# Ontology-Based Extraction and Summarization of Protein Mutation Impact Information

**Nona Naderi and René Witte**
Department of Computer Science and Software Engineering
Concordia University, Montréal, Canada

## 1 Introduction

NLP methods for extracting mutation information from the bibliome have become an important new research area within bio-NLP, as manually curated databases, like the Protein Mutant Database (PMD) (Kawabata et al., 1999), cannot keep up with the rapid pace of mutation research. However, while significant progress has been made with respect to mutation detection, the automated extraction of the *impacts* of these mutations has so far not been targeted. In this paper, we describe the first work to automatically summarize impact information from protein mutations. Our approach is based on populating an OWL-DL ontology with impact information, which can then be queried to provide structured information, including a summary.

## 2 Background

Mutations are alterations, rearrangements, or duplications of genetic material, impacting protein properties like stability or activity. For example:

> *H86A/E/F/K/Q/W decreased the enzyme stability at 60° C by up to 95% and the transition temperature by 2.5° C to 5.8° C.*

Impacts are described through other concepts, since mutational events may cause changes to physical quantities such as *pH* and *temperature*. As presented in the above example, the named mutations (H86A/E/F/K/Q/W) made changes to the thermostability by 2.5–2.8 degrees Celsius. Hence, we extract (i) units of measurement, e.g., *%*, *degree Celsius*, *min*; (ii) protein properties: *stability*, *activity* and others; and (iii) impact words, including *increase*, *stabilize*, and *reduce*.

Measurable impacts can thus be classified based on the type of effect (increase, decrease or destabilize) on the protein property.

## 3 Related work

Little previous work exists on automatically detecting and extracting mutation impacts. An exception is EnzyMiner (Yeniterzi and Sezerman, 2009), which performs document classification for disease-related mutations. This work differs significantly from ours, as we are concerned with sentence-level impact detection and summarization.

## 4 Mutation Impact Detection

Our main contribution for impact detection and summarization consists of two major parts: an ontology describing impacts on a semantic level, and an NLP pipeline for detecting impacts in documents in order to populate the ontology. Further analysis, including summarization, can then be performed on this NLP-populated ontology through ontology queries and reasoning.

**Ontology Design.** Our *Mutation Impact Ontology* conceptualizes impacts and the mutations associated with them. The main concepts are: Mutation: An alteration or a change to a gene and developing a different offspring. UnitOfMeasurement: A class for measurement units. MutImpact: Mutation effect on protein properties. ProteinProperty: A class for properties of "Protein" and subclassed by different properties like "Activity" and "Stability." To design the Mutation Impact Ontology, information about several other elements is needed: Text elements, biological entities and entity relations. The relations between these entities are expressed as OWL object properties.

**Mutation Impact Extraction.** Impacts are detected through a combination of an *OntoGazetteer* annotating impact words, measurement units, etc., and JAPE grammar transducers, e.g.:

```
Rule: MutationImpact
({Lookup.majorType == "onto_impact"}):impact --> {
try {
    // get Impact annotations
    gate.AnnotationSet impactSet = (gate.AnnotationSet)bindings.get("impact");
    ...
}
```

Here, the impact word that is marked as "Lookup" with a feature of "majorType," "onto_impact" is



*Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010*, pages 128–129,
Uppsala, Sweden, 15 July 2010. 

annotated as "MutImpact." Accordingly, "Protein-Property" and "UnitOfMeasurement" are annotated through similar JAPE grammars. Finally, each sentence is annotated as containing impact information or not. All the units of measurement and protein properties (ProteinProperty) existing in that sentence (impact) are recorded for subsequent ontology export.

**Mutation-Impact Relation Extraction.** When the entities such as *mutations* and *impacts* are identified and annotated, the sentence containing the impact word expressions (MutImpact) is associated with the nearest "Mutation," making the simple assumption that the nearest mutation invokes the impacts mentioned. The complete sentence is then considered as an impact sentence.

For each mutation-impact relation, we record the connection together with a number of properties, including units of measurement and effects.

**Ontology Population.** After preprocessing the documents and extracting the entities, the ontology is populated with the extracted entities such as *mutations*, *mutation impact* and their relations *mutation impact relations*.

