SWAT4LS, Berlin, 10 Dec. 2010

A renaissance for the point mutation:

from legacy data to semantic web services .....

Christopher J. O. Baker

University of New Brunswick

Outline

- Problem Statement
- Mutation Databases
- Text Mining
- Grounding
  - Mutations
  - Molecular Functions
- Mutation Impact Extraction
  - NAR Subset
- Performance Metrics
- Reuse Scenarios
- Impact Prediction Data
- 3 DEMOS reusing mutation annotations and impacts
- Knowlegator / mSTRAP
- PubMed Semantic Assistant
- SADI Web Services
  - 3 sample queries

Data Integration Challenges:

(i) publishing of salient mutation impact descriptions in unstructured text.
(ii) prevalence of boutique databases of mutation information with a many years of latency.
(iii) errors within manually populated mutation databases.
(iv) mining of mutations from scientific documents for denovo database creation.

Types of Mutations

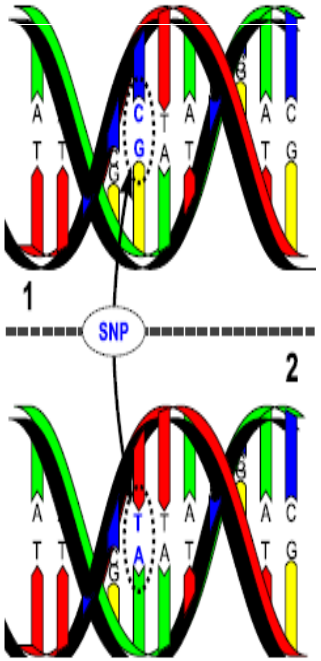
- Genotype
- Missense/nonsense
- Splicing
- Regulatory
- Small deletions
- Small insertions
- Gross deletions
- Gross insertions
- Repeat variations

- Single Amino Acid polymorphisms (exp.)
- SNP / nsSNP

Single Nucleotide Polymorphism (SNP)

Wikipedia[1]:

- "DNA sequence variation occurring when a single nucleotide—A, T, C, or G—in the genome differs between members of a species"
- "Successful" point mutations in a population: > 1%



[1] http://de.wikipedia.org/wiki/Single_Nucleotide_Polymorphism

Data Types Stored

- Residues
- Coordinates
- DNA / Protein
- Publication Links
- Experimental conditions
- Organisms

- Impact Annotations
  - Phenotype
  - Modified Properties
    - Stability,
    - Protein Interaction
    - Molecule conformation
    - Pathway Dynamics
  - Units of measurement

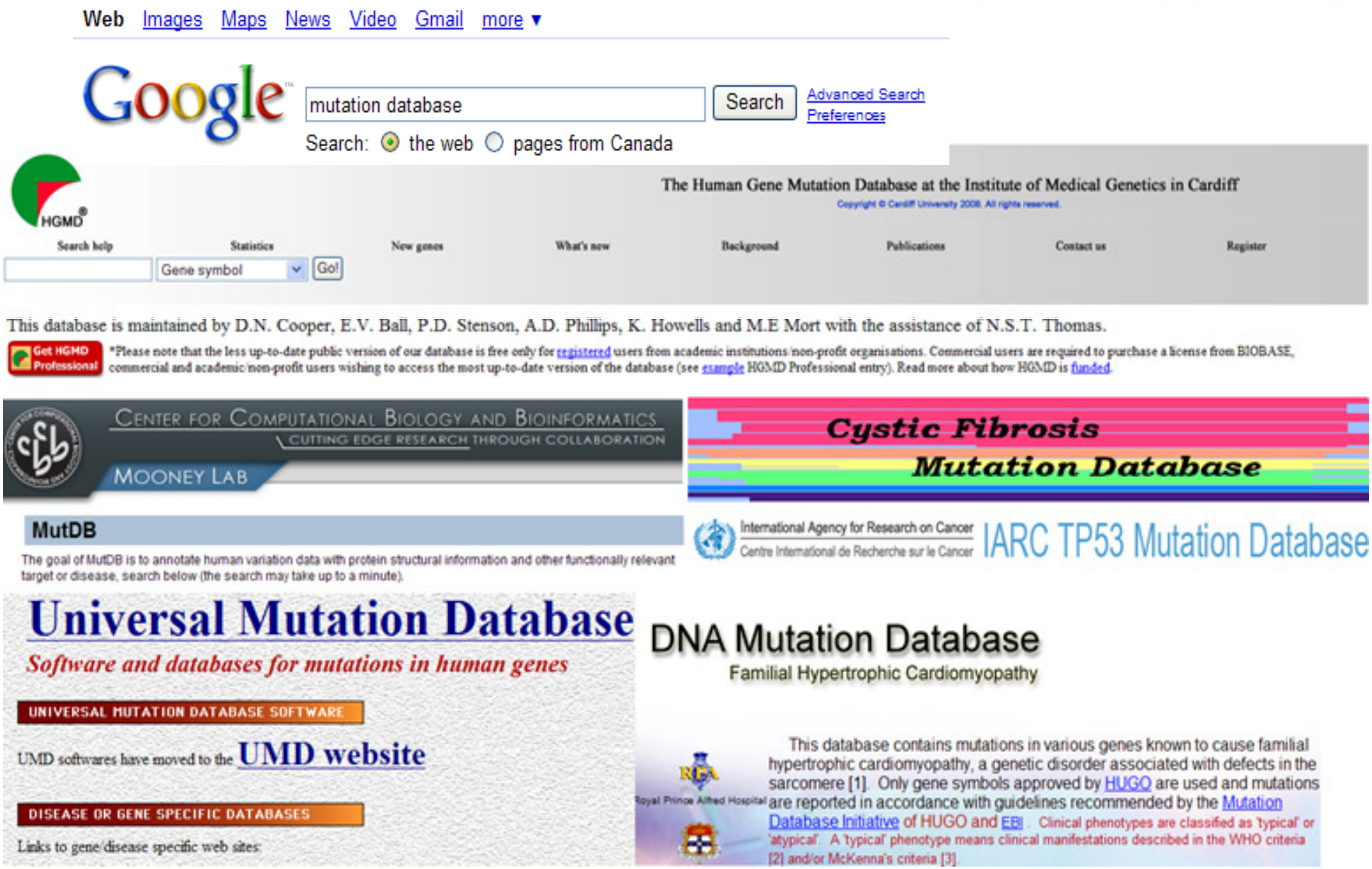
Relevance of non-synonymous SNPs

- Changes to the protein lead to changes in
  - molecular function (e.g., impaired signaling)
  - metabolism (e.g., cystic fibrosis)
  - cellular phenotype (e.g., neurofilament aggregation)
  - physiological phenotype (e.g., QT prolongation)
  - the human being
- SNPs make human beings different
  - The collection of SNPs form the haplotype
  - SNPs and haplotypes are the basis to understand human variability

Future Relevance of SNPs

- Increasingly single individuals scanned for their genotypic variability
- Comparison against standard genomes:
  - Human genome project
  - 1,000 genomes project (www.1000genomes.org)
- Commercialization of sequencing
  - High speed and cost efficiency
  - Sequencing of individuals for a reasonable amount of money
- **=> ….. How do we make sense of the data and reuse it …..**

400 + Mutation Databases



Mutation Databases

- Specialized databases:
  - focus on an individual disease / phenotype
- General databases:
  - Omim (NCBI): monogenetic diseases, keeping track of genetic variability and disease implication, gathered from the literature
  - GAD: SNPs from associations studies, gathered from the literature – identifies medically relevant polymorphism from the large volume of polymorphism and mutational data
  - dbSNP (NCBI): broad range of SNPs - 51,312,474 variations for 43 different organisms

Different in scope and scale

	Diseases	Distinct genes	Associations
MIM	1864	2708	4851
GAD	2303	1740	7945

- Omim makes reference to less diseases and to more genes
- GAD makes reference to more diseases, gene-disease associations, but monitors less genes

Mining Mutation Annotations

- From databases
  - High number of specialized databases
  - No uniform format => data mining is difficult
- From the literature:
  - Advantage:
    - Contextual information supports interpretation
    - Universal resource => different types of SNPs available, i.e. disease related and experimental SNPs
  - Disadvantage:
    - Normalisation of SNPs is not straight forward
    - Language variability

Maintenance of Mutation Databases

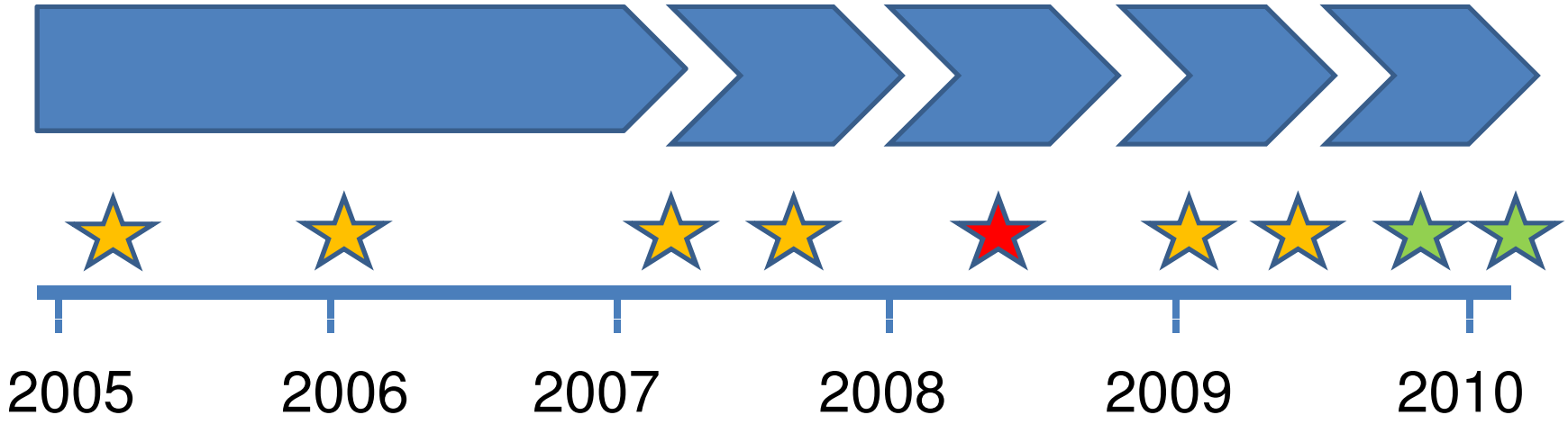
- Update
  - Scientist reading papers and entering data (PMD 1999)
  - Authors depositing information

- Curation
  - Often out of date (backlog of unprocessed papers)
  - Funding dries up ?
  - PDB ~40 % inaccurate wrt mutations*
  - What about other techniques

Intrinsic Evaluation of Text Mining Tools May Not Predict Performance on Realistic Tasks

- *J. Gregory Caporaso, Nita Deshpande, J. Lynn Fink, Philip E. Bourne, K. Bretonnel Cohen, Lawrence Hunter*; Pacific Symposium on Biocomputing 13:640-651(2008)

Mutation Mining: a history



Front ISR 2005 - Baker and Witte - Mutation Mining
JBCB 2007 - Kanagasabai etal – mSTRAP Workflow
JBCB 2007 - Witte and Baker - Systematic Evaluation Mutation Extraction
IJBRA. 2007 - Witte, Kappler, Baker – *En. semantic access* to protein eng. Lit.
ISMB 2008 - Baker etal - Towards Ontology-driven Navigation of Mutation Lit.
BMC Bioinf. 2009 - Baker and Rebholz-Schuhmann Special Issue
AMIA 2009 - ISCB Workshop - Interpretation of *Mutations* with *Semantic Supp*
DILS 2010 - Laurila etal - Algorithm for Mutation Grounding
BMC Geno 2010 - Laurila etal - Semantic Infrastructure for Mutant Impact Extraction

Single Nucleotide Polymorphism—examples

BACKGROUND AND PURPOSE: The collagen alpha2(I) gene (COL1A2) on chromosome 7q22.1, a positional and functional candidate for intracranial aneurysm (IA), was extensively screened for susceptibility in Japanese IA patients. METHODS: Twenty-one single nucleotide polymorphisms (SNPs) of COL1A2 were genotyped in genomic DNA from 260 IA patients (including 115 familial cases) (mean age, 59.9 years) and 293 controls (mean age, 61.6 years). Differences in allelic and genotypic frequencies between the patients and controls were evaluated with the chi(2) test. Circular dichroism spectrometry was monitored with collagen-related peptides that mimic triple-helical models of type I collagen with Ala-459 and Pro-459 to estimate the conformation and stability of alterations. RESULTS: Significant genotypic association in the dominant model was observed between an exonic SNP of COL1A2 and familial IA patients (chi(2)=11.08; df=1; P=0.00087; odds ratio=3.19; 95% CI, 2.22 to 6.50). This SNP induces Ala to Pro substitution at amino acid 459, located on a triple-helical domain. Circular dichroism spectra showed that the Pro-459 peptide had a higher thermal stability than the Ala-459 peptide. CONCLUSIONS: The variant of COL1A2 could be a genetic risk factor for IA patients with family history.

maps to rs number

states, locations, genes, types

PMID: 14739420, T. Yoneyama et. al: "Collagen type I alpha2 (COL1A2) is the susceptible gene for intracranial aneurysms".

Roman Klinger etal . Identifying gene specific variations in biomedical text. Journal of Bioinformatics and Computational Biology, *Special* Issue: Making Sense of Mutations requires Knowledge Management, December 2007.

Text Mining Systems for Mutations: Performance

2004 – 2009

- **MuteXt** (Protein Point Mutation) (P=87.9% R=85.8%)
  (P=49.3% R= 64.5%)
- **MEMA** (Regex DNA / Protein, HUGO) (P=75% R=98%)
- **MutationFinder** (Regex) + rules (P=98%, R=81%)

- **ProMiner** SNPExtraction and normalization / grounding (P=78%, R=67%)
- **mSTRAP** RegEx plus protein or organism name, (P=94.5% R=79.6%)
- **mSTRAP** Grounding / Normalization to db (P=91.8% R=80.9%)

- **VTag**: (CRF approach) in special context of cancer, no mapping to database
- **OSIRIS:** Query expansion: for all SNPs of a found gene: PubMed query) slow, limited to results of PubMed search engine (P=99% R=82 %)

Ongoing Challenges

- Normalization (Manual nom to dbSNP R=61%)
- Defining appropriate metrics and definitions for individual tasks for system benchmarking
- Mining Impacts / Causality from sentences
  - Protein Engineering – Impact
  - Disease Studies - Causality

<hasContent>It led to the amino acid sequence change of **H1047R,** which was found to be a **gain-of-function** mutation at the **kinase domain**. (d) FISH analysis showing DNA copy number gains and amplifications of **PIK3CA** locus on the primary tumor cases T10 and T24. </hasContent></Sentence>

- Cottage Industry ? – Scale up to publishing of mutation annotations according to Semantic Metadata for incorporation into systems level approaches and prediction tools e.g.
  - pathway analyses based on SNP annotations
  - impact prediction for un annotated residues
  - reuse scenarios

