First Release of the Open Mutation Miner (OMM) System
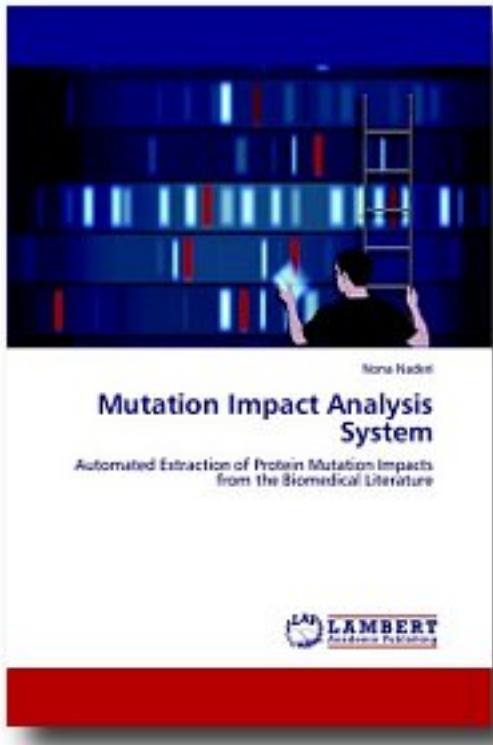
- Corpora
- GATE Components
- Open Mutation Miner
- OrganismTagger
- Bioinformatics
- Text Mining

We are happy to announce the first major public release of our protein mutation impact analysis system, *Open Mutation Miner* (OMM), together with a new open access publication [1]: Naderi, N., and R. Witte, "Automated extraction and semantic analysis of mutation impacts from the biomedical literature", *BMC Genomics*, vol. 13, no. Suppl 4, pp. S10, 06/2012.

OMM is the first comprehensive, fully open source system for extracting and analysing mutation-related information from full-text research papers. Novel features not available in other systems include: the detection of various forms of mutation mentions, in particular mutation series, full mutation impact analysis, including linking impacts with the causative mutation and the affected protein properties, such as molecular functions, kinetic constants, kinetic values, units of measurements, and physical quantities. OMM provides output options in various formats, including populating an OWL ontology, Web service access, structured queries, and interactive use embedded in desktop clients. OMM is robust and scalable: we processed the entire PubMed Open Access Subset (nearly half a million full-text papers) on a standard desktop PC, and larger document sets can be easily processed and indexed on appropriate hardware.

The current OMM release (http://www.semanticsoftware.info/open-mutation-miner) can be installed directly from within GATE Developer through the Plugin Manager and contains the following open source components and resources:

I. Text Mining

Open Mutation Miner (OMM) is a component-based system that integrates multiple sub-systems, all based on the General Architecture for Text Engineering (GATE):

OMM Mutations detects and annotates mutation mentions, including SNPs and mutation series, such as double-/triple-mutants. Basic mutations are detected through MutationTagger, which is implemented with JAPE-Plus. Additionally, the MutationFinder system can be integrated.

OMM Organisms provides for the recognition, normalization, and grounding of species names, by integrating our high-performance OrganismTagger.

OMM Impacts detects impact information of mutations, annotates them, and links the impacts to their originating mutation. It currently contains the sub-components:

- **ImpactTagger:** detection of impact mentions
- **ProteinPropertiesTagger:** detection of protein properties, including molecular functions, kinetic constants, kinetic values, units of measurements, and physical quantities
- **ImpactGrounding:** detection of the relation between impacts and their originating mutation

OMM Impacts integrates the Gene Ontology (GO) for the analysis of protein properties. Detected molecular functions are annotated with their GO ID.

II. Resources

In addition to the resources included with the text mining subsystems, such as RDF triples and gazetteering lists, we developed a number of general-purpose resources for the domain of mutation analysis:

OMM Ontology: The ontology is implemented in the standard Web Ontology Language (OWL) format. It formally models concepts and their relations in the mutation analysis domain, providing semantic knowledge during the text mining process. For example, the restrictions placed on possible slot fillers for a specific protein property allows to filter out erroneous value assignments during the analysis process, thereby improving overall precision. It can be automatically populated through the OwlExporter output option.

OMM Corpora: Two manually annotated corpora, containing full-text articles, are used for evaluation: (i) a corpus of 11 documents on protein engineering for evaluating the extraction of mutation series. (ii) a corpus of biomedical documents on enzymes for impact extraction and grounding evaluations.

III. System Output

Open Mutation Miner is based on a flexible architecture that facilitates multiple use cases, including batch-processing of a large number of publications in order to build a knowledge base or interactive use for supporting literature analysis. At present, the following output formats are supported (in addition to plain XML result export).

OMM Ontology Export: Results of the analysis can be exported into the OMM Ontology (ontology population) through the OwlExporter component. The resulting ontology can then be queried to facilitate knowledge discovery. Note that exported entities such as mutations, impacts, protein properties, or physical quantities are all linked to the originating paper from which they were extracted, which enables query-based literature navigation.

OMM Impact Summarization: Automatic summarization of a single publication or a corpus of papers, is a particular application

example of the ontology export functionality in OMM: Using a custom SPARQL query, all detected mutations can be extracted for a publication and displayed in various formats.

IV. Services

For accessing OMM results and integration into existing web information system infrastructure, a number of services are included.

OMM Web Service: Both the mutation tagging and impact analysis pipelines can be deployed as standard W3C web services using the Semantic Assistants server. This allows to remotely execute one of the analysis pipelines through standard SOAP or REST web service request on a single or set of documents, and receive the results in XML format.

OMM Query: The structured information extracted by OMM can be queried together with the unstructured natural language text in a unified query language based on GATE Mímir, allowing you to ask queries like *"Show me all mutations that increased the thermostability of a protein."* Indexed documents can be queried through a web interface or through a RESTful web service. For an example, using documents from the PMC Open Access Corpus, please visit our online OMM Query page at http://www.semanticsoftware.info/omm-query

OMM Semantic Assistants: The Semantic Assistants architecture also includes a number of clients that allow to consume the web services with a number of standard desktop clients. For details on currently supported clients, please visit the Semantic Assistants page at http://www.semanticsoftware.info/semantic-assistants.

The Open Mutation Miner (OMM) system, including its components and resources, is published under the GNU Affero General Public License v3 (AGPL3).

Additional information on OMM is available in the thesis [2], which is also available as a printed book [3].

References

1. Naderi, N., and R. Witte, "Automated extraction and semantic analysis of mutation impacts from the biomedical literature", *BMC Genomics*, vol. 13, no. Suppl 4, pp. S10, 06/2012.
2. Naderi, N., "Automated Extraction of Protein Mutation Impacts from the Biomedical Literature", *Department of Computer Science and Software Engineering*, M. Comp. Sc., Montreal : Concordia University, 09/2011.
3. Naderi, N., Mutation Impact Analysis System: Automated Extraction of Protein Mutation Impacts from the Biomedical Literature, LAP LAMBERT Academic Publishing, pp. 1-224, 2012.

Source URL (retrieved on *2026-01-16 05:30*): https://www.semanticsoftware.info/first-open-mutation-miner-release