Benchmarking infrastructure for mutation text mining

Artjom Klein*[1], Alexandre Riazanov[1] , Matthew M Hindle[2] and Christopher JO Baker[1]

[1]Computational Statistics And Science Department, University of New Brunswick,Saint John, Canada
[2] Synthetic and Systems Biology, Edinburgh University, Edinburgh, UK

Email: Artjom Klein*- aklein@unb.ca; Alexandre Riazanov - alexr@unb.ca; Matthew M Hindle -matthew.hindle@ed.ac.uk; Christopher JO Baker - bakerc@unb.ca;

*Corresponding author

Abstract

Background: Research work on the automatic extraction of information about mutations from texts is greatly hindered by the lack of consensus evaluation facilities and easy-to-use infrastructure for testing and benchmarking of mutation text mining systems.

Results: We propose a community-oriented annotation and benchmarking infrastructure to support development, testing, benchmarking, and comparison of mutation text mining systems. The design is based on semantic standards, where RDF is used to represent the annotations, an OWL ontology provides an extensible schema for the data and SPARQL is used to compute various performance metrics, so that in many cases programming is not needed to analyze system results. While large benchmark corpora for biological entity and relation extraction are focused mostly on gene, proteins, diseases, and species, our benchmarking infrastructure fills the gap for mutation information. The core infrastructure comprises of: 1) an ontology for modelling annotations, 2) SPARQL queries for performance metrics computation, and 3) a sizeable collection of manually curated documents, that can minimally support mutation grounding and mutation impact extraction.

Conclusion: This is the first example of benchmarking infrastructure for mutation text mining. It is designed for community uptake.

Introduction

Mutation text mining. The use of knowledge derived from text mining for mentions of mutations and their consequences is increasingly important for systems biology, genomics and genotype-phenotype studies. Mutation text mining facilitates a wide range of activities in multiple scenarios including the modelling of cell signalling pathways [1], protein structure annotation [2, 3] , the expansion of disease-mutation database annotations [4] and the development of tools predicting the impacts of mutations [5, 6]. The types of useful text mining tasks specific to mutations range from the relatively simple identification of mutation mentions [7], to very complex tasks such as linking (*"grounding"*) identified mutations to the corresponding genes and proteins [8], or identifying mutation impacts [9, 10] and related phenotypes [11].

Benchmarking and evaluation difficulties. Although the demand for mutation text mining software has lead to a significant growth of the experimental research in this area, the development of such systems and the publication of results is greatly hindered by the lack of adequate benchmarking facilities. In the first place, developers of mutation text mining systems need input data – texts annotated with target information – to simply test new versions of their implementations. They further need facilities to benchmark different versions of their systems in order to monitor development progress. Finally, the developers need to be able to convincingly evaluate their systems performance by comparing their results with extensive gold standard data and results of other systems.

Ideally, there should be community-based consensus corpora and utilities to make such benchmarking and evaluation easy. However, such facilities currently do not exist and developers are forced to spend time and effort on creating *ad hoc* corpora and scripts. As a result, the time required for benchmarking in the total development work is disproportionally high. Moreover, since only relatively small corpora are affordable to many research groups, the quality of evaluation suffers too. In developing a mutation grounding system [8] showing an encouraging level of performance accuracy, 0.73, on a homogeneous corpus of 76 documents, the authors achieved only 0.13 on a heterogeneous corpus of larger size. When the system was reimplemented (see, [12]), the authors encountered another challenge – the evaluation of the new system by comparing it to the state-of-the-art was practically unaffordable, despite the existence of similar systems, due to the lack of consensus benchmarking infrastructure. The lack of adequate community-based benchmarking infrastructure is a great hindrance to progress in the area of mutation text mining. We propose to improve this situation by developing a publicly accessible infrastructure.