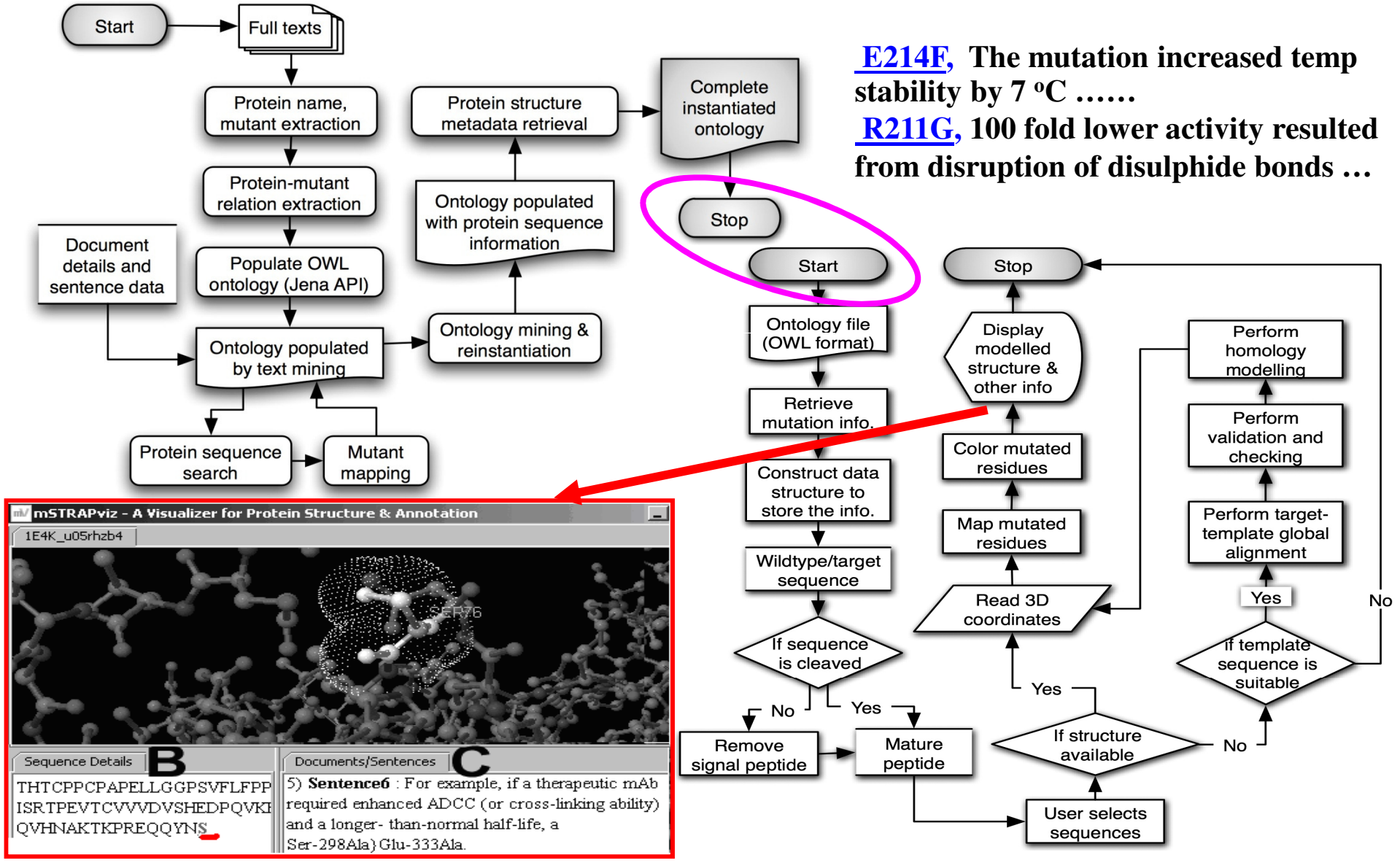
Mutation Reuse Scenarios

- Annotation
  - Cancer Genome (Forbes etal 2010)
  - Pathway (Bauer Mehren etal 2009)
  - **Pubmed Abstracts (Laurila etal 2010)**
  - **Protein Structure (Kanagasabai etal 2007)**
- Impact Prediction
  - SNAP (Bromberg and Rost 2008)
  - Membrane protein stability (Winnenburg etal 2009)
- Federated Query over Mutation Triplestores
  - **SADI Semantic Web Services (Riazanov etal 2010)**

mSTRAP

A System for Mutation Extraction and Structure Annotation Visualization

Home | Download | Guide | Contact

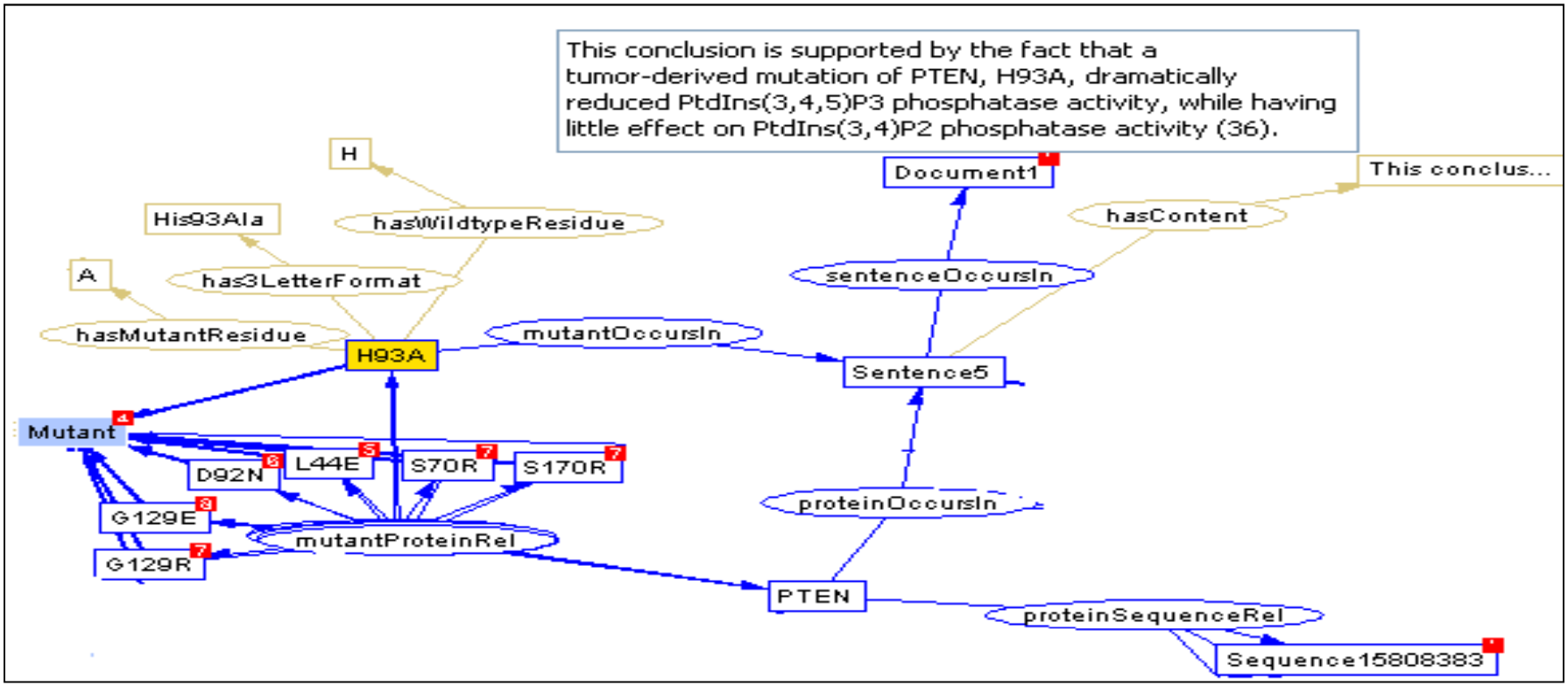


E214F, The mutation increased temp stability by 7 °C ……

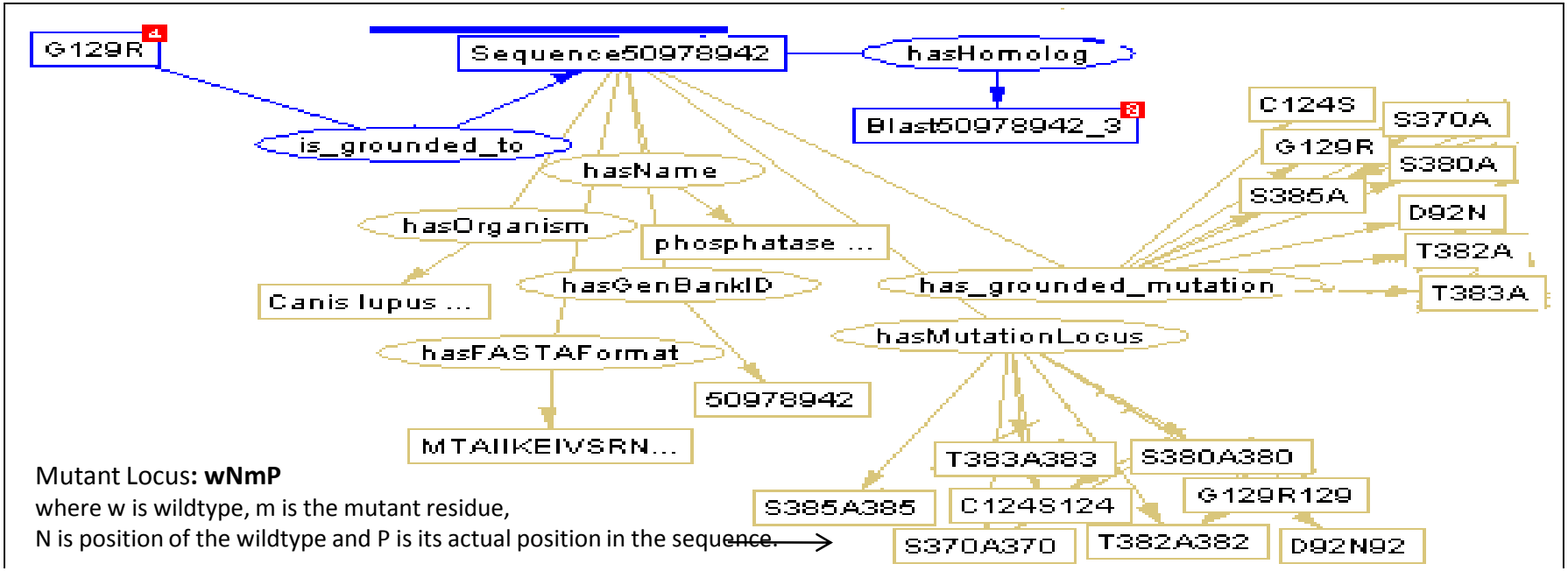
R211G, 100 fold lower activity resulted from disruption of disulphide bonds …

- COSMIC database content used for building a gold standard corpus for evaluation of mutation extraction and grounding
- Target families:
  1. PIK3CA - 42 papers out of 78 online
  2. FGFR3 - 37 papers out of 59 online
  3. MEN1 - 19 papers out of 72 online
- Protein–Mutation Tuple (extraction, normalisation)
- Mutation Grounding to sequences
- Sequences then checked by pair wise sequence alignment with gold standard protein sequence with >99% homology

A



B



Mutant Locus: **wNmP**
where w is wildtype, m is the mutant residue,
N is position of the wildtype and P is its actual position in the sequence.

mSTRAP

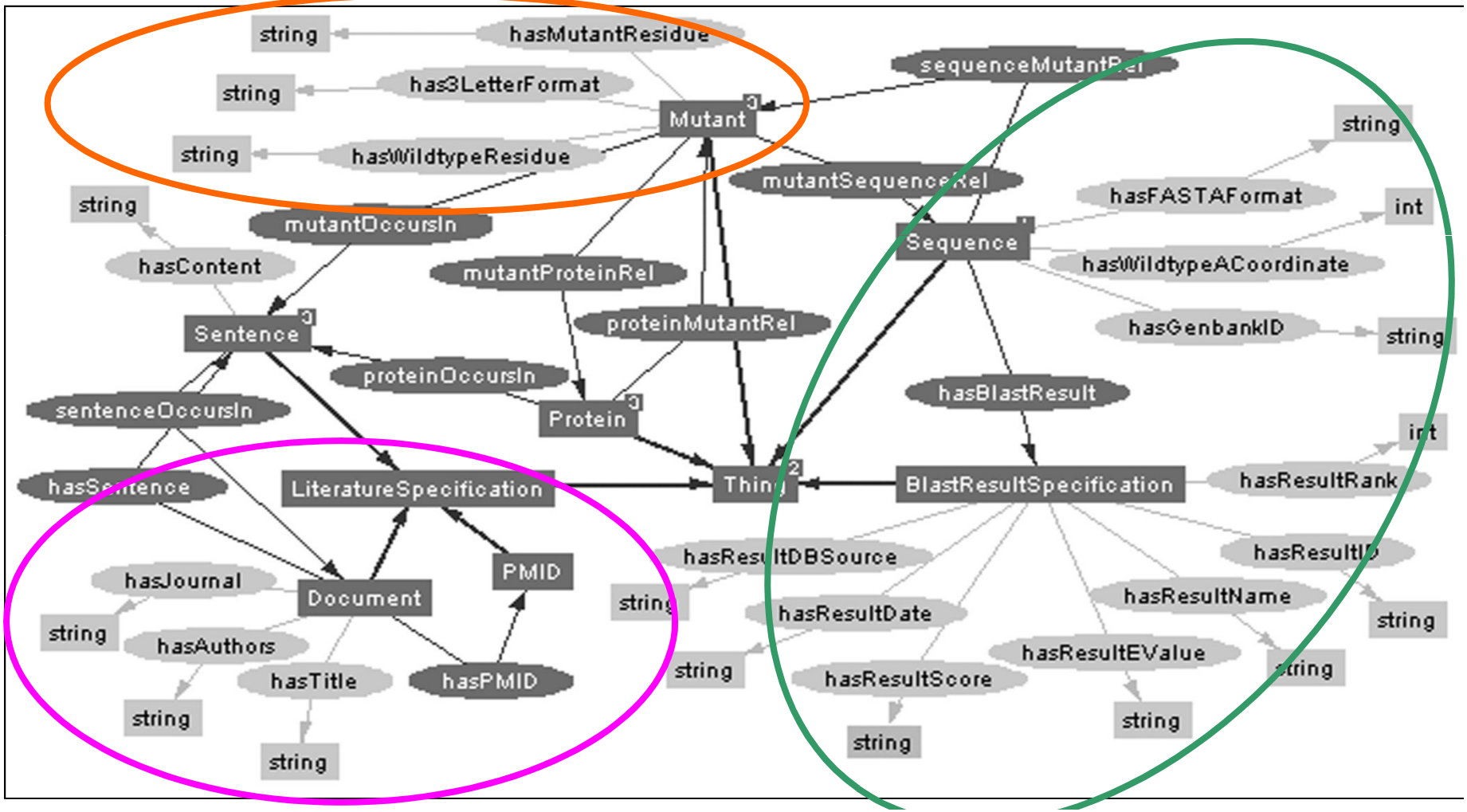
A System for Mutation Extraction and Structure Annotation Visualization

Home | Download | Ontologies | Guide | References | Contact

Sample Ontologies

Listed below are some sample ontologies that we have created:

- Phosphotase Mutation
- BRCA1/BRCA2 Mutation

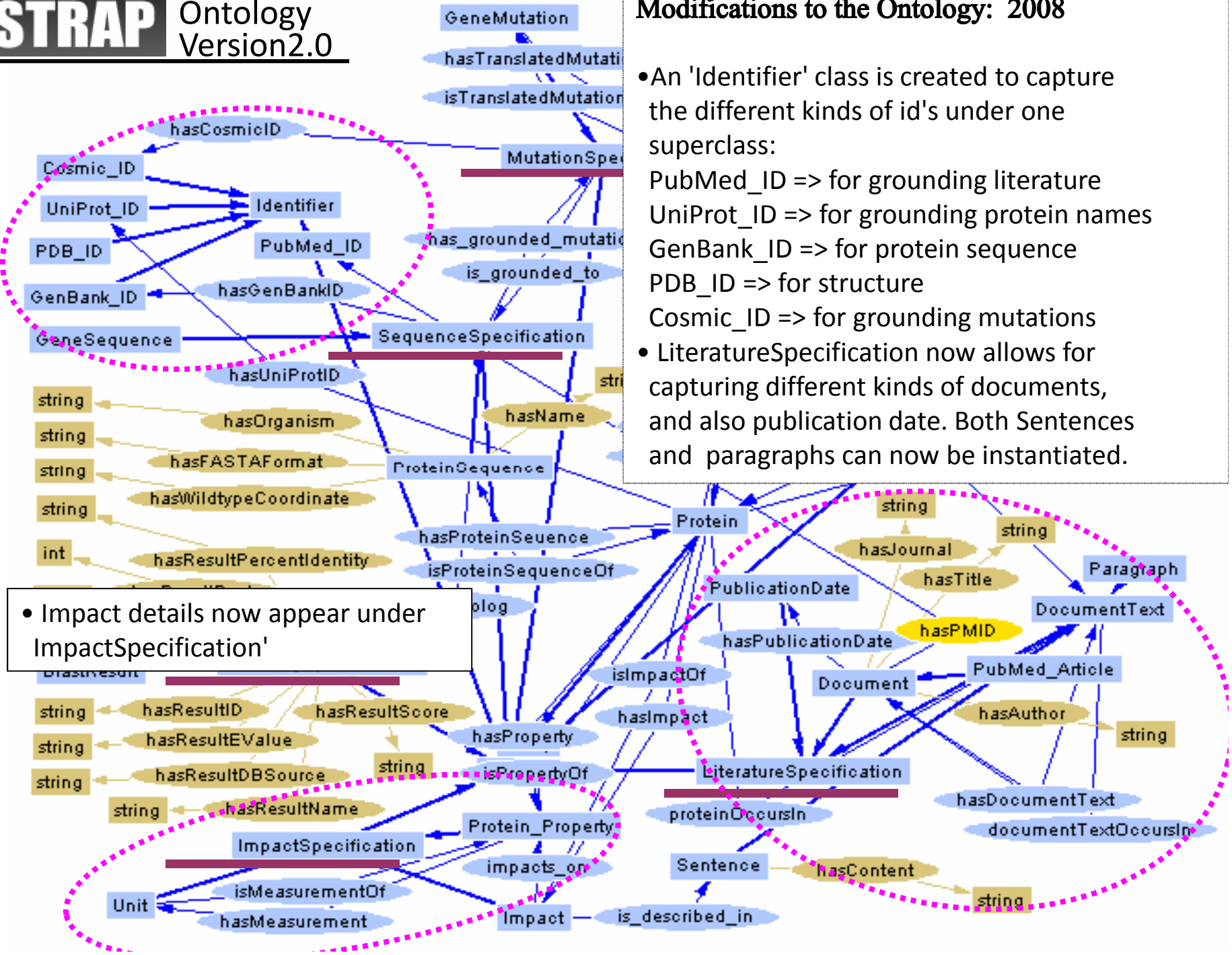


mSTRAP Ontology Version2.0

Modifications to the Ontology: 2008

- An 'Identifier' class is created to capture the different kinds of id's under one superclass:
  PubMed_ID => for grounding literature
  UniProt_ID => for grounding protein names
  GenBank_ID => for protein sequence
  PDB_ID => for structure
  Cosmic_ID => for grounding mutations
- LiteratureSpecification now allows for capturing different kinds of documents, and also publication date. Both Sentences and paragraphs can now be instantiated.