## 5 Impact Summarization

The exported, populated OWL impact ontology can be queried using the SPARQL query language. To summarize impacts for a certain mutation, we can simply query the ontology for all detected impacts and extract the corresponding impact sentences:

```
PREFIX onto: <http://www.owl−ontologies.com/unnamed.owl#>
SELECT ?sentence
FROM <http://www.owl−ontologies.com/unnamed.owl#>
WHERE { ?document onto:containsSentence ?sentence.
                ?sentence onto:contains ?MutImpact.
                ?Mutation  onto:mutationMutImpactRel ?MutImpact}
ORDER BY DESC (?document) DESC (?Mutation)
```

These are then collected into a textual summary providing the mutations with their impacts for the user, as shown in Fig. 1.

## 6 Evaluation

The performance of the system was evaluated on the abstracts of four different mutation corpora, each on a specific protein family: *Xylanase* (19 documents), *Haloalkane Dehalogenase* (23 documents), *Subtilisin* (5 documents), and *Dioxygenase* (11 documents). Altogether, 58 documents were manually annotated with their impacts. For each annotation "Sentence," a binary feature "impact" is considered. As long as an impact exists in the sentence, the feature "impact" is set to "Yes;" otherwise to "No." The results are shown in the Table below; here, #C, #P, #M, and #S correspond to the correct, partially correct, missing, and spurious impact sentences, respectively; and $P$, $R$, $F$ are the precision, recall, and F-measure:

| **PMID 10860737** | |
|---|---|
| Mutation | Impacts |
| N35D | As predicted from sequence comparisons, substitution of this asparagine residue with an aspartic acid residue (N35D BCX) shifts its pH optimum from 5.7 to 4.6, with an 20 % increase in activity.. . . |
| **PMID 8855954** | |
| Mutation | Impacts |
| E123A | Mutation of a third conserved active site carboxylic acid (E123A) resulted in rate reductions of up to 1500-fold on poorer substrates,... |
| E127A | Elimination of the acid/base catalyst (E127A) yields a mutant for which the deglycosylation step is slowed some 200-300-fold as a consequence of removal of general base catalysis, but with little effect on the transition state structure at the anomeric center. Effects on the glycosylation step due to removal of the acid catalyst depend on the aglycon leaving group ability, with minimal effects on substrates requiring no general acid catalysis but large (>105-fold) effects on substrates with poor leaving groups... |
| . . . | . . . |

Figure 1: Impact Summaries (Excerpts)

| **Impact detection evaluation results on four corpora** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Corpus** | **#C** | **#P** | **#M** | **#S** | ***P*** | ***R*** | ***F*** |
| Haloalkane D. | 171 | 2 | 24 | 22 | 0.882 | 0.873 | 0.877 |
| Xylanase | 140 | 2 | 19 | 17 | 0.886 | 0.875 | 0.881 |
| Dioxygenase | 77 | 0 | 13 | 14 | 0.846 | 0.855 | 0.850 |
| Subtilisin | 32 | 2 | 9 | 10 | 0.750 | 0.767 | 0.758 |

The evaluation of associating the mutations with their impacts has so far been performed on the "Xylanase" corpus:

| | Precision | Recall | F-Measure |
|---|---|---|---|
| Lenient (Partial matches included) | 88% | 80% | 91% |
| Average (of Lenient and Strict) | 86% | 76% | 80% |
| Strict (Partial matches not counted) | 51.8% | 46.6% | 49.06% |

## 7 Discussion

Our Mutation Impact Ontology models mutation impacts in the biomedical domain, linking them to the texts where they are found. Although the detection of mutation impacts has shown to be successful by this simple proximity heuristic to some extent, in some cases the impacts are missing or detected partially. Also, in cases where the impacts caused by a set of mutations, just one mutation (the nearest one) is considered, and the remaining mutations are ignored. Impacts are not always the result of the nearest mutation; However, automatically analysing the text and specifying the correct mutation associated with the impacts needs more complex analysis.

## References


T. Kawabata, M. Ota, and K. Nishikawa. 1999. The Protein Mutant Database. *Nucleic Acids Res*, 27(1):355–357.

S. Yeniterzi and U. Sezerman. 2009. Enzyminer: automatic identification of protein level mutations and their impact on target enzymes from pubmed abstracts. *BMC Bioinformatics*, 10(Suppl 8):S2.