Requirements. To guide our work, we impose the following requirements on the infrastructure to be created:

- To maximize its utility for system testing and evaluation, the infrastructure must include as big a gold standard corpus (a collection of annotated texts) as possible. It must also contain results of the runs of different systems to facilitate comparison of their performance.
- To be useful to a larger community, the infrastructure should support multiple mutation-related text mining tasks, such as identifying mutations both on DNA and protein levels, mutation grounding to gene and proteins, identifying effects of mutations, etc.
- The infrastructure must be easy to use requiring only minimal effort from system developers. Ideally, many development tasks should be facilitated so that the developers do not need to create new data formats or write additional scripts in order to leverage the infrastructure.
- The infrastructure should not only be publicly available but also support a sufficiently straightforward submission of both gold standard annotations and system results by the mutation text mining community.

Content overview. In this paper we report the results on the design and implementation of a community-oriented annotation and benchmarking infrastructure to support development, testing, benchmarking, and comparison of systems for mining information about mutations. The paper is outlined as follows. The Methods section starts by describing the motivation for the choice of representation format, it continues to outline a specification of ontology for modelling annotations and describes a method to calculate evaluation metrics. The Results and Discussion section presents details of seed corpora, methods for calculation of performance metrics, and utilities supporting benchmarking infrastructure. At the end of this section, we outline a testing infrastructure use-case and future work. Finally we provide a Conclusion summarizing results and highlight the availability of the benchmarking infrastructure.

Methods

Representing gold standard annotations and system results in RDF

Typically, document annotations intended for text mining system testing and evaluation, are represented in various custom XML-based or tabular formats. XML is a standard and widely used format for corpora annotation which comes with a large number of tools. Nevertheless, the processing of complex annotations in XML – parsing, storing, querying, evaluation – is usually practically impossible with off-the-shelf XML tools [13]. Developers need to create schema-specific parsers and processing scripts and change them each time the schema is changed or extended. This was the primary reason we chose RDF over custom

XML-based formats, because the reusability and extensibility of data are among the design goals of RDF. We also use OWL ontologies as highly extensible data schemas. An existing example of the successful use of RDF with OWL for representing biological data is the BIOPAX [14] format for representing biological pathway data.

The advantages of using the RDF/OWL bundle can be summarized as follows:

- **Extensibility.** Since the benchmarking infrastructure is going to be used for different mutation text mining tasks and all requirements can not be foreseen, we need extensible representations. Moreover, the same data may be used for different tasks (e. g., we have reused mutation impact corpora for improving mutation grounding system [12]).

  The use of RDF data with classes and properties defined in OWL ontologies makes it possible to support easy integration of new corpora with annotation schemas that need not be identical, as long as they are compatible. This simply amounts to *using compatible OWL ontologies and modelling patterns for RDF*. Data defined modulo one ontology can be *simply merged* with data modulo another ontology. Moreover, additional alignments between the ontologies can be provided by the annotation providers – corpus curators or text mining system developers.

- **Tool availability.** RDF and OWL are popular open formats and supported by a large number of open source and commercial tools. The following types of tools can be leveraged for the purpose of text mining annotation processing:

  - OWL reasoners can be used for data integrity checking.
  - RDF and OWL APIs for multiple programming languages, including Java, C++, Perl and Python, facilitate easy programmatic generation and manipulation of annotations or RDF data representing text mining results.
  - The SPARQL query language can be directly used for calculating system performance metrics as well as for various searches in the gold standard corpora. There is no need to implement custom querying mechanisms.
  - Multiple implementations of RDF databases (*triplestores*) are available that facilitate efficient storing and querying of large volumes of annotations.

  The diversity of available RDF tools enables out-of-the-box use of the annotation data in the main use scenarios, such as system testing and evaluation.