- Impact details now appear under ImpactSpecification'



Toward a Richer Representation of Sequence Variation In the Sequence Ontology (2010)

Michael Bada and Karen Eilbeck

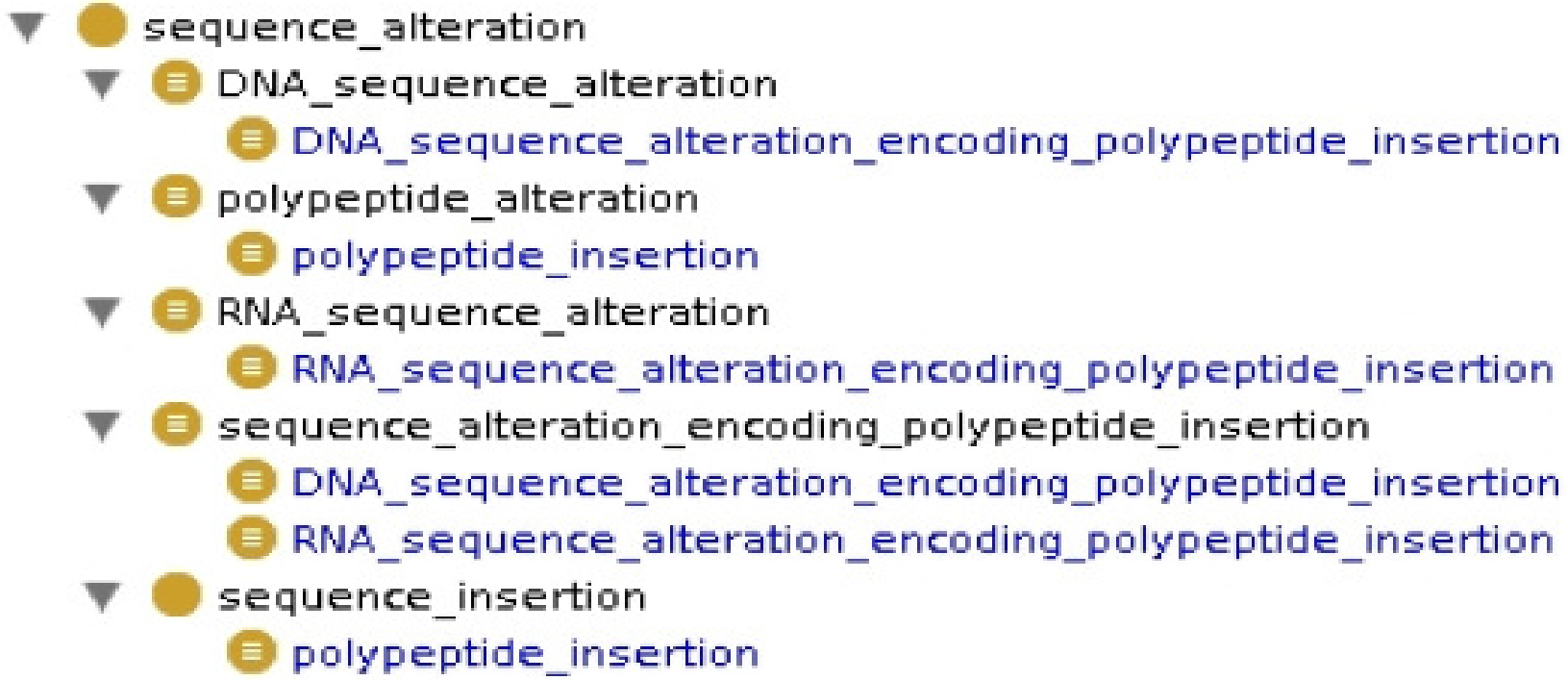
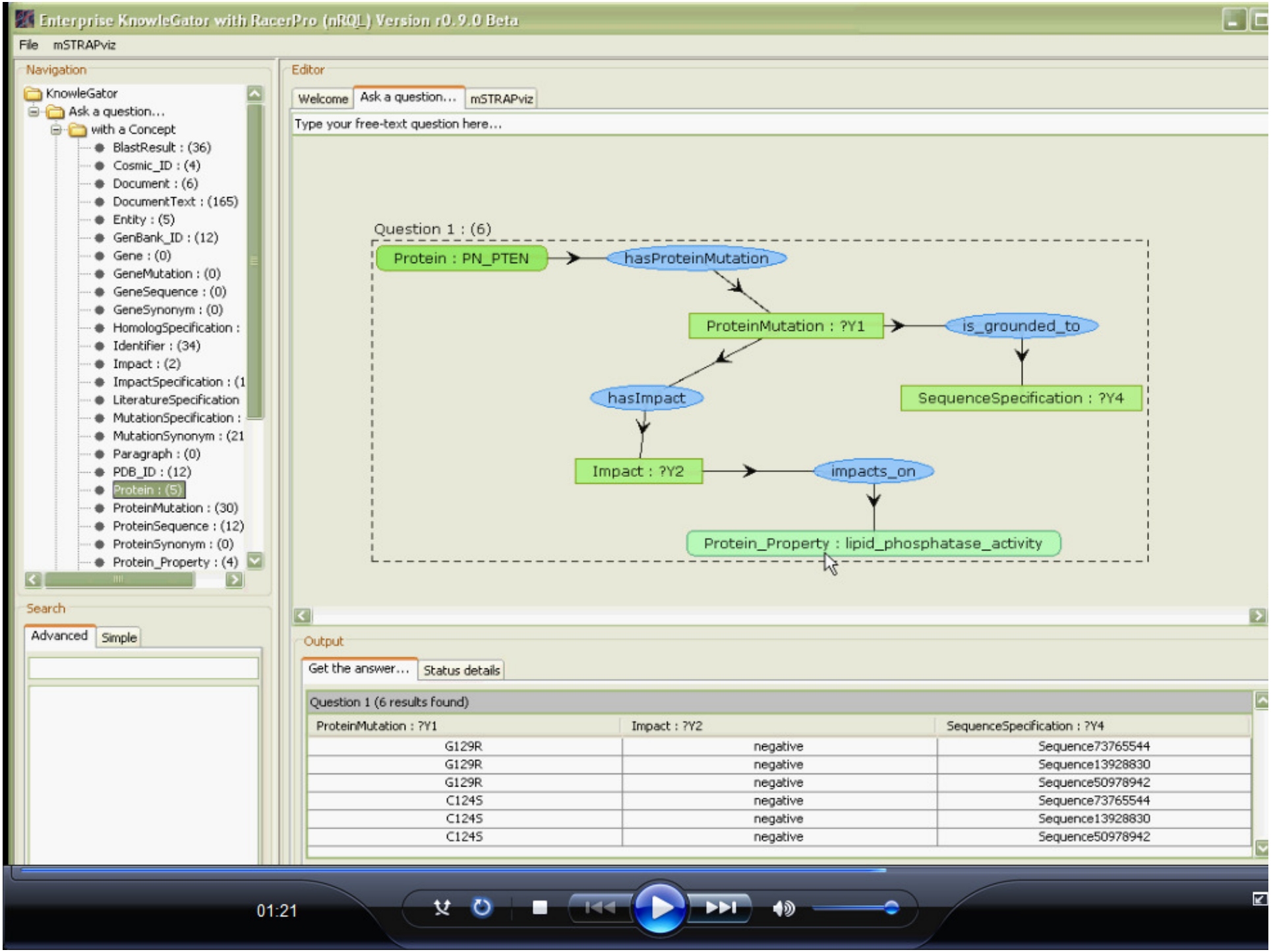
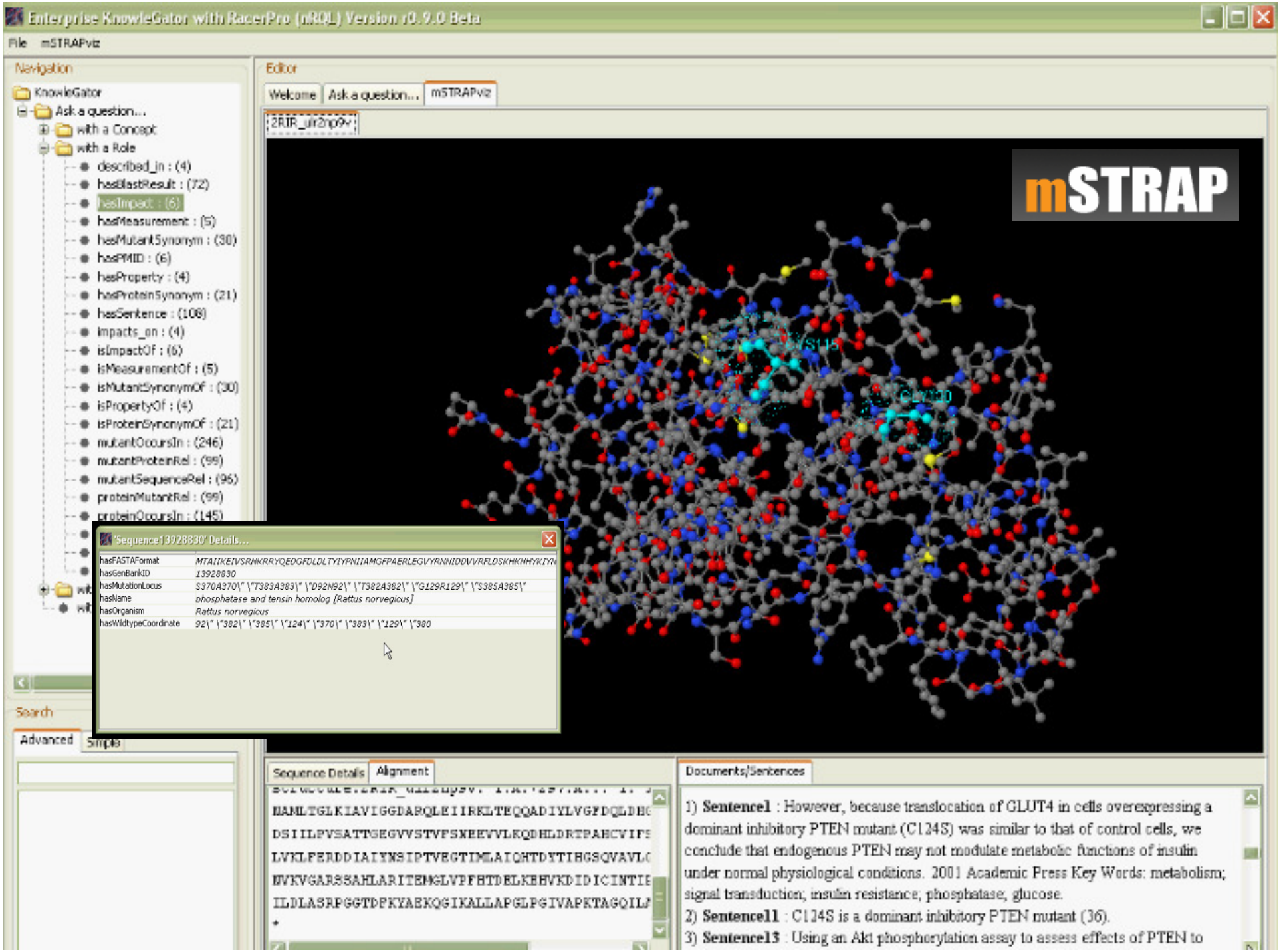


Figure 5. Automatically classified hierarchy of sequence alterations.

Proceedings of the Workshop on Annotation, Interpretation and Management of Mutations (AIMM-2010), Annotation, Interpretation and Management of Mutations A workshop at ECCB10. Ghent, Belgium, September 26th, 2010.





mSTRAP

A System for Mutation Extraction and Structure Annotation Visualization

News

*Kanagasabai R., Choo K. H., Ranganathan S. and *Baker CJO, Workflow for Mutation Extraction and Structure Annotation, Journal of Bioinformatics and Computational Biology, **(2007)**, vol 5, 6:1319-1337*

- mSTRAPviz™ is now integrated into Knowlegator™ (pic1, pic2), an ontology-driven knowledge integration and navigation tool. This system will be presented at Toronto during the ISMB2008 Highlights Track

```
ISMB 2008 Highlights Track: HL17
Towards Ontology-driven Navigation of the Lipid Bibliosphere and Mutation Literature
Monday, July 21 - 10:45 a.m. - 11:10 a.m.
Room: 718A
Presented by: Christopher J. O. Baker, Institute for Infocomm Research, SG
Session Chair: Andrey Rzhetsky
```

DEMO clips of the Knowlegator™+mSTRAPviz™ system:

Clip #1 : Query & Navigation in Knowlegator (25MB)

Clip #2 : Results from query fed into mSTRAPviz for automated homology modeling (17MB)

- Release of v1.1 (9Jul08) - skip repeated modeling of same sequences when reselected

Mutation Impact Extraction

Laurila *et al. BMC Genomics* 2010, **11**(Suppl 4):S24
http://www.biomedcentral.com/1471-2164/11/S4/S24

BMC Genomics

PROCEEDINGS Open Access

Algorithms and semantic infrastructure for mutation impact extraction and grounding

Jonas B Laurila[1], Nona Naderi[2], René Witte[2], Alexandre Riazanov[1], Alexandre Kouznetsov[1], Christopher JO Baker[1*]

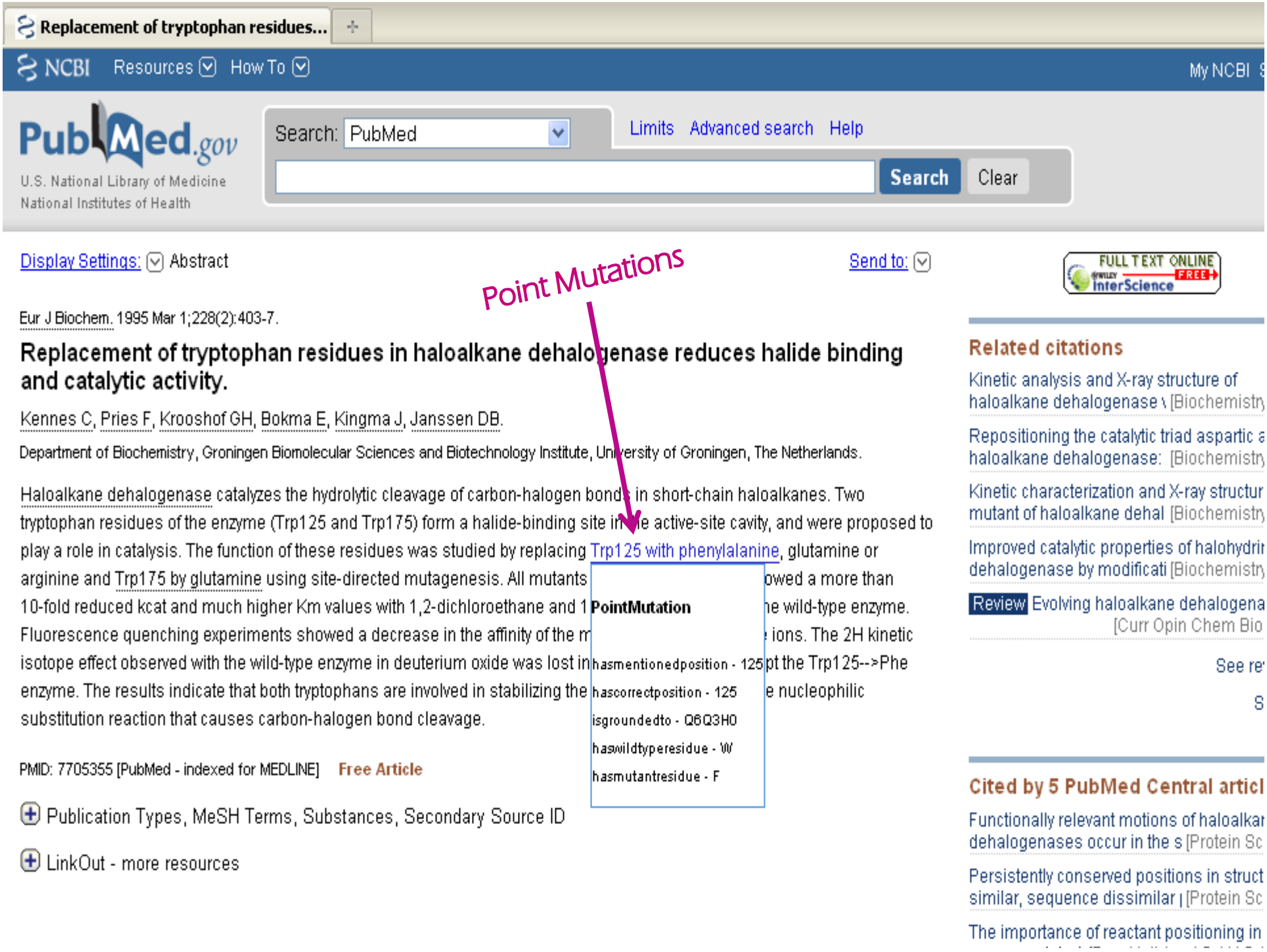
From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics (InCoB2010)
Tokyo, Japan. 26-28 September 2010

A semantic assistant for mutation mentions in PubMed abstracts.

