	-	-

Triage in SABIO-RK database

- Search should be based on the full text



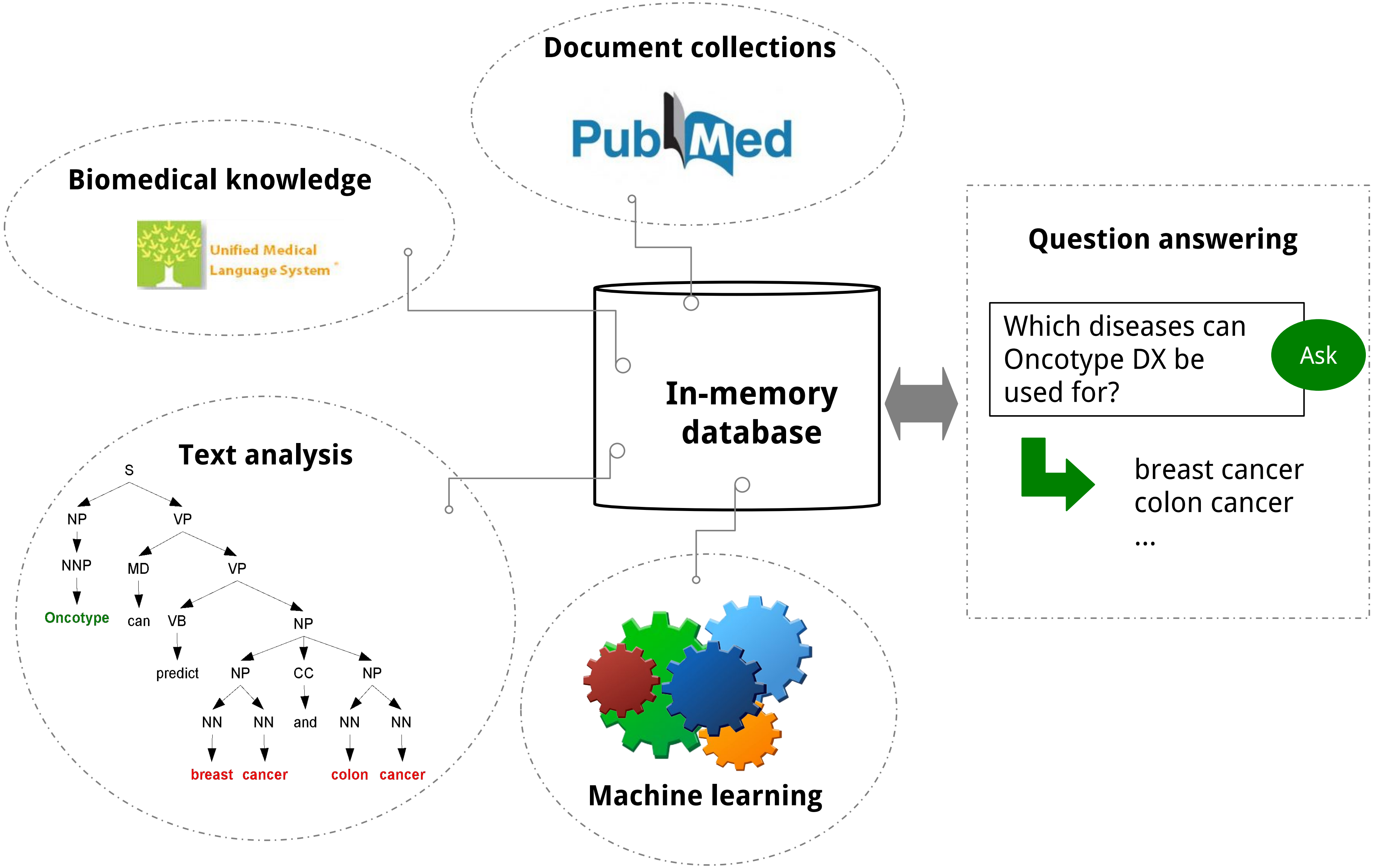
Because it is well known that *in silico* function identification can be very misleading [15], we looked for a biochemical identification of the activity. Using protocols derived from the work of Riscoe and co-workers [16], we set up a cell-free biochemical assay, comparing the wild type strain, a disrupted conditional mutant (BFS1850), and this same mutant grown in the present of IPTG (see also below), using radioactive ATP as the phosphate donor. As shown in Fig. 3, we found activity only in the wild type. Radioactive MTR-1-P was detected in the wild type when adding exogenous MTR, but not in its absence, nor in the *mtnK* of the enzyme for MTR has been mutant. The K_M approximately evaluated to be ca 60 μM (Fig. 4). We also found that in the preparations containing 5% glycerol, the wild type contained glycerol-1-P, whereas the mutant did not, showing that MtnK can phosphorylate glycerol (data not shown).

Figure 3

Autoradiograph showing the outcome of the MTR kinase assay. The assay was carried out for 90 min at 37°C (see Materials and Methods). Lane 1 corresponds to [•^{32}P]ATP as standard; lanes 2, 4 and 6: no MTR was added for the reactions with ...