Core Ontologies and Modelling

Ontologies

The Mutation Impact Extraction Ontology (MIEO) [15] is central to our infrastructure. It currently describes classes and properties necessary to represent information about mutations at protein level, identified in texts, and extracted mutation impacts on molecular functions. For example, `AminoAcidSequenceChange` is the class for mutations at protein level. Instances of `ProteinVariant` are most specific types of protein molecules that completely identify the corresponding amino acid sequences. Instances of `ProteinPropertyChange` are the identified changes of protein properties that can be linked to: the properties that change, the corresponding documents and specific text fragments, and the mutations they result from. To characterize a property change, e. g., as positive, which may correspond to increased activity, we can use the subclass `PositiveProteinPropertyChange`. Protein properties, such as molecular functions, are also modelled as individuals whose types are currently taken from the Gene Ontology [16]. Note that some of our target mutation tasks are related to the extraction of relations between entities rather than just identifying some entities of interest. We use custom reification for such relations, in particular to facilitate linking them to documents and more specific provenance information. For example, extracted statements of mutations impacting protein properties are represented as instance of the class `StatementOfMutationEffect`.

Note that our MIEO uses the Semanticscience Integrated Ontology (SIO) [17] as an upper ontology, and the LSRN ontology [18] to represent records and identifiers, as illustrated in the next section.

Modelling example.

We provide an RDF graph in pseudo-N3 as an example of how the gold standard corpus data and results of mutation impact text mining are represented in our infrastructure. Note that non-mnemonic ontological identifiers are replaced with pseudo-identifiers using the corresponding labels: e. g., `sio:SIO_000011` and `sio:SIO_000300` are replaced respectively with `sio:'has attribute'` and `sio:'has value'`.

```
# Description of a singular amino acid substitution N30A:
:singular_mutation1 rdf:type mieo:AminoAcidSubstitution .
:singular_mutation1 mieo:mutationHasWildtypeResidue mieo:Asparagine .
:singular_mutation1 mieo:mutationHasMutantResidue mieo:Alanine .
:singular_mutation1 mieo:mutationHasPosition :position1 .
:position1 rdf:type sio:'position' .
:position1 sio:'has value' "30"^^xsd:integer .

# Description of a singular amino acid substitution N50A:
:singular_mutation2 rdf:type mieo:AminoAcidSubstitution .
:singular_mutation2 mieo:mutationHasWildtypeResidue mieo:Asparagine .
:singular_mutation2 mieo:mutationHasMutanResidue mieo:Alanine .
:singular_mutation2 mieo:mutationHasPosition :position2 .
:position2 rdf:type sio:'position' .
:position2 sio:'has value' "50"^^xsd:integer .
```

```
# Combined mutation ("mutation series") consisting of the two singular mutations:
:mutation rdf:type mieo:CombinedAminoAcidChange .
:mutation sio:'has member' :singular_mutation1 .
:mutation sio:'has member' :singular_mutation2 .
:mutation sio:'has attribute' :number_of_singular_mutations .
:number_of_singular_mutations rdf:type sio:'count'
:number_of_singular_mutations sio:'has value' "2"^^xsd:integer .

# Mutation application ("grounding") to a specific protein:
:mutation_application rdf:type mieo:ProteinMutationApplication .
:mutation_application mieo:isApplicationOfMutation :mutation .
:mutation_application mieo:isApplicationOfMutationToProtein :protein .

# Description of the protein:
:protein rdf:type mieo:ProteinVariant . # it's a specific variant (uniquely identifies the sequence)
:protein mieo:proteinHasSequence :protein_sequence .
:protein sio:'is subject of' :uniprot_record .

# Standard SIO way to link entities, DB records and IDs:
:uniprot_record rdf:type lsrn:UniProt_Record .
:uniprot_record sio:'has attribute' :uniprot_record_id .
:uniprot_record_id rdf:type lsrn:UniProt_Identifier .
:uniprot_record_id sio:'has value' "P22635" .

# Provenance is mostly done with sio:'refers to' :
:document rdf:type sio:'article' .
:document sio:'refers to' :singular_mutation1 .
:document sio:'refers to' :singular_mutation2 .
:document sio:'refers to' :mutation .
:document sio:'refers to' :mutation_application .
:document sio:'refers to' :protein .

:document sio:'has unique identifier' :document_identifier .
:document_identifier rdf:type mieo:PubMedURI . # subclass of mieo:URI
:document_identifier sio:'has value' "http://www.ncbi.nlm.nih.gov/pubmed/17526795"^^xsd:anyURI .
```

Note that, for simplicity, RDF data in this example are in "flat" RDF. In practice this is not convenient because we need to somehow separate the gold standard data from system results. Moreover, it is necessary to separate results coming from different systems or different experiments. We use *named graphs* [19] for this purpose: results from different experiments, and even gold standard data from different corpora, are placed in different named graphs.