PubMed.gov

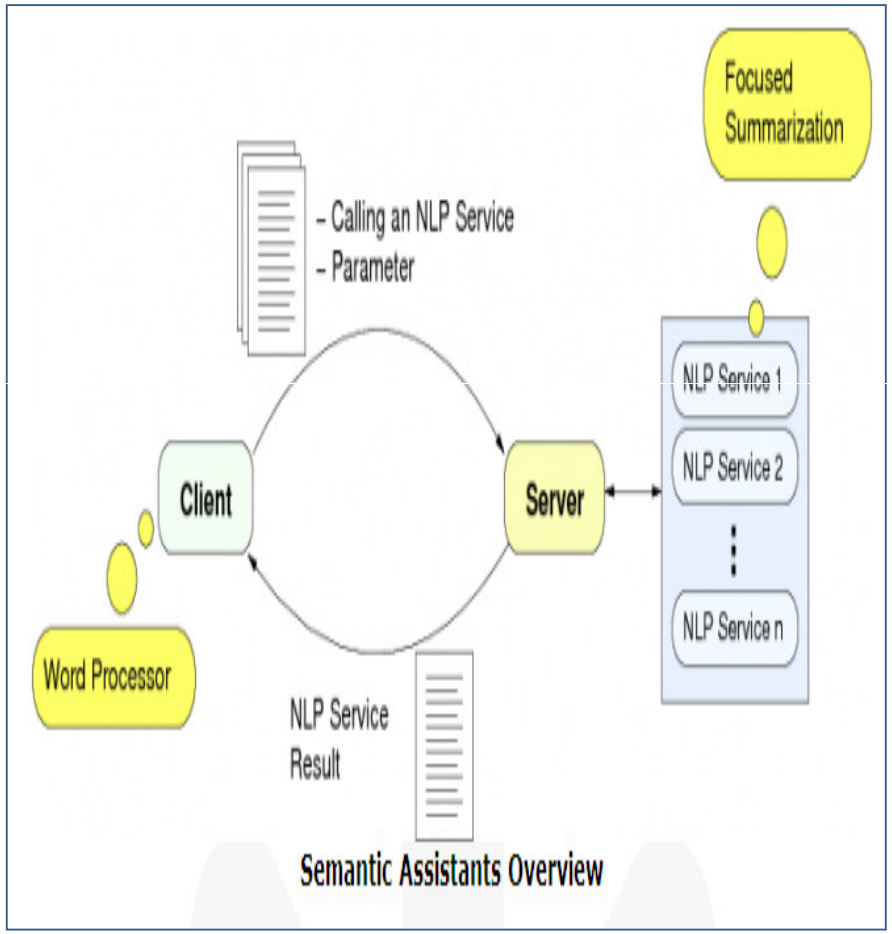
Jonas B Laurila[1], Alexandre Kouznetsov[1] and Christopher J O Baker*[1]

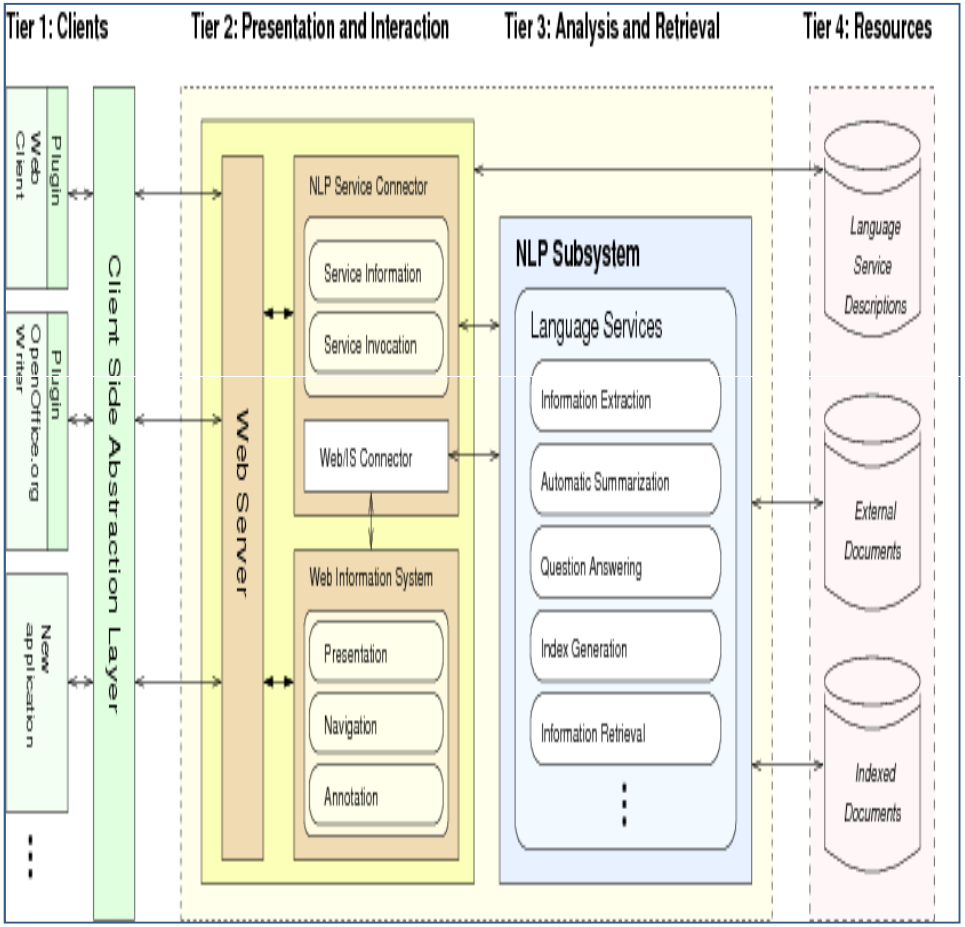
AIMM-2010 @ ECCB 2010
Annotation, Interpretation and Management of Mutations 2010 **Ghent, Belgium.**



http://www.ncbi.nlm.nih.gov/pubmed/10508409

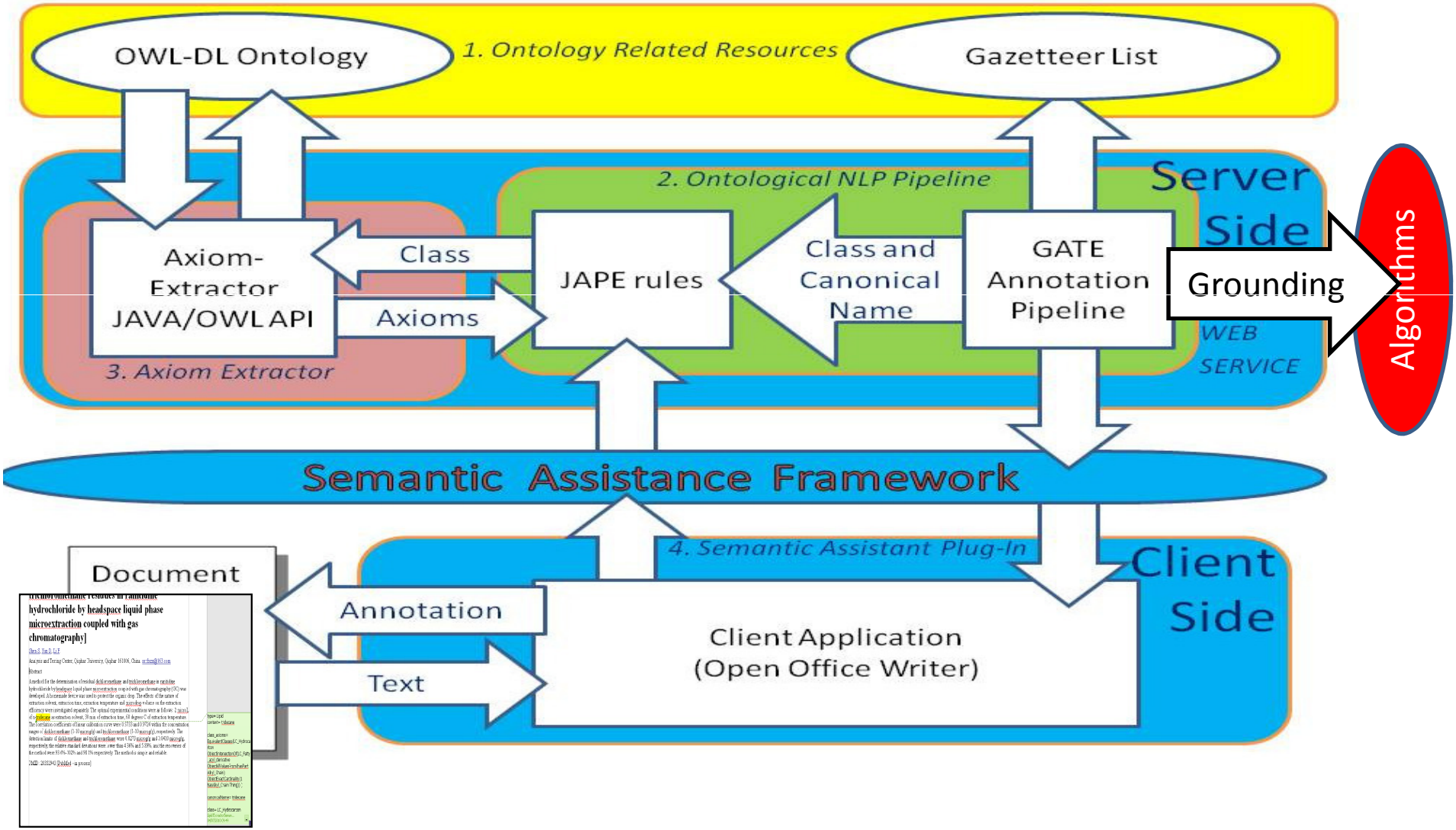
Semantic Assistant Framework





René Witte and Thomas Gitzinger.
Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients.
3rd Asian Semantic Web Conference (ASWC 2008), February 2–5, 2009, Bangkog, Thailand.
Springer LNCS 5367, pp. 360–374. *(Acceptance rate: 31%)*

Semantic Assistant: Mutations



Mutation Grounding

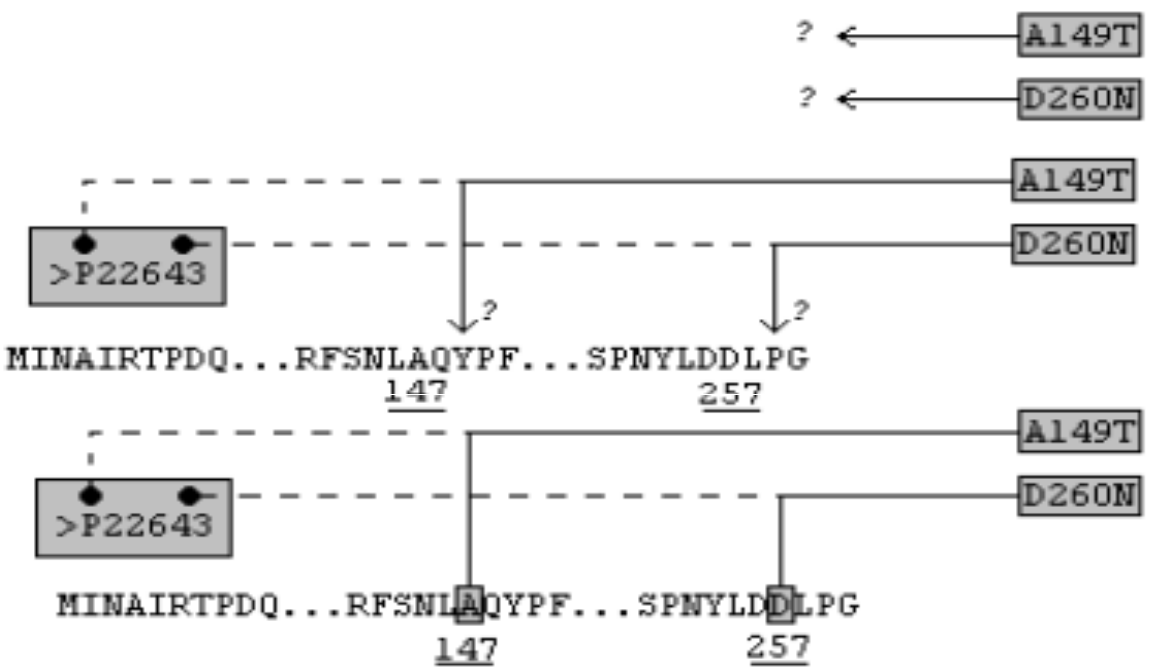
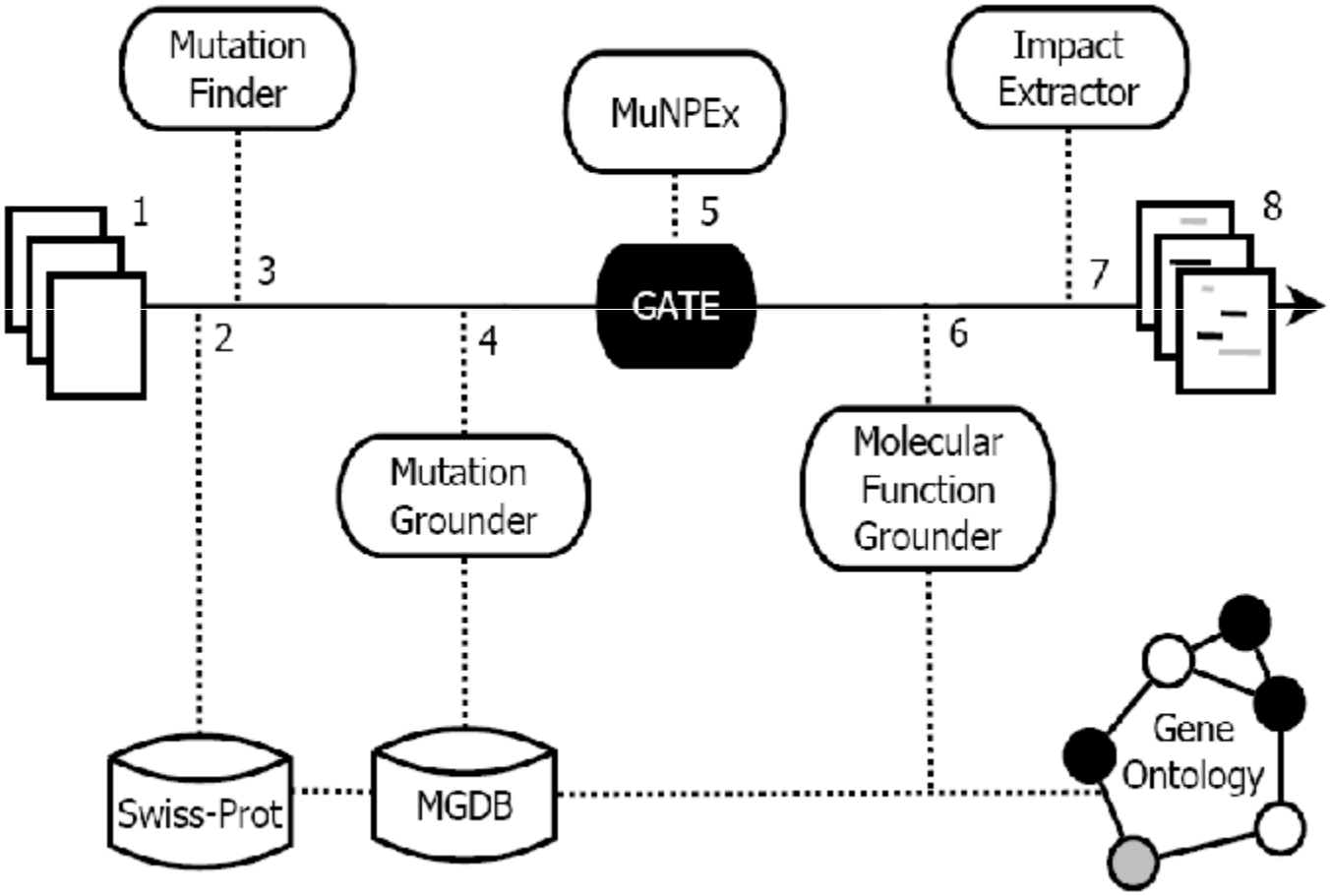


Fig. 1. Degrees of Mutation Grounding. *Uppermost*, mutation mentions are extracted but their relation to the appropriate protein sequence is not (no grounding). *Middle*, the related protein is found and mappings are established to the sequence if both text and database follow the same numbering scheme. *Lowermost*, the mutations are properly grounded, i.e. mapped to the correct position on the amino acid sequence of the related protein. Systems performing and not performing mutation grounding are displayed to the right.

