A renaissance for the point mutation:
from legacy data to semantic web services

Christopher J. O. Baker
University of New Brunswick
• 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)**

Wikipedia¹:

- "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%

¹http://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.....
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

<table>
<thead>
<tr>
<th></th>
<th>Diseases</th>
<th>Distinct genes</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIM</td>
<td>1864</td>
<td>2708</td>
<td>4851</td>
</tr>
<tr>
<td>GAD</td>
<td>2303</td>
<td>1740</td>
<td>7945</td>
</tr>
</tbody>
</table>

- 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
Mutation Mining: a history

2005 2006 2007 2008 2009 2010

Front ISR 2005 - Baker and Witte - Mutation Mining
JBCB 2007 - Kanagasabai etal – mSTRAP Workflow
JBCB 2007 - Witte and Baker - Systematic Evaluation Mutation Extraction
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.

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

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
    - 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.

- 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 et al. 2010)
  – Pathway (Bauer Mehren et al. 2009)
    – Pubmed Abstracts (Laurila et al. 2010)
    – Protein Structure (Kanagasabai et al. 2007)

• Impact Prediction
  – SNAP (Bromberg and Rost 2008)
  – Membrane protein stability (Winnenburg et al. 2009)

• Federated Query over Mutation Triplestores
  – SADI Semantic Web Services (Riazanov et al. 2010)
The mutation increased temp stability by 7 °C …

**E214F**, 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
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.
Sample Ontologies

Listed below are some sample ontologies that we have created:

- Phosphotase Mutation
- BRCA1/BRCA2 Mutation
Ontology Version 2.0

**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.


mSTRAP

[Image of a molecular structure]

1) **Sentence 1**: However, because translocation of GLUT4 in cells overexpressing a dominant inhibitory PTEN mutant (C124S) was similar to that of control cells, we conclude that endogenous PTEN may not modulate metabolic functions of insulin under normal physiological conditions. 2001 Academic Press Key Words: metabolism, signal transduction, insulin resistance, phosphatase, glucose.

2) **Sentence 2**: C124S is a dominant inhibitory PTEN mutant (36).

3) **Sentence 3**: Using an Akt phosphorylation assay to assess effects of PTEN to
**mSTRAP**

* A System for Mutation Extraction and Structure Annotation Visualization

**Home | Download | Ontologies | Guide | References | Contact**

---

**News**


- 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
Algorithms and semantic infrastructure for mutation impact extraction and grounding

Jonas B Laurila¹, Nona Naderi², René Witte², Alexandre Riazanov¹, Alexandre Kouznetsov¹, Christopher JO Baker¹*

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.

Jonas B Laurila¹, Alexandre Kouznetsov¹ and Christopher J O Baker*¹

AIMM-2010 @ ECCB 2010

Annotation, Interpretation and Management of Mutations 2010 Ghent, Belgium.
Replacement of tryptophan residues in haloalkane dehalogenase reduces halide binding and catalytic activity.

Kennes C, Prieß F, Krooshof GH, Bakma E, Kingma J, Janssen DB.

Department of Biochemistry, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands.

Haloalkane dehalogenase catalyzes the hydrolytic cleavage of carbon-halogen bonds in short-chain haloalkanes. Two tryptophan residues of the enzyme (Trp125 and Trp175) form a halide-binding site in the active site cavity, and were proposed to play a role in catalysis. The function of these residues was studied by replacing Trp125 with phenylalanine, glutamine or arginine and Trp175 by glutamine using site-directed mutagenesis. All mutants showed a more than 10-fold reduced kcat and much higher Km values with 1,2-dichloroethane and 1,2-dibromoethane than the wild-type enzyme. Fluorescence quenching experiments showed a decrease in the affinity of the mutants for deuterated molecules. The 2H kinetic isotope effect observed with the wild-type enzyme in deuterium oxide was lost in the mutants. The results indicate that both tryptophans are involved in stabilizing the nucleophilic substitution reaction that causes carbon-halogen bond cleavage.
Point Mutations

Crystallographic and kinetic evidence of a collision complex formed during halide import in haloalkane dehalogenase.

Pikkemaat MG, Ridder IS, Rozeboom HJ, Kalk KH, Dijkstra BW, Janssen DB.

Laboratory of Biochemistry, BIOSON Research Institute, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands.