Benchmarking with SPARQL

An infrastructure intended for benchmarking and evaluation must support the computation of performance metrics, such as precision and recall. Note that different flavours of these statistics are used by system developers: e. g., [20] proposes over 15 different metrics to evaluation protein mutation extraction systems. Moreover, text mining results sometimes need to be evaluated with different granularities, e. g., the mutant protein property change may be evaluated by considering binary outcomes (*has effect* vs *no effect*) or with higher granularity when the outcome may also identify the direction of the effect – e. g., *positive effect* or *negative effect*.

Our infrastructure has to be sufficiently flexible to accommodate many such uses. This is achieved by using SPARQL to retrieve entities, such as different flavours of true and false positives, that need to be counted

in order to calculate a particular metric. The current version of SPARQL (1.1) offers a sufficient degree of flexibility. In particular, the *negation-as-failure* related features – `FILTER NOT EXISTS` and `MINUS` – allow, e. g., for easy qualification of some results as *false positives* by checking whether they are absent from the gold standard data.

Design of the seed corpora

To facilitate a preliminary evaluation of our infrastructure, we seeded it with several corpora supporting at least two mutation text mining tasks: mutation grounding to proteins and extraction of mutation impacts on molecular functions of proteins.

The document annotations for mutation grounding identify extracted mutations and proteins, and relations between them. The annotations for mutation impact extraction additionally identify molecular function of proteins and changes of these properties causally linked to some mutations, and provide references to supporting text fragments.

Results and Discussion

Contents of the corpora

EnzyMiner-based corpus.

One of our seed corpora is based on an extract from the EnzyMiner [21] abstract database. It was annotated manually and comprises 38 semi-randomly selected full text documents with 176 different singular mutations linked to 48 different protein sequences. The selection was adjusted to ensure maximal diversity by having documents with proteins from all enzyme families and 24 different species. The corpus currently contains 488 statements (occurrences of impact information in text), 61 molecular functions and 29 combined mutations.

In what follows, we call it simply "the EnzyMiner corpus".

We annotated documents with mutation impact information which includes:

- Studied protein-level **mutations**, in the form of singular amino acid substitutions. They are represented as triples specifying the wild type and mutant residues, and the absolute positions of the mutations on the corresponding amino acid sequences. For situations when the effects of several simultaneous amino acid substitutions are studied, we allow them to be expressed as *combined mutations*.

- **Proteins** to which the mutations are related, identified with UniProt IDs. The host organisms and sets of specific protein sequences can be identified via the UniProt IDs.
- **Protein properties** specified as Gene Ontology Molecular function classes.
- **Mutation impacts** qualified as *Positive*, *Negative* or *Neutral*.
- **Text fragments** the information was extracted from. Typical fragments contain mentions of protein properties, impact directionality words, such as "increased" or "worse", mutation mentions, protein and organism names, etc.
- **Documents** identified with PubMed IDs.

DHLA corpus.

This is a small corpus comprising 13 documents with 52 unique per document mutations on Haloalkane Dehalogenases, manually annotated similarly to the EnzyMiner documents (see [2]).

COSMIC-based corpus.

We have an extract from the COSMIC database [22] containing 63 documents for three target genes: FGFR3, MEN1 and PIK3CA. Unlike the EnzyMiner and DHLA corpora, this corpus does not identify mutation impacts, although it links mutations to proteins and, thus, is suitable for mutation grounding benchmarking.

KinMutBase-based corpus.

We retrieved 201 documents annotated with singular amino acid substitutions grounded to proteins, from the KinMutBase [23] database. We additionally curated the selection by running MutationFinder, which is a reliable tool for this purpose due to its very high recall, and comparing the results with the annotations in the database. Based on this comparison, we discarded about 70 documents that appear annotated with protein-level mutations that they don't seem to mention directly, although this may be due to the translation from SNPs made by the curators. The final size of the corpus is 128 documents. In total, we have 271 mutations linked to 26 different UniProt identifiers.