Implementation
A GATE pipeline



Sample sentences

“As expected, complete loss in activity of **W109L** and sustained activity of **F151W** were observed.”

”In order to further understand the catalytic mechanism we constructed an **Asp-124->Asn** mutant enzyme.”

“DhlA shows only a small decrease in activity when **Trp-125 is replaced with phenylalanine**.”

“The **W125F** mutant showed only a slight reduction of activity (Vmax) and a larger increase of Km with 1,2-dibromoethane."

"**Haloalkane dehalogenase** (**DhlA**) from **Xanthobacter autotrophicus** GJI0 hydrolyses terminally chlorinated and brominated n-alkanes to the corresponding alcohols."

Mutation **Description** **Protein name** **Gene name** **Organism name**

Sample sentences

“As expected, complete **loss** in **activity** of W109L and **sustained activity** of F151W were observed.”

”In order to further understand the catalytic mechanism we constructed an Asp-124->Asn mutant enzyme.”

“DhlA shows only a small **decrease** in **activity** when Trp-125 is replaced with phenylalanine.”

“The W125F mutant showed only a slight **reduction** of **activity** (Vmax) and a larger **increase** of **Km** with 1,2-dibromoethane."

“Haloalkane dehalogenase (DhlA) from Xanthobacter autotrophicus GJI0 hydrolyses terminally chlorinated and brominated n-alkanes to the corresponding alcohols."

Direction **Protein Property**

Named entities of interest

- Mutations
- Proteins
- Genes
- Organisms
- Protein properties
  - Protein functions *(activity, binding etc)*
  - Kinetic variables *(Km, kcat etc)*
  - (stability)

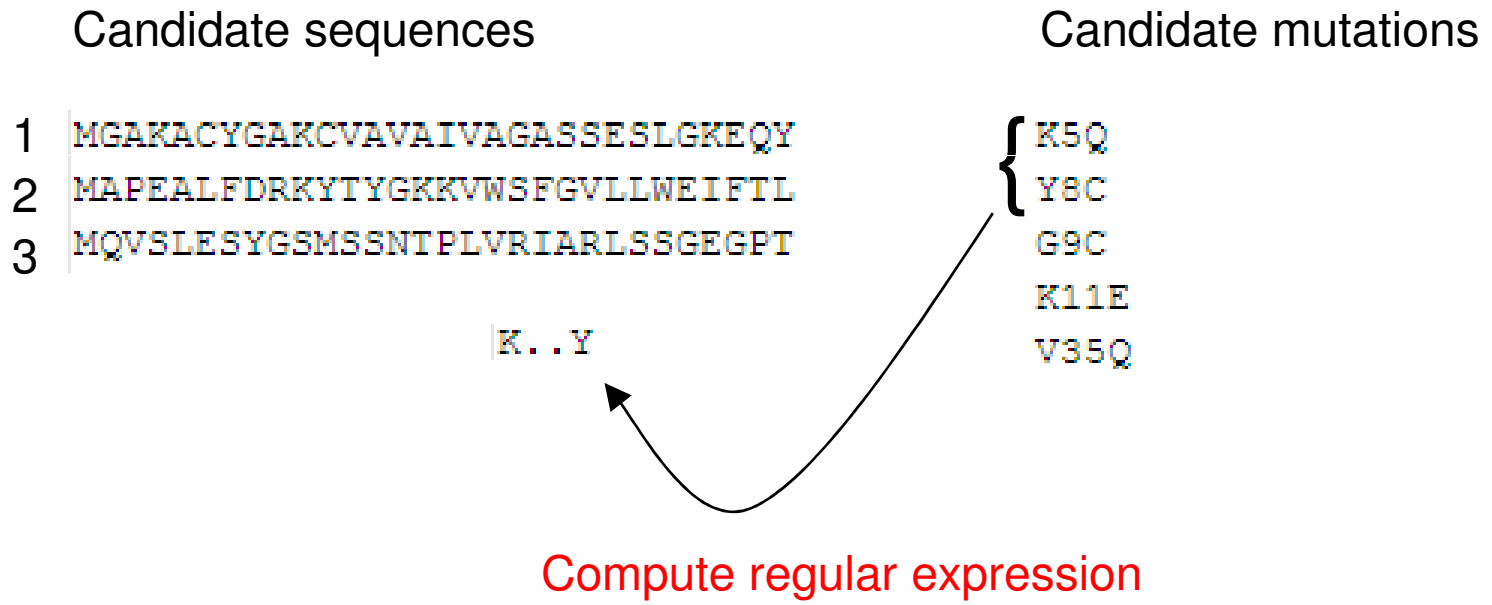
Mutation Grounding

1. Retrieve and normalize mutations
2. For each candidate sequence
    1. For each pair of mutations
        1. Make regexp $w_1.(N_2\text{-}N_1)w_2$
        2. Match regexp to sequence
        3. Check remaining residues at corrected positions.
3. Ground proteins and mutations to the same AC / sequence

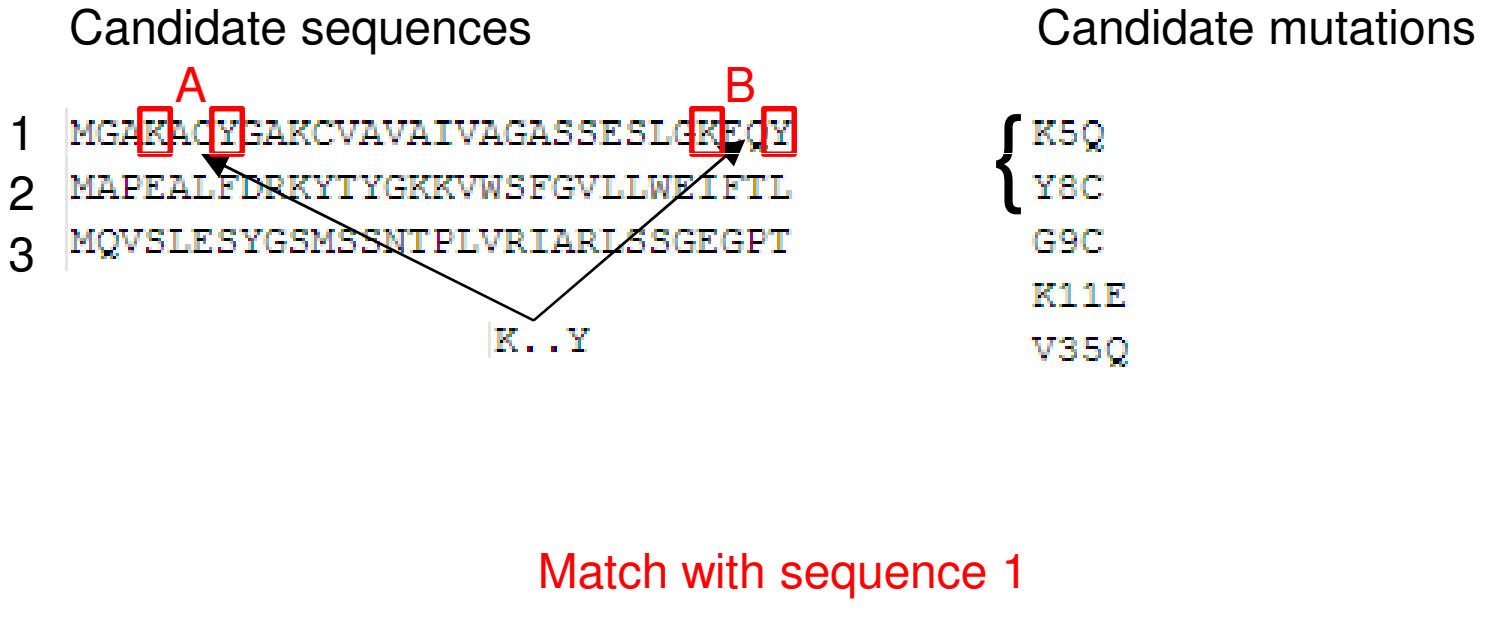
Mutation Grounding: Example

	Candidate sequences	Candidate mutations
1	MGAKACYGAKCVAVAIVAGASSESLGKEQY	K5Q
2	MAPEALFDRKYTYGKKVWSFGVLLWEIFTL	Y8C
3	MQVSLESYGSMSSNTPLVRIARLSSGEGPT	G9C
		K11E
		V35Q

Mutation Grounding: Example

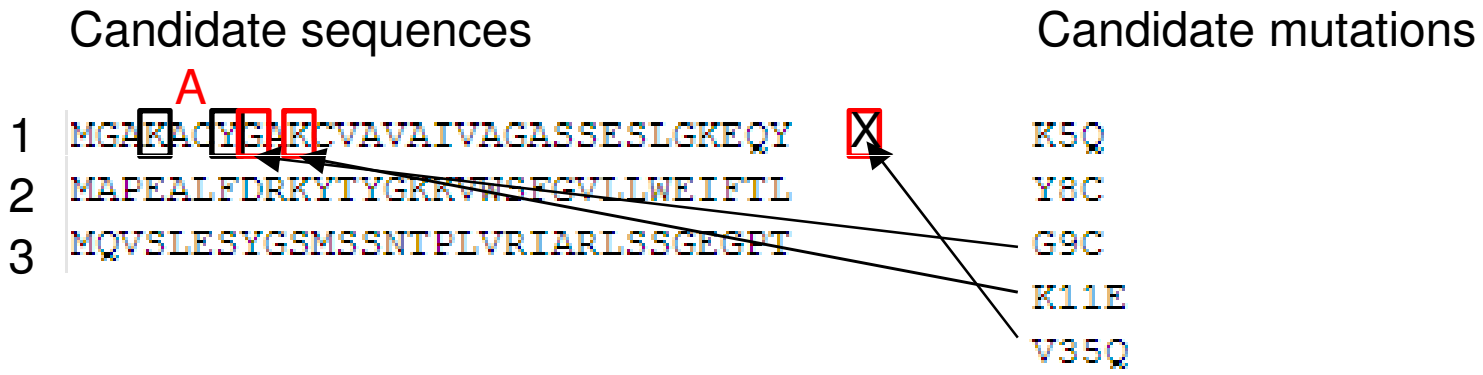


Mutation Grounding: Example



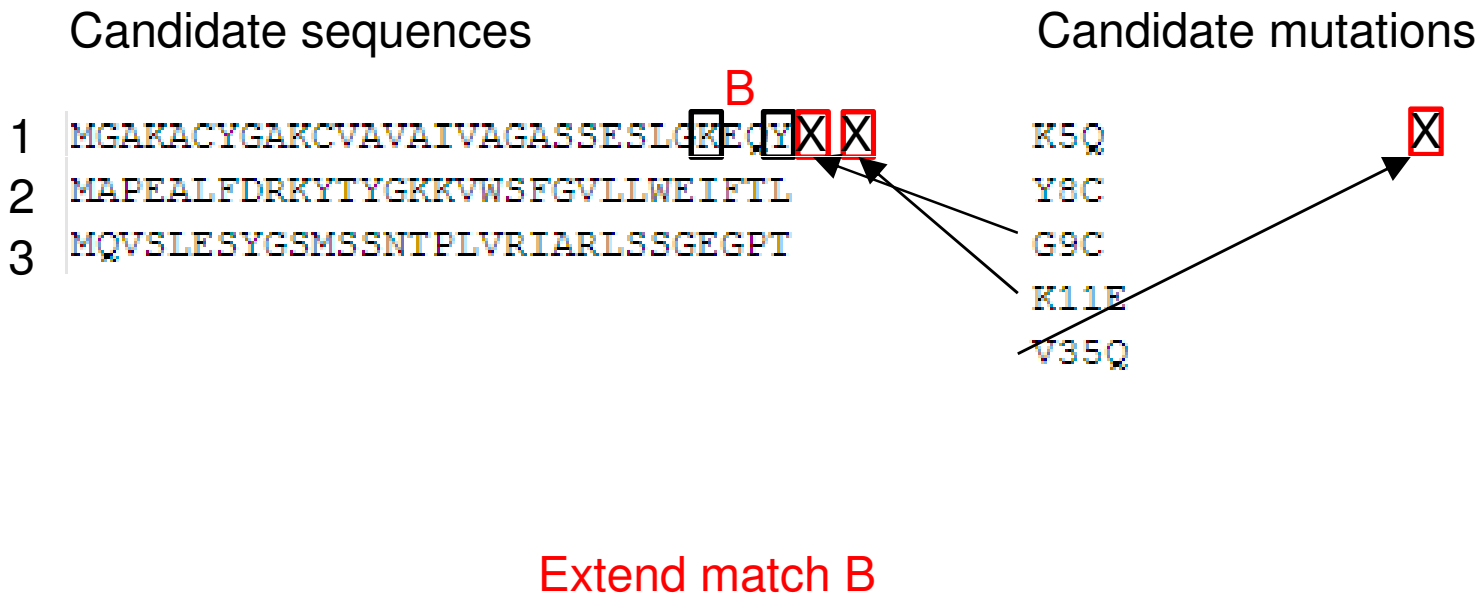
Match with sequence 1

Mutation Grounding: Example



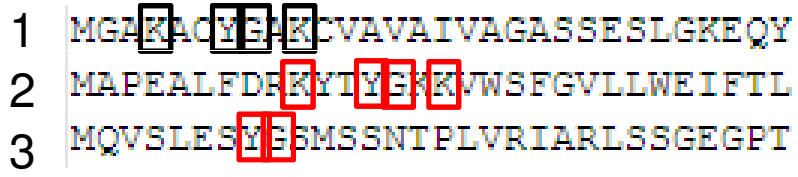
Extend match A

Mutation Grounding: Example



Mutation Grounding: Example

Candidate sequences

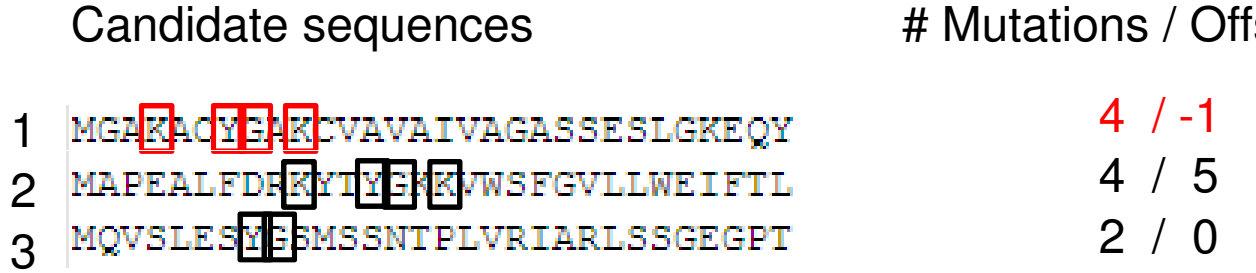


Candidate mutations

K5Q
Y8C
G9C
K11E
V35Q

Match with sequence 2 & 3

Mutation Grounding: Example



Choose best candidate sequence:

1. Most grounded mutations
2. Least absolute offset

Evaluation

- COSMIC
  - Catalogue Of Somatic Mutations In Cancer
  - PIK3CA, FGFR3, MEN1
  - 63 documents
- Haloalkane dehalogenases
  - Protein engineering literature
  - 13 documents

Evaluation

Corpus	Precision	Recall	Corpus size
PIK3CA	0.86	0.70	30
FGFR3	0.89	0.66	26
MEN1	0.54	0.32	7
Haloalkane Dehalogenases	0.83	0.73	13
Average	0.84	0.65	76

Grounding of named entities

- Protein grounding
  - Assign the correct UniProt id to each detected protein entity.
- Mutation grounding
  - Verify and, if necessary, correct each mutation location to match its corresponding protein's sequence as obtained from UniProt.
- Protein function grounding
  - Assign the correct gene ontology id to detected protein functions

