Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Database**

xGENIA: A comprehensive OWL ontology based on the GENIA corpus

Rafal Rak[1,*], Lukasz Kurgan[1], Marek Reformat[1]

[1]University of Alberta, ECERF, 9107 116 Street, Edmonton, AB, T6G 2V4 Canada;

Rafal Rak* - Email: rrak@ece.ualberta.ca; * Corresponding author

received February 28, 2007; accepted March 19, 2006; published online March 20, 2007

Abstract:

The GENIA ontology is a taxonomy that was developed as a result of manual annotation of a subset of MEDLINE, the GENIA corpus. Both the ontology and corpus have been used as a benchmark to test and develop biological information extraction tools. Recent work shows, however, that there is a demand for a more comprehensive ontology that would go along with the corpus. We propose a complete OWL ontology built on top of the GENIA ontology utilizing the GENIA corpus. The proposed ontology includes elements such as the original taxonomy of categories, biological entities as individuals, relations between individuals using verbs and verb nominalizations as object properties, and links to the UMLS® Metathesaurus concepts.

Keywords: biological ontology, biological entity extraction, OWL, MEDLINE

Availability: http://www.ece.ualberta.ca/~rrak/ontology/xGENIA/

Background:

The GENIA corpus consists of a set of 2000 annotated abstracts from MEDLINE database concerning "transcription factors in human blood cells". The corpus along with the corresponding taxonomy (ontology) was developed to provide a reference material for bio-textmining. **[1]** Since its development the GENIA corpus and ontology have been intensively used by researchers for biological entity recognition **[2]**, ontology creation and population **[3]**, and query processing. **[4]** However, recent work on biological name recognition and query processing **[4]** demonstrates a demand for a more comprehensive and complete ontology that would go along with the GENIA corpus. Other researchers **[5]** also suggested utilizing an ontology in the information extraction process, which is not feasible with a basic taxonomy only.

We propose xGENIA, an ontology that is based on the GENIA corpus and ontology created by. **[1]** This ontology, developed in OWL **[6]**, can be used as a golden standard and a knowledge base for biological information extraction.

Methodology:

The biological entities in the GENIA corpus were preprocessed before we added them to the xGENIA ontology as individuals. The first step of biological entity extraction involves decomposition of nested tags and terms involving ellipsis in coordinated clauses. The decomposed entities are further preprocessed with a set of manually developed rules, a common approach used in biological entity extraction. **[3, 4, 5]** We created our own set of rules putting special emphasis on the unification of entities carrying identical concepts yet being slightly different in form. Processing entities with the rules involves removing unnecessary white spaces, dividing words and word sequences into separate instances, and removing acronyms embedded in the sequence of words representing their full form.

In order to extract relations we used a set of verbs and verb nominalizations from. **[4]** To preserve generality we replaced inflectional variants of verbs and verb nominalizations with their canonical form (e.g., *activate*, *activates*, *activating*, and *activated* were replaced with *activate*). That way we reduced the original list of verbs and verb nominalizations from 246 to 142.

We manually assigned the *rdfs:subPropertyOf* element between verbs and verb nominalizations with prepositions and their canonical forms and between verb nominalizations and their root verbs as well as *owl:inverseOf* between two verbs of inverse meaning. The relations between the entities were found by searching for two entities appearing in the neighborhood on opposite sides of the verb in the same sentence. The sequence of words that includes the verb and is located between the subject and object entities must not be interrupted by a coma or a semicolon.

To properly identify UMLS® Metathesaurus CUIs, the extracted biological entities were normalized using *norm*, a tool provided by NLM **[7]**, which is used to create indices on the Metathesaurus database. The normalized entities were then compared against the Metathesaurus MRXNS ENG file, one of the Metathesaurus' indices, and, if found, CUIs were fetched and added to the ontology as the *hasCUI* datatype property.

Overview of the xGENIA ontology:

OWL integrates a taxonomy and instances (called individuals in OWL) of the taxonomy. xGENIA utilizes a variety of the OWL as well as RDF and RDFS (the languages OWL is based on) elements. They include classes (the GENIA's original taxonomy), individuals (biological entities), object properties (relations between the entities), datatype properties (unique identifiers), and others. The core of xGENIA consists of the original taxonomy of 47 categories as described in. **[1]** This taxonomy of categories is represented by classes (*owl:Class*) in our OWL ontology.