Corpora statistics.

The statistics for the corpora are summarized in Table 1.

	Corpus size	UniProt IDs	Mutations (unique per document)
EnzyMiner	38	49	176
KinMutBase	128	26	271
DHLA	13	4	52
PIK3CA	30	1	169
FGFR3	26	1	174
MEN1	7	1	22

Table 1: Corpus Statistics.

RDF database

The RDF files representing our corpora are already relatively large, so for the purposes of efficient SPARQL querying we deploy the data to a Sesame triplestore. Users have the option of downloading the RDF data and using their own querying machinery, or accessing our DB via a public SPARQL endpoint. The details can be found in [24].

SPARQL queries for performance metrics

To test the idea of using SPARQL for performance metrics computation, we have formulated several SPARQL queries sufficient for computing precision and recall for systems implementing two text mining tasks: mutation grounding to proteins and the extraction of impacts of mutations on protein properties. For each task we wrote (1) a SPARQL query that selects relevant annotations in the gold standard data, representing correct cases, (2) a query that selects all relevant results of the text mining system being evaluated, and (3) a query that selects only correct results. These selections are enough to calculate precision and recall.

We illustrate this by presenting a slightly simplified version of the query used to select the correct results from mutation-impact extraction, which can be used for evaluation according to the metric definitions from, e. g., [2, 10]. According to the definitions, a result is a set – *document*, *protein*, *mutation*, *protein property* changed by the mutation, and a *direction* of the property change. If the gold standard data contain the same set, the result is considered *correct*. Technically we have to compare two RDF graphs and get the corresponding intersection. Note that the query assumes that the gold standard data is kept in the named graph `http://example.com/gold-standard.rdf` and the system results come from another named graph `http://example.com/experiment.rdf`.