Protein & mutation grounding

- Combined into one method

1. A pool of accession numbers is created based on occurence of protein and gene names
2. Mutations are matched to candidate sequences, going from min to max amount of mutations.
3. Sequence with most grounded mutations is considered correct for the entire paper

Protein function grounding

1. Retrieve go:mf concepts related to previously grounded proteins
2. ground noun phrases, with *activity, binding, affinity* or *specificity* as head nouns, to retrieved go:mf concepts.
3. score them according to lexical similarity with the retrieved go:mf concepts.
4. use scores to solve contradictions in output protein function grounding and impact information

Protein function grounding

1. Retrieve go:mf concepts related to previously grounded proteins
2. ground noun phrases, with *activity, binding, affinity* or *specificity* as head nouns, to retrieved go:mf concepts.
3. score them according to lexical similarity with the retrieved go:mf concepts.
4. use scores to solve contradictions in output protein function grounding and impact information

Protein function grounding Example

Grounded protein

(P22643) Haloalkane dehalogenase

Related gene ontology molecular function concepts

(GO:0018786) haloalkane dehalogenase activity

Found noun phrases

anhydrase activity

dehalogenase activity

activity **First stem!**

Protein function grounding Example

compare – $similarity = \frac{|N \cap G|^2}{|N|\,|G|}$

haloalk dehalogen activ

anhydr activ	s = 1 / (2*3) = 1/6
dehalogen activ	s = 4 / (2*3) = 4/6
activ	s = 1 / (1*3) = 2/6

Protein function grounding Example:

ground *to a certain degree*

(GO:0018786) haloalk dehalogen activ

anhydr activ	→((GO:0018786) , 0.17)	(3)
dehalogen activ	→((GO:0018786) , 0.67)	(1)
activ	→((GO:0018786) , 0.33)	(2)

Impact direction term lists

Positive	Negative	(cont.)	Neutral	Negation	Non-Neutral
increase	abolish	loose	identical	without	affect
-increases	decrease	defect	similar	no	effect
-increased	reduce	disrupt	full	not	alter
-increasing	lower	diminish			differ
enhance	inhibit				
higher	impair				
improve					

Relation detection

1. Impacts
    1. Directionality+Protein property
    2. «**sustained activity** of F151W were observed»
2. Mutants
    1. Set of grounded mutations
3. Mutant+Impact

- Relations found in text by the use of rules

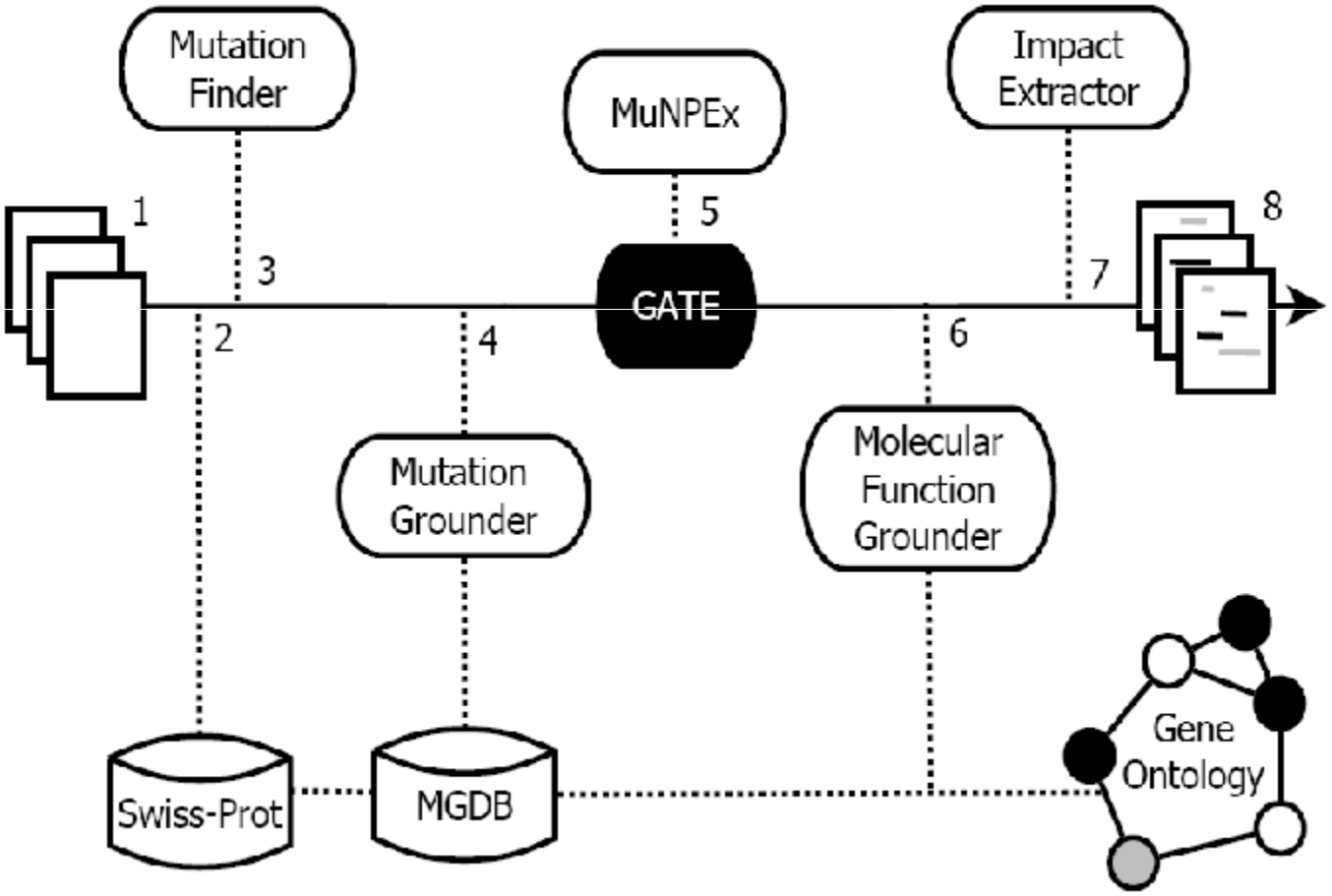
Impact rules - an example

#negative-impact-rule
*If (***Sentence** *contains* **Protein Function** *and*
Sentence *contains* **Negative Direction** *and*
Sentence *not contains* **Positive Direction**)
→
Markup **Sentence** as **Negative Impact** on **Protein Function**

DhlA shows only a small **decrease** in **activity** when Trp-125 is replaced with phenylalanine.

A larger **increase** of **Km** with 1,2-dibromoethane.

Implementation: a GATE pipeline



Access to mutation information

1. Export mutation impact information from text to an RDF triple store
2. Provide a SPARQL endpoint as a query interface to the triple store
3. Make both the pipeline and triple store available through semantic web services (SADI)

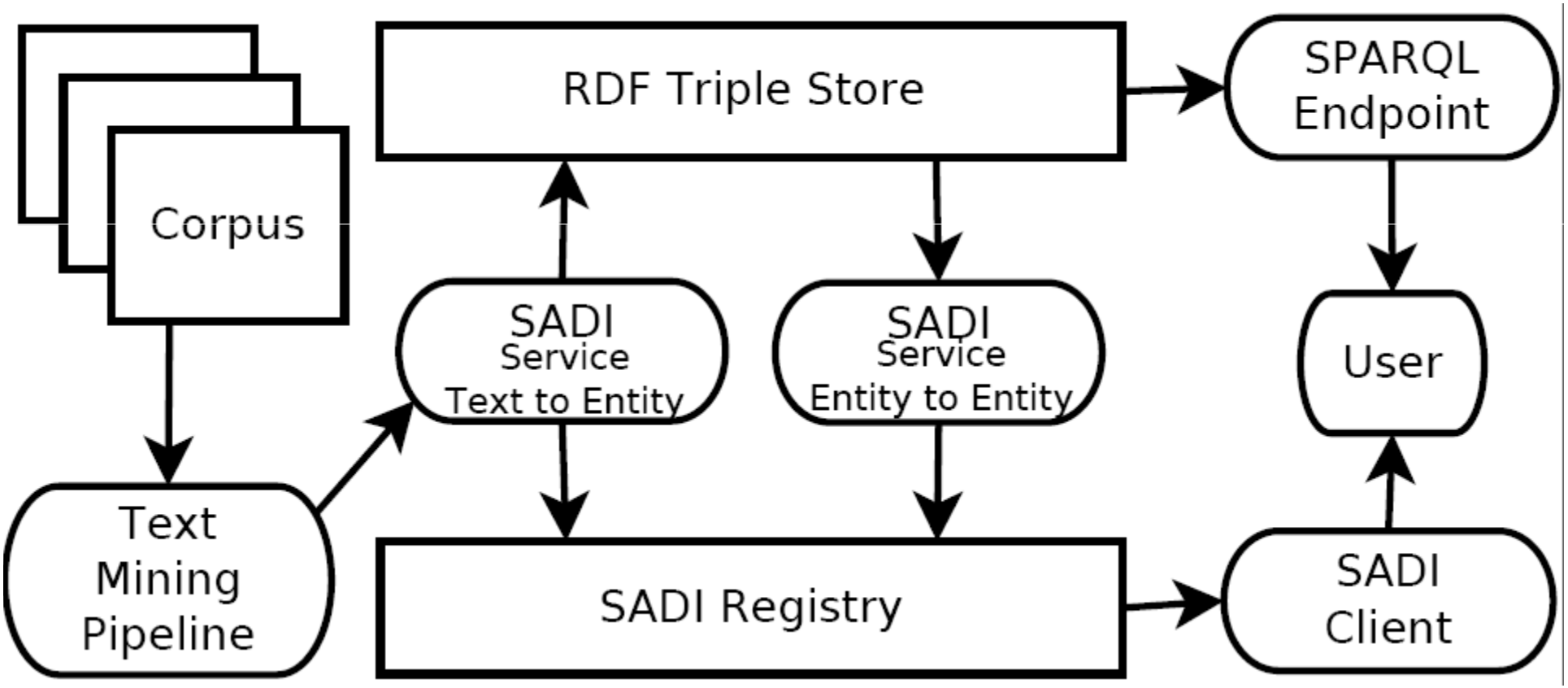
http://sadiframework.org/

Example queries

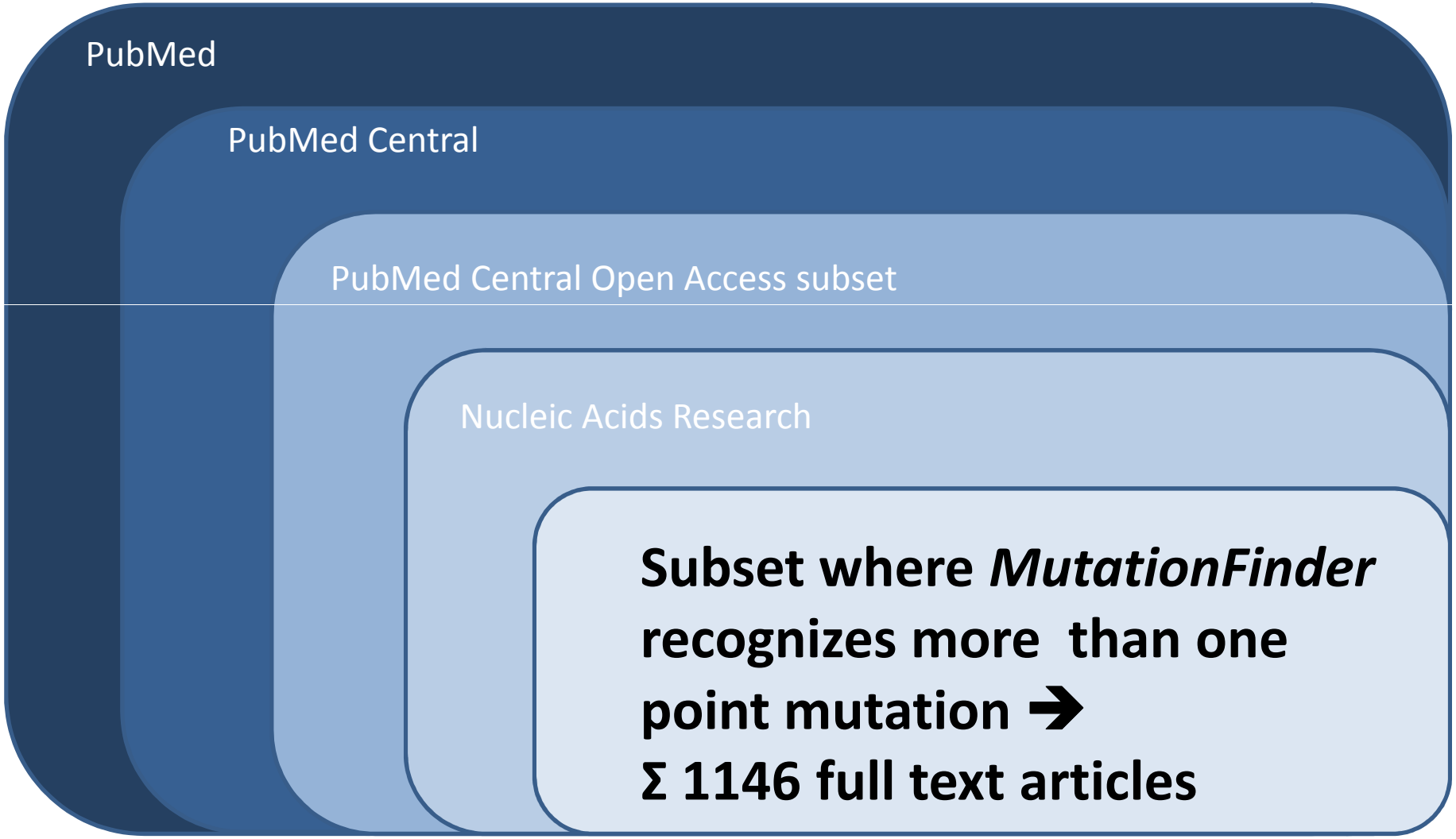
- Retrieve all reported mutations and their impacts on haloalkane dehalogenase.
- Find all mutants with an increased Ca^{2+} affinity.
- Has the F137T mutation of carbonic anhydrase II previously been studied?

http://sadiframework.org/

Access to mutation information - summary



The NAR Subset



Statistics of NAR subset

- Mutations could be grounded in 733 of the 1146 relevant documents by our system.

  – A total of 4008 *GPM-occurs-in-Document* triples
  – 2977 *unique/distinct GPMs*
  – 759 *Protein-occurs-in-Document* triples
  – 995 *Impacts* (191 positive, 120 neutral, 684 negative)
  – 110 *unique/distinct* Gene Ontology Terms

  …were extracted from these documents.

Impact Extraction for Reuse in SNP Prediction

Predicted

Extracted

#	Mut	$SNAP_{MAT}$ Pred	Reference	Func Change/ Impact	GO ID	Agreement	GO Decrip
	R165W		34	**positive**	GO_0004016		adenylate cyclase activity
17	F51L	Non-neutral	16	negative	GO_0004016	Y	adenylate cyclase activity
36	I102T	Neutral	16	negative	GO_0004016	N	adenylate cyclase activity
70	A154D	Non-neutral	16	negative	GO_0004016	Y	adenylate cyclase activity
	S127L		43	**positive**	GO_0004144		diacylglycerol O-acyltransferase activity
	T11S		13	**negative**	GO_0004144		diacylglycerol O-acyltransferase activity
	R18C		13	**negative**	GO_0004144		diacylglycerol O-acyltransferase activity
	F121R		19	**negative**	GO_0004977		melanocortin receptor activity
44	N123A	Non-neutral	38	negative	GO_0004978	Y	adrenocorticotropin receptor activity
45	N123D	Non-neutral	38	negative	GO_0004978	Y	adrenocorticotropin receptor activity
	D126A		38	neutral	GO_0004978		adrenocorticotropin receptor activity
156	I316T	Non-neutral	33	negative	GO_0005000	Y	Vasopressin receptor activity
15	E49A	Non-neutral	29	negative	GO_0005184	Y	neuropeptide hormone activity
26	D90A	Non-neutral	29	negative	GO_0005184	Y	neuropeptide hormone activity
	E100A		29	**negative**	GO_0005184		neuropeptide hormone activity
	E100C		1	**negative**	GO_0005515		protein binding

In silico mutagenesis: a case study of the melanocortin 4 receptor.
Bromberg Y, Overton J, Vaisse C, Leibel RL, Rost B – FASEB 2009

Impact Extraction: Deployed as Semantic Web Service

SADI, SHARE, and the *in silico Scientific Method* - Mark D Wilkinson1§*, Luke McCarthy1*, Benjamin Vandervalk1*, David Withers1*, Edward Kawas1*, Soroush Samadian1* BMC Bioinformtics (2010)

Deploying the Mutation Impact mining pipeline with SADI: an exploratory case study, Alexandre Riazanov, Jonas Bergman Laurila and Christopher J O Baker

Proceedings of the Workshop on Annotation, Interpretation and Management of Mutations (AIMM-2010)
Annotation, Interpretation and Management of Mutations. A workshop at ECCB10.
Ghent, Belgium, September 26th, 2010.