Haloalkane dehalogenase (DhIA) converts haloalkanes to their corresponding alcohols and halide ions. The rate-limiting step in the reaction of DhIA is the release of the halide ion. The kinetics of halide release have been analyzed by measuring halide binding with stopped-flow fluorescence experiments. At high halide concentrations, halide import occurs predominantly via the rapid formation of a weak initial collision complex, followed by transport of the ion to the active site. To obtain more insight in this collision complex, we determined the X-ray structure of DhIA in the presence of bromide and investigated the kinetics of mutants that were constructed on the basis of this structure. The X-ray structure revealed one bromide ion firmly bound in the active site and two bromide ions weakly bound on the surface of the enzyme. One of the weakly bound ions is close to Thr197 and Phe294, near the entrance of the earlier proposed tunnel for substrate import. Kinetic analysis of bromide import by the Thr197Ala and Phe294Ala mutants of DhIA at high halide concentration showed that the rate constants for halide binding no longer displayed an increase with increasing bromide concentrations. This is in agreement with an elimination of a surface-located halide-binding site. Likewise, chloride binding kinetics of the mutants indicated with wild-type enzyme. The results indicate that Thr197 and Phe294 are involved in the formation of an halide import in DhIA and provide experimental evidence for the role of the tunnel in substrate and

Semantic Assistant Framework

René Witte and Thomas Gitzinger.

*Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients.*


Springer LNCS 5367, pp. 360–374. *(Acceptance rate: 31%)*
**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
“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 GJ10** hydrolyses terminally chlorinated and brominated n-alkanes to the corresponding alcohols."
“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 GJ10 hydrolyses terminally chlorinated and brominated n-alkanes to the corresponding alcohols.”

<table>
<thead>
<tr>
<th>Direction</th>
<th>Protein Property</th>
</tr>
</thead>
</table>

Named entities of interest

- Mutations
- Proteins
- Genes
- Organisms
- Protein properties
  - Protein functions (*activity, binding etc*)
  - Kinetic variables (*Km, kcat etc*)
  - (stability)
1. Retrieve and normalize mutations

2. For each candidate sequence
   1. For each pair of mutations
      1. Make regexp $w_1.(N_2-N_1)w_2$
      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

1. MGAKACYGAKCVAVAIVAGASSESLGKEQY
2. MAPEALFDKTYGKVVWSFGVLLWEITL
3. MQVSLEYGSMSSNTPVRIARLSSGEGPT

Candidate mutations

K5Q
Y8C
G9C
K11E
V35Q
Mutation Grounding: Example

Candidate sequences

1. MGAKACYGAKCVAVAIVAGASAESLGEQY
2. MAPEALFDRKYTYGKKVWSFGVLLWEIFTL
3. MQVSLESYGSMSNTPLVRIARLSSGEGPT

Candidate mutations

{ K5Q, Y8C, G9C, K11E, V35Q }

Compute regular expression
Mutation Grounding: Example

Candidate sequences

1. MGAKACY3AKCVAVAIWAGASSESLGKFLY
2. MAPEALDFRKYTYGKAVWSFVGVLWEITFL
3. MQVSLRYGSMSSNTPLVRIARLSSGEGPT

Candidate mutations

{ K5Q, Y8C, G9C, K11E, V35Q }

Match with sequence 1
Mutation Grounding: Example

Candidate sequences:

| 1 | MGA | KCHBKCVAVAIVAGASSSLGKEQY |
| 2 | MAPEALFDRKTYGARVNSCFQULLWEIFTL |
| 3 | MQVLSESYGSMSNTPLVRAIRLSSGEPF |

Candidate mutations:

| 1 | K5Q |
| 2 | Y8C |
| 3 | G9C |
| 3 | K11E |
| 3 | V35Q |

Extend match A
Mutation Grounding: Example

Candidate sequences

1. MGAKACYGAKCVAVAIVAGASSESLEEQY
2. MAPEALFDROYTYGKVWSFGVLLWEIFTL
3. MQVSLESYGSSNTPLVRIARLSSGEGPT

Candidate mutations

K5Q, Y8C, G9C, K11E, V35Q

Extend match B
Mutation Grounding: Example

Candidate sequences

1. MGAKAGLQKCVAVAIVAGASESLSGKEQY
2. MAPEALFDQKTYTHKVKWSFGVLLWEITL
3. MQVSLESYKSMSSNTPLVRIARLSSGEQPT

Candidate mutations

- K5Q
- Y9C
- G9C
- K11E
- V35Q

Match with sequence 2 & 3
Mutation Grounding: Example

<table>
<thead>
<tr>
<th>Candidate sequences</th>
<th># Mutations / Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGAKA\underline{C}Y3-KCVAVAIVAGASSESLGKEQY</td>
<td>4 / -1</td>
</tr>
<tr>
<td>MAPEAL\underline{F}DREYTHK\underline{R}K\underline{S}FGVLLWEIFTL</td>
<td>4 / 5</td>
</tr>
<tr>
<td>MQVSELE\underline{S}S\underline{M}SSNTPLVRIARL\underline{S}SGE\underline{G}PT</td>
<td>2 / 0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Precision</th>
<th>Recall</th>
<th>Corpus size</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIK3CA</td>
<td>0.86</td>
<td>0.70</td>
<td>30</td>
</tr>
<tr>
<td>FGFR3</td>
<td>0.89</td>
<td>0.66</td>
<td>26</td>
</tr>
<tr>
<td>MEN1</td>
<td>0.54</td>
<td>0.32</td>
<td>7</td>
</tr>
<tr>
<td>Haloalkane</td>
<td>0.83</td>
<td>0.73</td>
<td>13</td>
</tr>
<tr>
<td>Dehalogenases</td>
<td>0.84</td>
<td>0.65</td>
<td>76</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.84</strong></td>
<td><strong>0.65</strong></td>
<td><strong>76</strong></td>
</tr>
</tbody>
</table>
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 occurrence 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
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  \quad s = \frac{1}{2 \times 3} = \frac{1}{6}
dehalogen activ  \quad s = \frac{4}{2 \times 3} = \frac{4}{6}
activ  \quad s = \frac{1}{1 \times 3} = \frac{2}{6}
Protein function grounding
Example:

ground to a certain degree

\((\text{GO:0018786})\) haloalk dehalogen activ

\[
\begin{align*}
\text{anhydr activ} & \rightarrow ((\text{GO:0018786}), 0.17) \quad (3) \\
\text{dehalogen activ} & \rightarrow ((\text{GO:0018786}), 0.67) \quad (1) \\
\text{activ} & \rightarrow ((\text{GO:0018786}), 0.33) \quad (2)
\end{align*}
\]
### Impact direction term lists

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>(cont.)</th>
<th>Neutral</th>
<th>Negation</th>
<th>Non-Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>increase</td>
<td>abolish</td>
<td>loose</td>
<td>identical</td>
<td>without</td>
<td>affect</td>
</tr>
<tr>
<td>-increases</td>
<td>decrease</td>
<td>defect</td>
<td>similar</td>
<td>no</td>
<td>effect</td>
</tr>
<tr>
<td>-increased</td>
<td>reduce</td>
<td>disrupt</td>
<td>full</td>
<td>not</td>
<td>alter</td>
</tr>
<tr>
<td>-increasing</td>
<td>lower</td>
<td>diminish</td>
<td></td>
<td></td>
<td>differ</td>
</tr>
<tr>
<td>enhance</td>
<td>inhibit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>higher</td>
<td>impair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>improve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
#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

DhIA 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?
The NAR Subset

Subset where *MutationFinder* recognizes more than one point mutation ➔ Σ 1146 full text articles
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 $GPM$s
  - 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