Note that as in modelling example for readability we replace non-mnemonic SIO identifiers with their labels.

```
PREFIX sio:<http://semanticscience.org/resource/>
PREFIX lsrn:<http://purl.oclc.org/SADI/LSRN/>
SELECT DISTINCT ?pubmed_id ?wt_residue ?position_value ?mut_residue ?uniprot_record_id ?
  property_change_class ?protein_property_class
WHERE {
GRAPH <http://example.com/gold-standard.rdf> {
      ?document a sio:'article' .
      ?document sio:'is subject of' ?pubmed_record .
      ?pubmed_record sio:'has attribute' ?pubmed_identifier .
      ?pubmed_identifier sio:'has value' ?pubmed_id .

      ?document sio:'refers to' ?mutation_application .
      ?mutation_application a mieo:ProteinMutationApplication .
      ?mutation_application mieo:isApplicationOfMutation ?mutation .
      ?mutation_application mieo:isApplicationOfMutationToProtein ?protein .
      ?document sio:'refers to' ?mutation .
      ?mutation a mieo:CombinedAminoAcidSequenceChange .
      ?mutation sio:'has member' ?singular_mutation .
      ?singular_mutation mieo:mutationHasWildtypeResidue ?wt_residue .
      ?singular_mutation mieo:mutationHasMutantResidue ?mut_residue .
      ?singular_mutation mieo:mutationHasPosition ?position .
      ?position a sio:'position' .
      ?position sio:'has value' ?position_value .

      ?document sio:'refers to' ?protein .
      ?protein sio:'is subject of' ?uniprot_record .
      ?uniprot_record a lsrn:UniProt_Record .
      ?uniprot_record sio:'has attribute' ?uniprot_record_identifier .
      ?uniprot_record_identifier a lsrn:UniProt_Identifier .
      ?uniprot_record_identifier sio:'has value' ?uniprot_record_id .

      ?document sio:'refers to' ?property_change .
      ?mutation_application mieo:mutationApplicationCausesChange ?property_change .
      ?property_change a ?property_change_class .

      ?document sio:'refers to' ?protein_property .
      ?property_change mieo:propertyChangeAppliesTo ?protein_property .
      ?protein_property sio:'is property of' ?protein .
      ?protein_property a ?protein_property_class .
}
GRAPH <http://example.com/experiment.rdf> {
      ?document2 a sio:'article' .
      ?document2 sio:'is subject of' ?pubmed_record2 .
      ?pubmed_record2 sio:'has attribute' ?pubmed_identifier2 .
      ?pubmed_identifier2 sio:'has value' ?pubmed_id .

      ?document2 sio:'refers to' ?mutation_application2 .
      ?mutation_application2 a mieo:ProteinMutationApplication .
      ?mutation_application2 mieo:isApplicationOfMutation ?mutation2 .
      ?mutation_application2 mieo:isApplicationOfMutationToProtein ?protein2 .
      ?document2 sio:'refers to' ?mutation2 .
      ?mutation2 a mieo:CombinedAminoAcidSequenceChange .
      ?mutation2 sio:'has member' ?singular_mutation2 .
      ?singular_mutation2 mieo:mutationHasWildtypeResidue ?wt_residue .
      ?singular_mutation2 mieo:mutationHasMutantResidue ?mut_residue .
      ?singular_mutation2 mieo:mutationHasPosition ?position2 .
      ?position2 a sio:'position' .
      ?position2 sio:'has value' ?position_value .

      ?document2 sio:'refers to' ?protein2 .
      ?protein2 sio:'is subject of' ?uniprot_record2 .
      ?uniprot_record2 a lsrn:UniProt_Record .
      ?uniprot_record2 sio:'has attribute' ?uniprot_record_identifier2 .
      ?uniprot_record_identifier2 a lsrn:UniProt_Identifier .
      ?uniprot_record_identifier2 sio:'has value' ?uniprot_record_id .

      ?document2 sio:'refers to' ?property_change2 .
      ?mutation_application2 mieo:mutationApplicationCausesChange ?property_change2 .
      ?property_change2 a ?property_change_class .

      ?document2 sio:'refers to' ?protein_property2 .
      ?property_change2 mieo:propertyChangeAppliesTo ?protein_property2 .
      ?protein_property2 sio:'is property of' ?protein2 .
      ?protein_property2 a ?protein_property_class .
}}
```

We comment briefly on the query composition, the two halves of the query (lines 5-35 and 36-66) correspond to the selection of relevant data from the gold standard corpora and from the experimental system results. Since our goal is to select only *correct* results, the two selections are joined on the instances of the variables `?pubmed_id` (identifying documents), `?wt_residue`, `?mut_residue` and `?position_value` (for the wildtype and mutant residues, and positions of the corresponding mutations), `?uniprot_record_id` (identifying proteins), `?protein_property_class` (identifying studied properties) and `?property_change_class` (identifying the direction of the property change).
Note that the query can only be used to implement *micro averaging* that treats the whole corpus as one large document. If, for some reason, we were interested in *macro averaging* we would have to additionally group the results by the PubMed ID values.

Utilities

As a part of our infrastructure, we created a small set of simple utilities, which facilitate data access:

- The *evaluator* utility calculates standard performance metrics by executing some user-provided SPARQL queries, counting the results and making necessary calculations. The user can supply the queries in a simple configuration file.
- The *Sesame loader* and *query client* are simple command line applications that allow loading RDF graphs into a Sesame triplestore and executing queries from files.
- The *provenance enhancement* utility helps in situations when the sources of annotation data only provide fragments of texts as provenance, without specifying their positions in the text. The utility simply searches the document texts for the corresponding fragments in order to provide more precise provenance information.

Testing the infrastructure

For concept validation, we have used our infrastructure for testing and iterative performance evaluation during a project dedicated to the development of a robust mutation impact extraction system [10], and for the evaluation of the mutation grounding subtask, intended for publication (see [12]). The purpose of the system is to identify protein-level mutations, ground them to the corresponding UniProt IDs and, most importantly, to extract the information about what properties of the proteins are affected and how, if this is described in the processed document.

