BioNLP 2010 Poster

Ontology-Based Extraction and Summarization of Protein Mutation Impact Information

Nona Naderi and René Witte
Semantic Software Lab, http://www.semanticsoftware.info
Department of Computer Science and Software Engineering, Concordia University, Montréal

Concordia University

1. INTRODUCTION

Our *Mutation Impact Miner* extracts mutation impacts from full-text papers and populates an OWL-DL ontology with impact information for querying and summarization.

Impacts are described through other concepts, in particular, changes to physical quantities such as *pH* and *temperature*:

Mutations — H86A/E/F/K/Q/W decreased the enzyme stability at 60°C by up to 80% and the transition temperature by 2.9°C to 5.8°C.
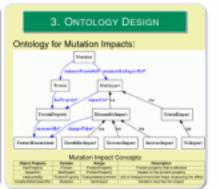Protein property — Unit of measurement — Impact word

2. MUTATION-IMPACT RELATIONS

Proximity heuristic to associate mutations with their impacts. For example:

H86A/E/F/K/Q, reduced the specific activity by 32-74%; whereas mutation H86W increased it by 18%.

The impact, *increased*, is caused by "H86W", which is nearer to the impact word than "H86A/E/F/K/Q". And in the same way, *reduced* is associated with "H86A/E/F/K/Q" rather than "H86W".

3. ONTOLOGY DESIGN

Ontology for Mutation Impacts:

Mutation Impact Concepts

5. DESIGN & IMPLEMENTATION

Mutation impacts are detected by an NLP pipeline developed based on the General Architecture for Text Engineering (GATE) [1]:

Preprocessing: Tokenizing and sentence splitting

Entity recognition: Ontogazetteering and JAPE grammars to capture *impact words*, *units of measurement* and *protein properties*

Relation extraction: Using rules and grammars mutations are associated with their impacts and sentences containing impacts are specified

Ontology population: Populating the OWL-DL ontology with the detected entities and relations

6. EVALUATION

The performance of the system was evaluated on the abstracts of four different mutation corpora. Altogether, 58 documents were manually annotated with their impacts. 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:
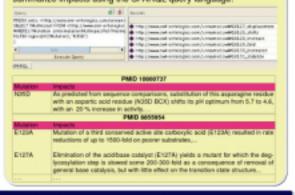
Impact detection evaluation results on four corpora

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 on the "Xylanase" corpus:

Relation detection evaluation results

	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%

6. APPLICATION

The OWL-DL ontology is populated with the entities and their relations to create a knowledge base, which can then be queried for a certain protein, mutation or protein property. For example, to summarize impacts using the SPARQL query language:

PMID 10660737

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 9655954

Mutation	Impacts
E123A	Mutation of a third conserved active site carboxylic acid (E123A) resulted in rate reductions of up to 1580-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...

References

[1] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: an Architecture for Development of Robust HLT Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, 2002. http://gate.ac.uk/.

Source URL (retrieved on *2025-12-23 11:26*): https://www.semanticsoftware.info/image/bionlp-2010-poster