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC55331/)

Text Mining platform at the HPI



Document collections

PubMed

Biomedical knowledge

Unified Medical Language System

In-memory database

Text analysis

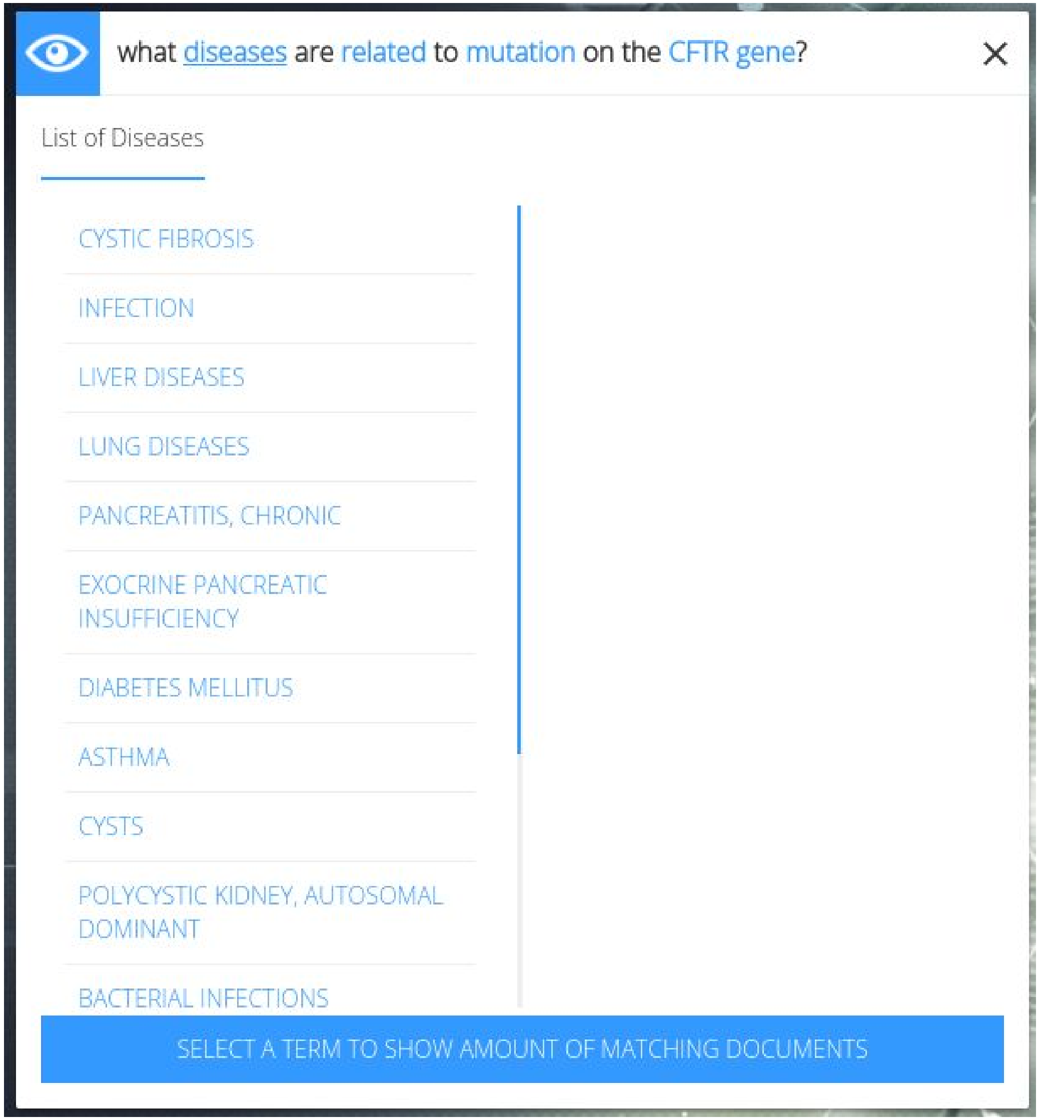
Machine learning

Question answering

Which diseases can Oncotype DX be used for?

breast cancer
colon cancer
…

Text mining platform



what diseases are related to mutation on the CFTR gene?

List of Diseases

- CYSTIC FIBROSIS
- INFECTION
- LIVER DISEASES
- LUNG DISEASES
- PANCREATITIS, CHRONIC
- EXOCRINE PANCREATIC INSUFFICIENCY
- DIABETES MELLITUS
- ASTHMA
- CYSTS
- POLYCYSTIC KIDNEY, AUTOSOMAL DOMINANT
- BACTERIAL INFECTIONS

Resources in Medicate

MeSH

Medical Subject Headings

Vocabulary thesaurus

229.000

PubMed

Medical Publications

References to articles from biomedical journals

16.000.000

UMLS

Unified Medical Language System

Health and biomedical vocabulary

7.400.000

PMC

PubMed Central

Open access full-text archive of biomedical literature

1.300.000

(slide from the Bachelor Project HP2)

Natural Language Processing

? What are treatments for hair loss?

↓ treatments (Token)

↓ hair loss (Token)

?

What	are	treatments	for	hair	loss?
(pronoun)	(verb)	(noun)	(preposition)	(noun)	(noun)

? Connection
Connected → connect

1. Question Answering
2. Named Entity Recognition
3. Part-of-speech tagging
4. Stemming
5. Summarization

(slide from the Bachelor Project HP2)

Tasks in the hackathon

- Analysis of the curation needs for the SABIO-RK database
- Evaluation of the HPI text mining platform
- Reproducibility of existing data

Thank you for your attention! Questions?

- E-mail: Mariana.Neves@hpi.de
- Web site: http://hpi.de/plattner/projects/in-memory-natural-language-processing.html
- Twitter: @marianalraneves