Since early versions of the system already produced output in RDF modelled according to an ontology similar to the MIEO, it was trivial to adjust the system to produce output in a format compatible with our infrastructure. This was the major prerequisite, which was required to enable the evaluation of the system on our gold standard corpora and the subsequent comparison of results from different versions of the mutation grounding system.

Although the system previously showed reasonable performance on the 76 documents, the performance on the larger and more representative data set comprising the Enzyminer and KinMutBase corpora, was very low. After an investigation in which we relied heavily on the analysis of system runs based on our annotations, including the provenance information, we have identified the mutation grounding module as a major performance bottleneck having only 0.32 precision and 0.08 recall. We focused our attention on the mutation grounding subtask, in which our infrastructure was very instrumental as the task is also supported by existing gold standard annotations, and eventually improved the performance to 0.83 precision and 0.82 recall. More details on this effort can be found in [12].

Future work

In current work we are defining a procedure for the submission of third-party human-curated annotations and system results.

In the future work we will further stress-test the infrastructure with text mining tasks other than mutation grounding and mutation impact extraction, and a third-party mutation text mining system. We plan to extend the ontology based on the new requirements identified through community involvement and our own research. In the near future, we plan to extend the infrastructure to include protein properties other than molecular functions, such as enzyme kinetics, and DNA-level mutations.

The ontologies we are using provide only very basic means for attaching provenance information to identified entities and relations by simply linking them to the documents they were mined from. We are planning to use one of the existing ontologies for modelling sentence level provenance (Annotation Ontology [25], CALBC semantic annotation schema [26], or NLP Interchange Format [27]) to provide more precise pointers to text fragments supporting annotations.

Conclusions

We report preliminary results on the development of a community-oriented benchmarking infrastructure intended to relieve the developers of mutation text mining software from the burden of developing *ad hoc*

corpora and scripts for testing, benchmarking and evaluation of multiple mutation-related text mining tasks. While large benchmark corpora for biological entity and relation extraction (such as CALBC [28], BioCreative [29], GENIA [30], etc.) are focused mostly on gene, proteins, diseases, and species, our benchmarking infrastructure fills the gap for mutation information. We have seeded the infrastructure with a sizeable gold standard corpus (242 documents). To maximize the reusability and extensibility of our infrastructure, we use RDF and OWL for annotation data representation and SPARQL queries as a means of flexible analysis of text mining results. The infrastructure was tested for benchmarking and evaluation of a mutation impact extraction system.

We have done this work with the goal of *initiating a community effort*, and the future evolution of the benchmarking infrastructure will be based on feedback and contributions from the community.

Availability

The benchmark corpora, ontology for modelling annotations, example output of our mutation text mining system, benchmarking SPARQL query templates, and infrastructure support tools are available on the project page [24].

Acknowledgement

This research was funded in part by the New Brunswick Innovation Foundation, New Brunswick, Canada; the NSERC, Discovery Grant Program, Canada and the Quebec-New Brunswick University Co-operation in Advanced Education - Research Program, Government of New Brunswick, Canada.

References

1. Bauer-Mehren A, Furlong LI, Rautschka M, Sanz F: **From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways.** *BMC Bioinformatics* 2009, **10**(Suppl 8):S6.
2. Baker CJO, Witte R: **Mutation Mining—A Prospector's Tale**. *Information Systems Frontiers (ISF)* 2006, **8**:47–57.
3. Kanagasabai R, Choo KH, Ranganathan S, Baker CJO: **A workflow for mutation extraction and structure annotation.** *J Bioinform Comput Biol* 2007, **5**(6):1319–37.
4. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, Kann MG: **Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature.** *Bioinformatics* 2011, **27**(3):408–15.
5. Bromberg Y, Overton J, Vaisse C, Leibel RL, Rost B: **In silico mutagenesis: a case study of the melanocortin 4 receptor.** *FASEB J* 2009, **23**(9):3059–69.
6. Winnenburg R, Plake C, Schroeder M: **Improved mutation tagging with gene identifiers applied to membrane protein stability prediction.** *BMC Bioinformatics* 2009, **10**(Suppl 8):S3.