<table>
<thead>
<tr>
<th>#</th>
<th>Mut</th>
<th>SNAPMAT Pred</th>
<th>Reference</th>
<th>Func Change/Impact</th>
<th>GO ID</th>
<th>Agreement</th>
<th>GO Decrip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R165W</td>
<td></td>
<td>34</td>
<td>positive</td>
<td>GO_0004016</td>
<td>Y</td>
<td>adenylate cyclase activity</td>
</tr>
<tr>
<td>17</td>
<td>F51L</td>
<td>Non-neutral</td>
<td>16</td>
<td>negative</td>
<td>GO_0004016</td>
<td>N</td>
<td>adenylate cyclase activity</td>
</tr>
<tr>
<td>36</td>
<td>I102T</td>
<td>Neutral</td>
<td>16</td>
<td>negative</td>
<td>GO_0004016</td>
<td>Y</td>
<td>adenylate cyclase activity</td>
</tr>
<tr>
<td>70</td>
<td>A154D</td>
<td>Non-neutral</td>
<td>16</td>
<td>negative</td>
<td>GO_00044144</td>
<td></td>
<td>diacylglycerol O-acyltransferase activity</td>
</tr>
<tr>
<td></td>
<td>S127L</td>
<td></td>
<td>43</td>
<td>positive</td>
<td>GO_00044144</td>
<td></td>
<td>diacylglycerol O-acyltransferase activity</td>
</tr>
<tr>
<td></td>
<td>T11S</td>
<td></td>
<td>13</td>
<td>negative</td>
<td>GO_00044144</td>
<td></td>
<td>diacylglycerol O-acyltransferase activity</td>
</tr>
<tr>
<td></td>
<td>R18C</td>
<td></td>
<td>13</td>
<td>negative</td>
<td>GO_00044144</td>
<td></td>
<td>diacylglycerol O-acyltransferase activity</td>
</tr>
<tr>
<td></td>
<td>F121R</td>
<td></td>
<td>19</td>
<td>negative</td>
<td>GO_0004977</td>
<td></td>
<td>melanocortin receptor activity</td>
</tr>
<tr>
<td>44</td>
<td>N123A</td>
<td>Non-neutral</td>
<td>38</td>
<td>negative</td>
<td>GO_0004978</td>
<td>Y</td>
<td>adrenocorticotropin receptor activity</td>
</tr>
<tr>
<td>45</td>
<td>N123D</td>
<td>Non-neutral</td>
<td>38</td>
<td>negative</td>
<td>GO_0004978</td>
<td>Y</td>
<td>adrenocorticotropin receptor activity</td>
</tr>
<tr>
<td></td>
<td>D126A</td>
<td></td>
<td>38</td>
<td>neutral</td>
<td>GO_0004978</td>
<td></td>
<td>adrenocorticotropin receptor activity</td>
</tr>
<tr>
<td>156</td>
<td>I316T</td>
<td>Non-neutral</td>
<td>33</td>
<td>negative</td>
<td>GO_0005000</td>
<td>Y</td>
<td>Vasopressin receptor activity</td>
</tr>
<tr>
<td>15</td>
<td>E49A</td>
<td>Non-neutral</td>
<td>29</td>
<td>negative</td>
<td>GO_0005184</td>
<td>Y</td>
<td>neuropeptide hormone activity</td>
</tr>
<tr>
<td>26</td>
<td>D90A</td>
<td>Non-neutral</td>
<td>29</td>
<td>negative</td>
<td>GO_0005184</td>
<td>Y</td>
<td>neuropeptide hormone activity</td>
</tr>
<tr>
<td></td>
<td>E100A</td>
<td></td>
<td>29</td>
<td>negative</td>
<td>GO_0005184</td>
<td>Y</td>
<td>neuropeptide hormone activity</td>
</tr>
<tr>
<td></td>
<td>E100C</td>
<td></td>
<td>1</td>
<td>negative</td>
<td>GO_0005515</td>
<td></td>
<td>protein binding</td>
</tr>
</tbody>
</table>

In silico mutagenesis: a case study of the melanocortin 4 receptor.

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.

AIMM publications

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


2010 CEUR: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WSVol-645/
Semantic Health And Research Environment (SHARE) prototype.
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.
Registered services  http://sadiframework.org/

<table>
<thead>
<tr>
<th>Service URL</th>
<th>Input Class</th>
<th>Output Class</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://sadiframework.org/services/getGOTerm">http://sadiframework.org/services/getGOTerm</a></td>
<td>GO_Record</td>
<td>getGO</td>
</tr>
<tr>
<td>Name getGOTerm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description gets the text-label for a GO Term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Properties attached <a href="http://sadiframework.org/ontologies/predicates.owl#hasTermName">http://sadiframework.org/ontologies/predicates.owl#hasTermName</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(with values from <a href="http://www.w3.org/2000/01/rdf-schema#Literal">http://www.w3.org/2000/01/rdf-schema#Literal</a>)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/examples/uniprot2go">http://sadiframework.org/examples/uniprot2go</a></td>
<td>UniProt_Record</td>
<td>Annot</td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/examples/uniprot2pdb">http://sadiframework.org/examples/uniprot2pdb</a></td>
<td>UniProt_Record</td>
<td>Annot</td>
</tr>
<tr>
<td>+ <a href="http://dev.biordf.net/~kawas/cgi-bin/getGOTermDefinitions">http://dev.biordf.net/~kawas/cgi-bin/getGOTermDefinitions</a></td>
<td>GO_Record</td>
<td>getGO</td>
</tr>
<tr>
<td>+ <a href="http://dev.biordf.net/~kawas/cgi-bin/getGeneInformation">http://dev.biordf.net/~kawas/cgi-bin/getGeneInformation</a></td>
<td>GeneID_Record</td>
<td>getGene</td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/services/getDrugBankByUniprot">http://sadiframework.org/services/getDrugBankByUniprot</a></td>
<td>UniProt_Record</td>
<td>getDrugBank</td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/services/getKEGIDFromUniprot">http://sadiframework.org/services/getKEGIDFromUniprot</a></td>
<td>UniProt_Record</td>
<td>getKEGID</td>
</tr>
<tr>
<td>+ <a href="http://dev.biordf.net/~kawas/cgi-bin/getKeggPathwaysByEnzyme">http://dev.biordf.net/~kawas/cgi-bin/getKeggPathwaysByEnzyme</a></td>
<td>ENZYME_Record</td>
<td>getKeggPathway</td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/services/getKEGGPathwayDiagram">http://sadiframework.org/services/getKEGGPathwayDiagram</a></td>
<td>KEGG_PATHWAY_Record</td>
<td>getKEGGPathwayDiagram</td>
</tr>
<tr>
<td>+ <a href="http://sadiframework.org/services/getDrugNames">http://sadiframework.org/services/getDrugNames</a></td>
<td>DrugBank_Record</td>
<td>getDrugNames</td>
</tr>
</tbody>
</table>
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"] .
The haloalkane dehalogenase from the nitrogen-fixing hydrogen bacterium *Xanthobacter autotrophicus G.J10* (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