AIMM2010

Annotation, Interpretation and Management of Mutations.
A workshop at ECCB10.

HOME

SCHEDULE

WORKSHOP THEMES

KEYNOTE SPEAKERS

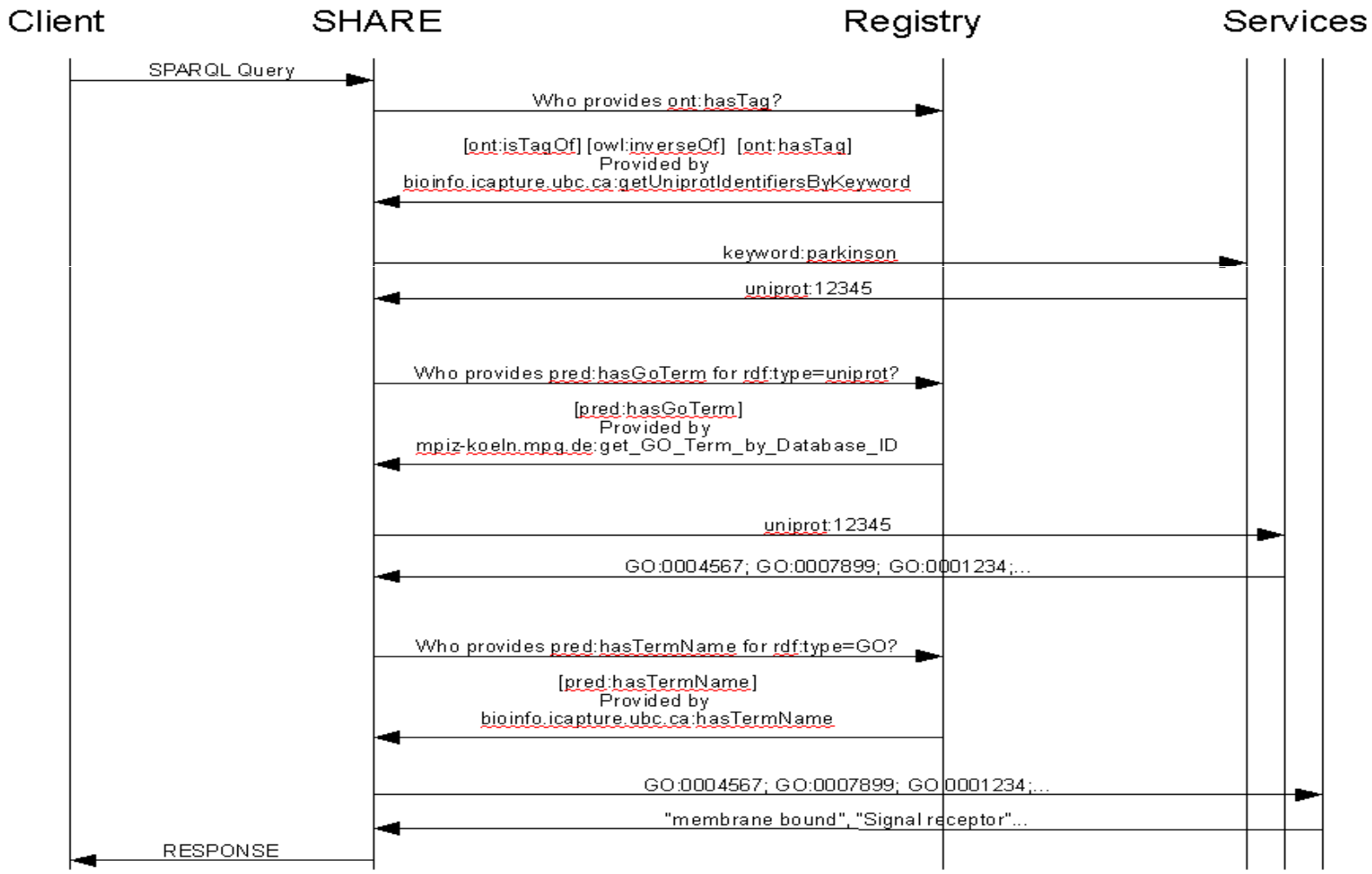
AIMM publications

2008 CEUR: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-429/

2008 BMC special issue: http://www.biomedcentral.com/bmcbioinformatics/10?issue=S8

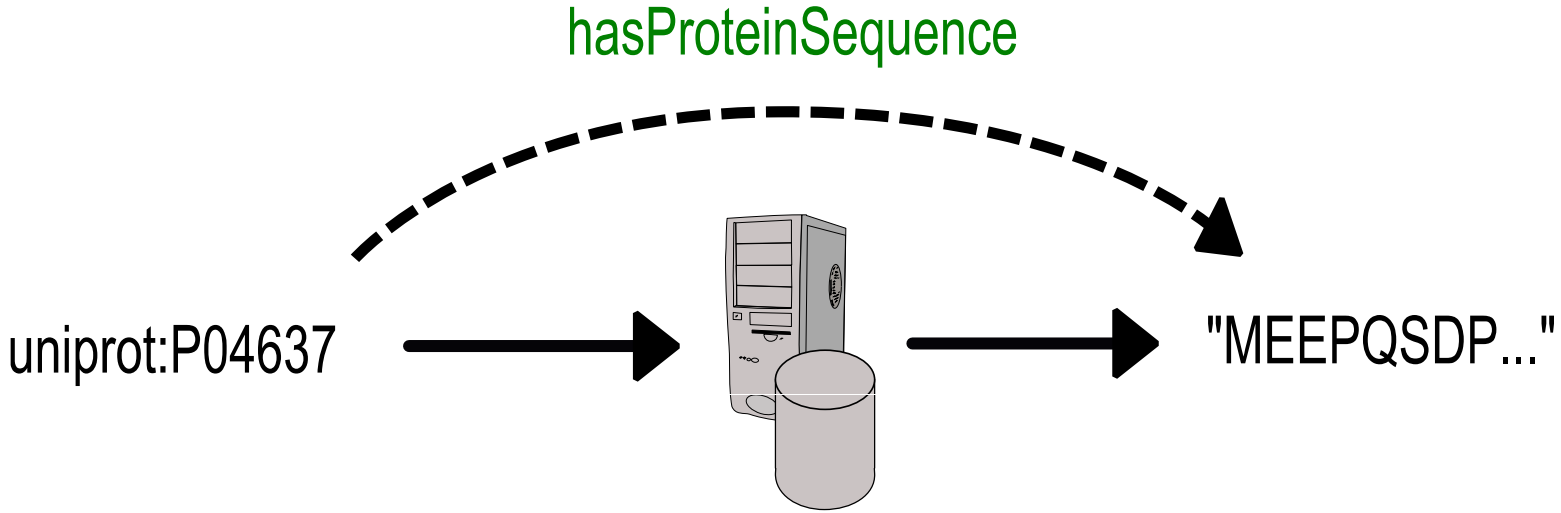
2010 CEUR: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-645/

Semantic Health And Research Environment (SHARE) prototype.



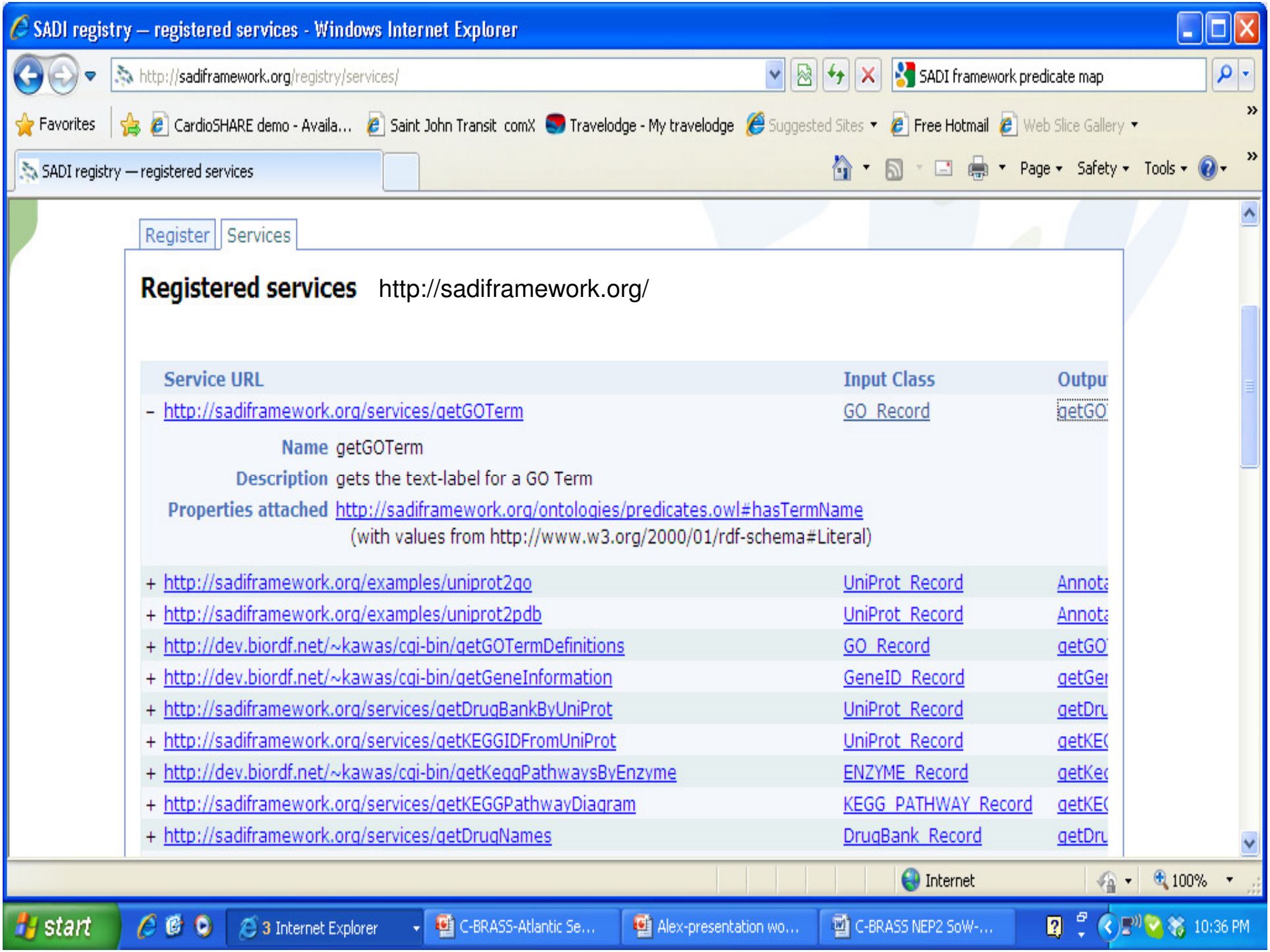
Slide by Michel Dumontier

hasProteinSequence



Input Data:	BRCA1 rdf:type Gene ID
Output Data:	BRCA1 hasDNASequence AGCTTAGCCA…
Registry Index:	Service provides “hasDNASequence” property to Gene IDs

Predicate-based web service invocation. Using the hasProteinSequence predicate in a query automatically invokes a web service capable of obtaining the amino acid sequence for UniProt entry P04637.



Input / Output Semantics: Example

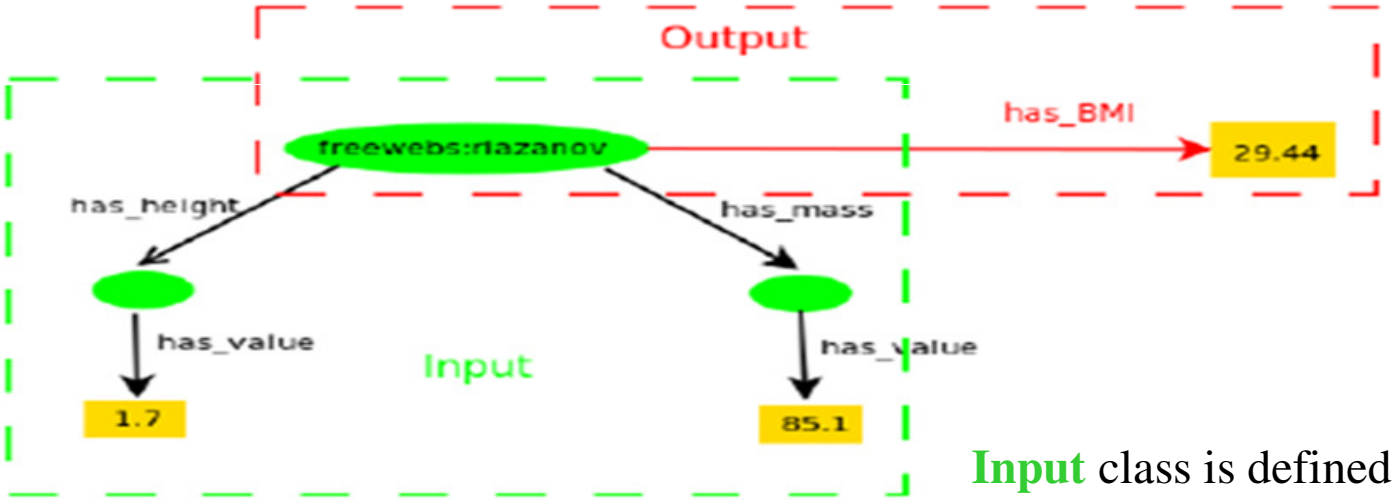
OWL class of the main input node specifies the service output.

- Uses existential property restrictions to define the predicates

of what the service does.

- Service computing BMI has this output class: (has_BMI exactly 1 xsd:float) advertising that it computes the Body Mass Index.
- Service annotates its input with the predicate has BMI with typical **output:**

<http://www.freewebs.com/riazanov> has_BMI float"29.44"



Input class is defined as by

(has_mass some (has_value some xsd:float))

(has_height some (has_value some xsd:float))

the service expects something like this in the input:

<http://www.freewebs.com/riazanov> has_height [has_value float"1.7"] .

<http://www.freewebs.com/riazanov> has_mass [has_value float"85.1"] .

Mutation Impact Pipeline

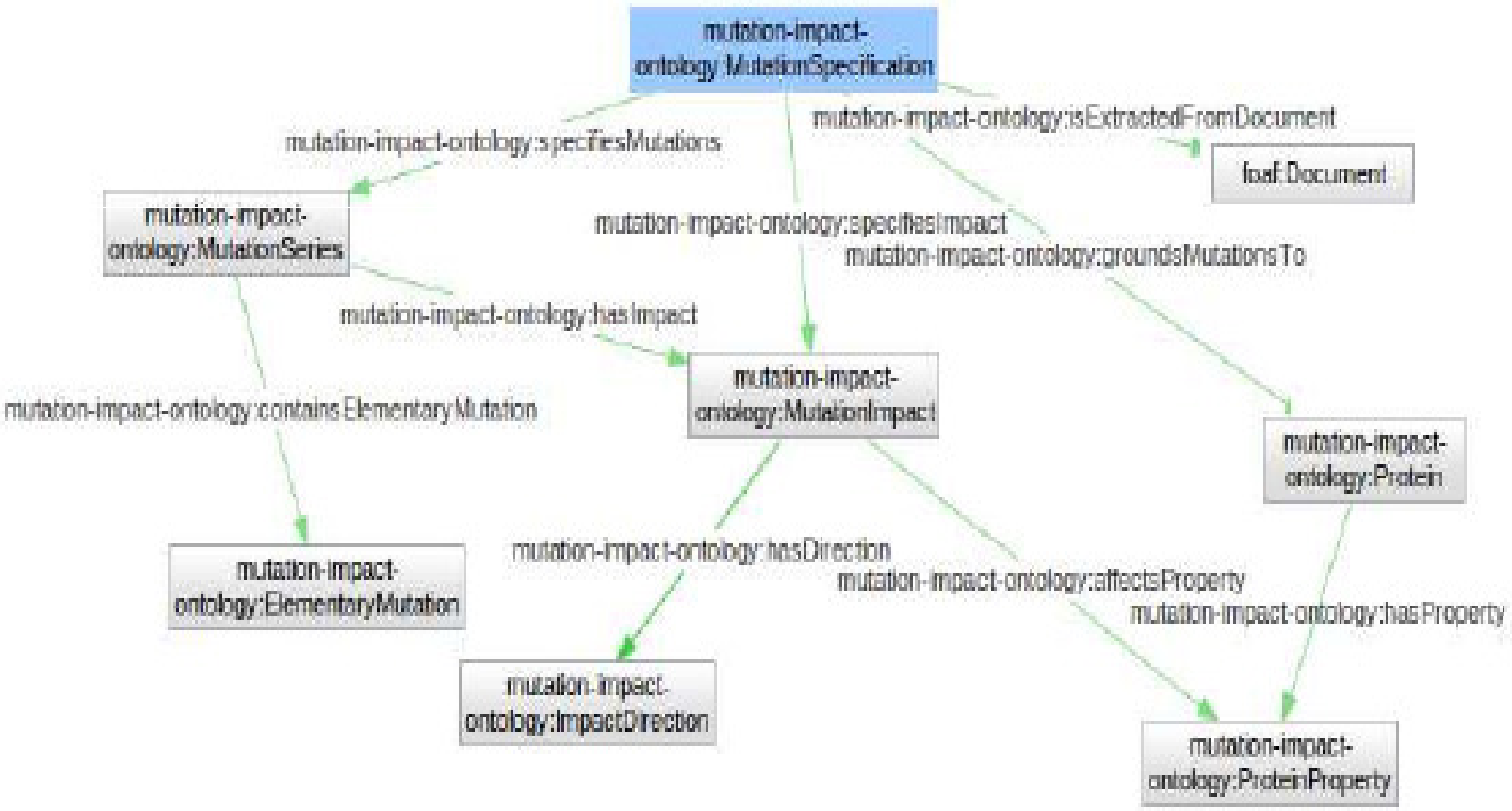
The haloalkane dehalogenase from the nitrogen-fixing hydrogen bacterium Xanthobacter autotrophicus GJ10 (DhlA) prefers 1,2-dichloroethane (DCE) as substrate and converts it to 2-chloroethanol and chloride ... DhlA shows only a small decrease in activity when Trp-125 is replaced with phenylalanine

$\Longrightarrow$

- "haloalkane dehalogenase" is a protein
- its UniProt Id is P22643
- "Trp-125 is replaced with phenylalanine" is the point mutation W125F
- "activity" is the protein property being affected, GO_00188786 in Gene Ontology
- "decrease" means the impact is negative

UNB-Concordia joint work. GATE pipeline wrapped as a Java library. Outputs Java objects representing mutation specifications: statements about mutations and their impacts on protein properties.