7. Caporaso JG, Jr WAB, Randolph DA, Cohen KB, Hunter L: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics* 2007, **23**(14):1862–1865.
8. Laurila JB, Kanagasabai R, Baker CJO: **Algorithm for grounding mutation mentions from text to protein sequences**. In *Proceedings of the 7th international conference on Data integration in the life sciences*, DILS'10, Berlin, Heidelberg: Springer-Verlag 2010:122–131.
9. Naderi N, Witte R: **Automated extraction and semantic analysis of mutation impacts from the biomedical literature**. *BMC Genomics* 2012, **13**(Suppl 4):S10.
10. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJO: **Algorithms and semantic infrastructure for mutation impact extraction and grounding.** *BMC Genomics* 2010, **11**(Suppl 4):S24.
11. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H: **Automatic extraction of mutations from Medline and cross-validation with OMIM.** *Nucleic Acids Res* 2004, **32**:135–42.
12. Klein A, Riazanov A, Al-Rababah K, Baker CJ: **Towards a next generation protein mutation grounding system for full texts**. *Accepted at SMBM2012*.
13. Eckart R: **Choosing an XML database for linguistically annotated corpora**. *Sprache und Datenverarbeitung* 2008, **32**:7–22.
14. Demir E, et al: **The BioPAX community standard for pathway data sharing**. *Nature Biotechnology* 2010, **28**(9):935–942.
15. **The Mutation Impact Extraction Ontology (MIEO)**. *http://unbsj.biordf.net/ontologies/mutation-impact-extraction-ontology.owl*.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.
17. **The Semanticscience Integrated Ontology (SIO)**. *http://semanticscience.org/ontology/sio.owl*.
18. **Life Science Record Name (LSRN)**. *http://lsrn.org*.
19. Carroll JJ, Bizer C, Hayes P, Stickler P: **Named graphs, provenance and trust**. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, New York, NY, USA: ACM 2005:613–622.
20. Witte R, Baker CJO: **Towards a systematic evaluation of protein mutation extraction systems.** *J Bioinform Comput Biol* 2007, **5**(6):1339–59.
21. Yeniterzi S, Sezerman U: **EnzyMiner: automatic identification of protein level mutations and their impact on target enzymes from PubMed abstracts.** *BMC Bioinformatics* 2009, **10**(Suppl 8):S2.
22. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A, Teague JW, Stratton MR, Futreal PA: **COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer.** *Nucleic Acids Res* 2010, **38**(Database issue):D652–7.
23. Ortutay C, Väliaho J, Stenberg K, Vihinen M: **KinMutBase: a registry of disease-causing mutations in protein kinase domains.** *Hum Mutat* 2005, **25**(5):435–42.
24. **Mutation text mining benchmarking infrastructure**. *http://code.google.com/p/mutation-text-mining*.
25. Ciccarese P, Ocana M, Garcia Castro LJ, Das S, Clark T: **An open annotation ontology for science on web 3.0.** *J Biomed Semantics* 2011, **2**:S4.
26. Croset S, Grabmüller C, Li C, Kavaliauskas S, Rebholz-Schuhmann D: **The CALBC RDF Triple Store: Retrieval over Large Literature Content**. In *SWAT4LS'10* 2010:–1–1.
27. Hellmann S, Lehmann J, Auer S: **NIF: An ontology-based and linked-data-aware NLP Interchange Format**[http://svn.aksw.org/papers/2012/WWW_NIF/public.pdf].
28. Rebholz-Schuhmann D, Yepes AJJ, Van Mulligen EM, Kang N, Kors J, Milward D, Corbett P, Buyko E, Beisswanger E, Hahn U: **CALBC silver standard corpus.** *J Bioinform Comput Biol* 2010, **8**:163–79.
29. **The BioCreAtIvE challenge evaluation**. *http://biocreative.sourceforge.net*.
30. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus–semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19**(Suppl 1):i180–2.