- “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.
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:speficifiesImpact  miodb:mutationImpact624
miodb:mutationImpact624  mio:hasDirection  mio:Positive
miodb:mutationImpact624  mio:affectsProperty  miodb:proteinProperty326
miodb:proteinProperty326  mio:hasType  GO:GO_0018786
• Wildtype protein → mutation specifications (complete description of)

• Mutant protein → mutation specifications

• Specific property of a specific protein → known mutation impacts

• Mutation impact (on a specific property of a specific protein) → mutation specifications

• Bio entity type (e.g., mio:Protein or mio:PointMutation) → known instances

• Set of elementary mutations → subsets described in the literature (with links to the corresponding mio:MutationSpecification)
Samples use cases queries are online:


and can be run at

Use Case 1

- A protein engineer is looking for mutations that can improve catalytic activity of an enzyme
- 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
  # grounded mutation
  # grounded mutation --> wildtype protein
  # grounded mutation --> point mutation series
  ?MutationSeries mio:mutationSeriesIsSpecifiedBy
  # point mutation series --> separate point mutations
  # grounded mutation
  ?Mutation mio:hasNormalizedForm ?NormalizedMutation .
  # grounded mutation
  --> Web page with Jmol applet call
}

<table>
<thead>
<tr>
<th>Normal</th>
<th>Protein</th>
<th>Visualisation</th>
</tr>
</thead>
</table>
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 an 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
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#>
WHERE {# enumerate known instances of go:GO_0008270
?Property mioe:proteinPropertyHasType go:GO_0008270 .
# check that the impact is positive ?Impact mio:hasDirection mio:Positive .
# wildtype protein --> interacting proteins ?Protein pred:hasMolecularInteractionWith ?InteractingProtein }

<table>
<thead>
<tr>
<th>DrugName</th>
<th>InteractingProtein</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-[3-BROMO-4-O-SULFAQYL]BENZYL)-(4-CYANOCYANIMIDE)</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/4657395">View on NCBI</a></td>
</tr>
<tr>
<td>4-[3-BROMO-4-O-SULFAQYL]BENZYL)-(4-CYANOCYANIMIDE)</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>AL6528</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>AL6528</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>4-[Aminosulfonyl]-N-[(2,3,4-Trifluorophenyl)Methyl]-E</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/4657395">View on NCBI</a></td>
</tr>
<tr>
<td>4-[Aminosulfonyl]-N-[(2,3,4-Trifluorophenyl)Methyl]-E</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>Cyanamide</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/4657395">View on NCBI</a></td>
</tr>
<tr>
<td>Cyanamide</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>Dansylamide</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/4657395">View on NCBI</a></td>
</tr>
<tr>
<td>Dansylamide</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/12711868">View on NCBI</a></td>
</tr>
<tr>
<td>2,6-Difluorobenzensulfonamide</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/4657395">View on NCBI</a></td>
</tr>
</tbody>
</table>
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.
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/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX

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 .
# MutationSeries mio:containsElementaryMutation ?Mutation .
#Document rss:link ?DocumentURL }

Query results

<table>
<thead>
<tr>
<th>DocumentURL</th>
<th>NormalizedMutation</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.freewebs.com/riazanov/15880580.pdf.txt">http://www.freewebs.com/riazanov/15880580.pdf.txt</a></td>
<td>G897C</td>
</tr>
</tbody>
</table>


SHORT REPORT Constitutive activating mutation of the FGFR3b in oral squamous cell carcinomas
Tan Zhang1, Yoshiko Hisaishi1, Hua Wang1, Ken-eku Swmii, Yasutaka Hayashidol, Shigeaki Torazani1,
• 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 Brunswick 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