Mutation Impact Ontology



Text -> Mutation Specifications

- Just wraps the Mutation Impact mining pipeline as a SADI service
- Identifies input text in the service input, submits it to the pipeline and converts the results to an RDF graph

Input:

```
<http://example.com/text1>   rss:link   anyURI"http://example.com/text1"
```

Output:

```
<http://example.com/text1>  foaf:topic              midb:mutationSpec243
miodb:mutationSpec243       mio:groundsMutationsTo  miodb:protein528
miodb:protein528            mio:hasSwissProtId      "P22643"
miodb:mutationSpec243       mio:specifiesImpact     miodb:mutationImpact624
miodb:mutationImpact624     mio:hasDirection        mio:Positive
miodb:mutationImpact624     mio:affectsProperty     miodb:proteinProperty326
miodb:proteinProperty326    mio:hasType             GO:GO_0018786
                  . . . . . .
```

Services: Mutation Impact DB

- Wildtype protein $\longrightarrow$ mutation specifications (complete description of)
- Mutant protein $\longrightarrow$ mutation specifications
- Specific property of a specific protein $\longrightarrow$ known mutation impacts
- Mutation impact (on a specific property of a specific protein) $\longrightarrow$ mutation specifications
- Bio entity type (e.g., mio:Protein or mio:PointMutation) $\longrightarrow$ known instances
- Set of elementary mutations $\longrightarrow$ subsets described in the literature (with links to the corresponding mio:MutationSpecification)

Use Case Queries

Samples use cases queries are online:

http://unbsj.biordf.net/mutation-impact/queries_Dec9.html

and can be run at

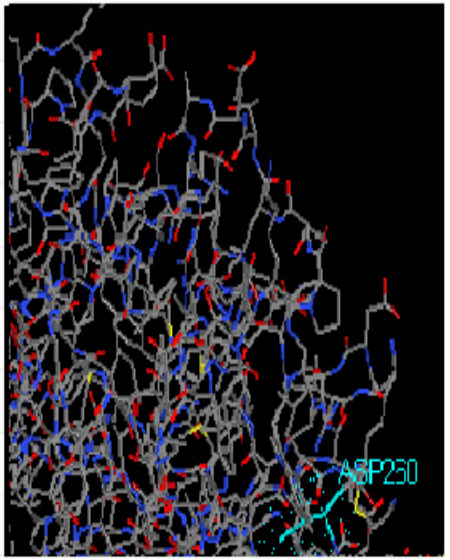
http://138.119.1.172:8080/cardioSHARE-mutations/

Use Case 1

- A protein engineer is looking for mutations that can improve catalitic activity of an enzime
- Query: find all mutations and the structure images of wild type proteins that were mutated, where the impact of the mutation is an enhanced haloalkane dehalogenase activity (GO_0018786)
- Predicates from our ontology take us from GO_0018786 to mutations and proteins: $proteinPropertyHasType + affectsProperty + hasDirection + specifiesImpact + containsElementaryMutation + groundMutationsTo$
- Two external SADI services provide $has3DStructure$ to link proteins to their structure descriptions, and $hasJmol3DStructureVisualization$ to link the structure to a 3D image file

Use Case 1 Query

```
PREFIX mio:<http://unbsj.biordf.net/ontologies/mutation-impact-ontology.owl#>
PREFIX mioe:<http://unbsj.biordf.net/ontologies/mutation-impact-ontology-extras.owl#>
PREFIX go:<http://purl.org/obo/owl/GO#>
PREFIX sio:<http://semanticscience.org/resource/>
PREFIX props:<http://sadiframework.org/ontologies/properties.owl#>
PREFIX objects:<http://sadiframework.org/ontologies/service_objects.owl#>
SELECT DISTINCT ?NormalizedMutation ?Protein ?Visualisation
FROM <http://unbsj.biordf.net/mutation-impact/service-data/protein_property_types.rdf>
WHERE { # impact <-- property instance ?Impact mio:affectProperty
?Property .
# protein property instance <-- GO_0018786 ?Property
mioe:proteinPropertyHasType go:GO_0018786 .
# check that the impact is positive
?Impact mio:hasDirection mio:Positive .
# grounded mutation
<-- impact ?MutationSpecification mio:specifiesImpact ?Impact .
# grounded mutation --> wildtype protein ?MutationSpecification
mio:groundMutationsTo ?Protein .
# grounded mutation --> point mutation series
?MutationSeries mio:mutationSeriesIsSpecifiedBy
?MutationSpecification .
# point mutation series  --> separate point mutations
?MutationSeries mio:containsElementaryMutation ?Mutation .
?Mutation mio:hasNormalizedForm ?NormalizedMutation .
# grounded mutation
--> Web page with Jmol applet call ?MutationSpecification
objects:hasJmol3DStructureVisualization ?Visualisation . }
```

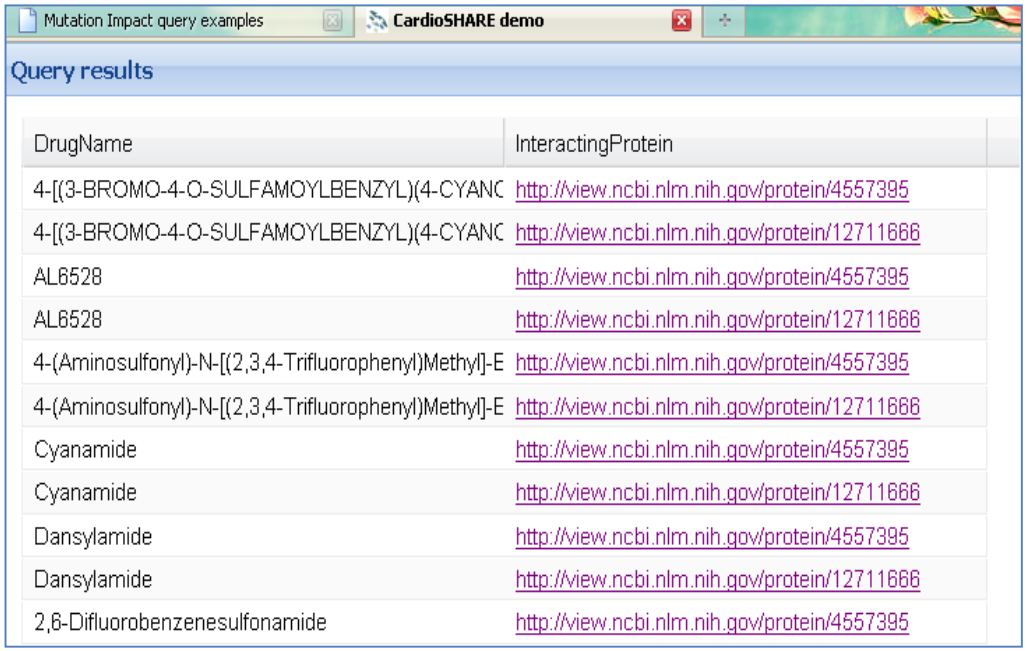
Normal	Protein	Visualisation
D260N	http://biordf.net/moby/UniProt/P22643	http://unbsj.biordf.net/mut-vis/visualiseMut
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	
D260N	http://biordf.net/moby/UniProt/P22643	

Use Case 3

- A researcher in drug discovery is looking for existing drugs targeting a new disease condition
- Query: find all drugs related to mutated proteins, together with their interaction partners, where the mutation impact is a increased carbonic anhydrase activity (GO_0008270)
- Our predicates link GO_0008270 to proteins via the instances of this protein properties, positive impacts and mutation specifications
- External ontologies facilitate the linking of proteins to the IDs of drugs affecting them
- Another external predicate, $hasMolecularInteractionWith$, links the proteins to proteins they interact with

Use Case 3 Query

```
PREFIX mio:http://unbsj.biordf.net/ontologies/mutation-impact-ontology.owl#>
PREFIX mioe:<http://unbsj.biordf.net/ontologies/mutation-impact-ontology-extras.owl#>
PREFIX go:<http://purl.org/obo/owl/GO#>
PREFIX objects:<http://sadiframework.org/ontologies/service_objects.owl#>
PREFIX pred: <http://sadiframework.org/ontologies/predicates.owl#>
SELECT ?DrugName ?InteractingProtein
FROM <http://unbsj.biordf.net/mutation-impact/service-data/protein_property_types.rdf>
WHERE {# enumerate known instances of go:GO_0008270
?Property mioe:proteinPropertyHasType go:GO_0008270 .
# impact <-- protein property instance ?Impact mio:affectProperty
?Property .
# check that the impact is positive ?Impact mio:hasDirection
mio:Positive .
# grounded mutation <-- impact ?MutationSpecification
mio:specifiesImpact ?Impact .
# grounded mutation --> wildtype protein
?MutationSpecification mio:groundMutationsTo ?Protein .
# wildtype protein --> drug ?Protein objects:isTargetOfDrug
?Drug . ?Drug objects:hasDrugGenericName ?DrugName .
# wildtype protein --> interacting proteins ?Protein
pred:hasMolecularInteractionWith ?InteractingProtein }
```



DrugName	InteractingProtein
4-[(3-BROMO-4-O-SULFAMOYLBENZYL)(4-CYANC	http://view.ncbi.nlm.nih.gov/protein/4557395
4-[(3-BROMO-4-O-SULFAMOYLBENZYL)(4-CYANC	http://view.ncbi.nlm.nih.gov/protein/12711666
AL6528	http://view.ncbi.nlm.nih.gov/protein/4557395
AL6528	http://view.ncbi.nlm.nih.gov/protein/12711666
4-(Aminosulfonyl)-N-[(2,3,4-Trifluorophenyl)Methyl]-E	http://view.ncbi.nlm.nih.gov/protein/4557395
4-(Aminosulfonyl)-N-[(2,3,4-Trifluorophenyl)Methyl]-E	http://view.ncbi.nlm.nih.gov/protein/12711666
Cyanamide	http://view.ncbi.nlm.nih.gov/protein/4557395
Cyanamide	http://view.ncbi.nlm.nih.gov/protein/12711666
Dansylamide	http://view.ncbi.nlm.nih.gov/protein/4557395
Dansylamide	http://view.ncbi.nlm.nih.gov/protein/12711666
2,6-Difluorobenzenesulfonamide	http://view.ncbi.nlm.nih.gov/protein/4557395

Use Case 4

In this use case, a genomics researcher asks for all known mutations reported in the literature for a protein containing a non-synonymous SNP. Here the researcher is primarily looking for any literature describing impacts of a nsSNP on a protein. By retrieving all known mutations for the protein in which the nsSNP is reported, the researcher can find out if any of these reported mutations corresponds to the location of the SNP in question. Minimally the researcher can retrieve the full set of mutations to the protein based on reported experimental analysis and their impacts, together with references to the supporting literature.

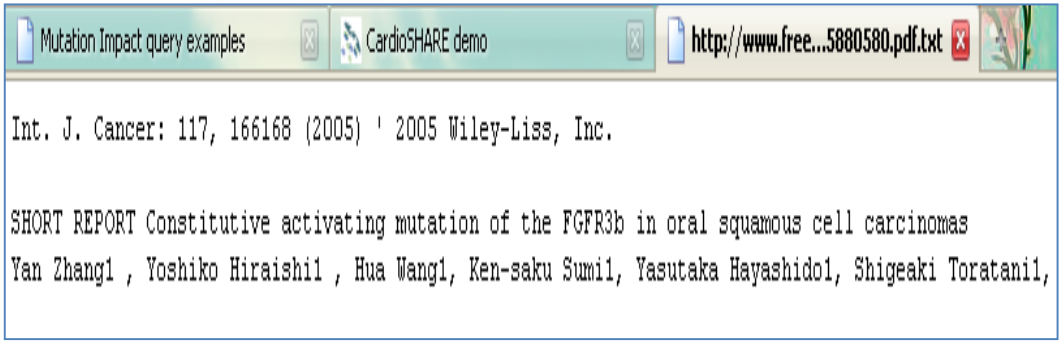
Use Case 4 Query

Query: Find all documented mutations of the protein with SNP rs2305178.

```
PREFIX mio:<http://unbsj.biordf.net/ontologies/mutation-impact-ontology.owl#>
PREFIX mioe:<http://unbsj.biordf.net/ontologies/mutation-impact-ontology-extras.owl#>
PREFIX dbsnp:<http://lsrn.org/dbSNP:>
PREFIX objects:<http://sadiframework.org/ontologies/service_objects.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rss: <http://purl.org/rss/1.0/>
PREFIX pred: <http://sadiframework.org/ontologies/predicates.owl#>
PREFIX sio: <http://semanticscience.org/resource/>
SELECT DISTINCT ?NormalizedMutation ?DocumentURL
WHERE { # SNP --> gene (Entrez)
# 'is variant of' dbsnp:rs2305178 sio:SIO_000272 ?EzGene .
# enumerate known proteins ?Protein mioe:biologicalEntityHasType
mio:Protein .
# proteins --> genes (KEGG) ?Protein pred:isEncodedBy?KeggGene
# gene (KEGG) --> reference sequence ?KeggGene
objects:hasRefSeqTranscript ?RefSeq .
# reference sequence --> gene (Entrez) ?RefSeq
objects:correspondsToEntrezGene ?EzGene .
# protein --> mutation info ?MutationSpecification
mio:groundMutationsTo ?Protein . ?MutationSeries
mio:mutationSeriesIsSpecifiedBy ?MutationSpecification .
?MutationSeries mio:containsElementaryMutation ?Mutation .
?Mutation mio:hasNormalizedForm ?NormalizedMutation .
# mutation --> literature reference ?Document foaf:topic
?MutationSpecification . ?Document rss:link ?DocumentURL }
```

Query results

DocumentURL	NormalizedMutation
http://www.freewebs.com/riazanov/15880580.pdf.txt	G697C



Summary

- Rule based mutation and impact extraction
- Methods for grounding of mutations to protein sequences and protein functions to Gene Ontology
- Algorithms deployed with Semantic Assistant
- Algorithms and mutation information exposed with semantic metadata and as semantic web services (SWS)
- Reuse of SWS Mutation services with multiple use cases

Acknowledgements

Jonas B. Laurila

Alexandre Riazanov

Alexandre Kouznetsov

*** Christopher J.O. Baker**

Computer Science & Applied Statistics

University of New Brunswick

Saint John, Canada

Nona Naderi

René Witte

Computer Science & Software Engineering

Concordia University

Montreal, Canada

This research was funded in part by:

- New Brunswcik Innovation Foundation, New Brunswick, Canada
- NSERC, Discovery Grant, Canada
- Quebec – New Brunswick Co-operation in Advanced Education – Research Program, Government of New Brunswick, Canada

Acknowledgements

Wee Tiong Ang
Justin Choo
Rajaraman Kanagasabai
Institute for Infocomm Research
A-STAR, Singapore

Luke McCarthy
Mark Wilkinson
University of British Columbia
Vancouver, Canada

This research was also funded in part by:

- Agency for Science, Technology and Research, Singapore
- CANADA'S Advanced Research And Innovation Network, CANARIE, Canada