Biological entities:
The biological entities annotated in the GENIA corpus constitute individuals of categories they are annotated to. In order to keep the xGENIA ontology coherent the annotated biological entities have been preprocessed (see Methodology) to form unique entities carrying the same concepts regardless of lexical and syntactic differences in the way they were written by the authors. To satisfy constraints on the names of individuals imposed by OWL, each biological entity (individual) is assigned a unique identifier being a concatenation of the name of the class it is assigned to and a consecutive number. (Although OWL allows for assigning more than one class to an individual, this is not the case in the GENIA corpus.) The real name of an entity is represented by the *rdfs:label* property. Examples of individuals are shown in Figure 1(a).

Relations between biological entities:
xGENIA is also equipped with relations between individuals (represented by hexagons in Figure 1(a)). Each such binary relation binds two individuals through a verb or verb nominalization. These relations come from the corpus and have been extracted using a method described in Methodology). Each predicate (verb or verb nominalization) is represented in ontology as *owl:ObjectProperty* and has its own domain (*rdfs:domain*) and range (*rdfs:range*), i.e., the set of classes being a subject and object, respectively, occurring with the given predicate. To further enrich the ontology, the properties were put in a hierarchy using the *rdfs:subPropertyOf* element to indicate that one property is a variant of another. Additionally, the *owl:inverseOf* element denotes that two properties have the same meaning but a different underlying direction (see Figure 1(b)).

Lexical decomposition of biological entities:
Using nested tags in the GENIA corpus, where the inner tags are the lexical stems of the outer tags, xGENIA incorporates a special object property, *stemsFrom*, which indicates the direction from the outer tags to the inner tags. Following this property for a given entity leads to the lexical root(s) of that entity, which in combination with the *unique identifiers* (see the next section) of atomic entities (entities that do not instantiate the *stemsFrom* property) allow for identification of the original entity (see Figure 1(a)).

Concept Unique Identifiers:
To allow for easier identification, some of the annotated entities were assigned to *concept unique identifiers* (CUI) as found in UMLS® Metathesaurus. **[7]** The CUIs were obtained by preprocessing of the entities to their normalized form recognizable by the Metathesaurus database (see Methodology). The entities are linked to the CUIs through datatype property (*owl:DatatypeProperty*) *hasCUI*. The number of direct and indirect (through the *stemsFrom* property) links to the CUIs is shown in Table 1.

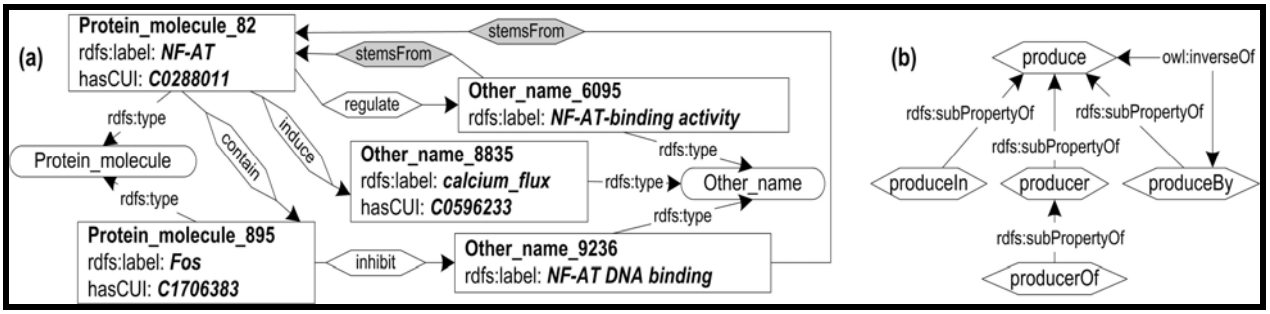

Figure 1: Fragment of xGENIA, (a) relations between biological entities, and (b) the taxonomy of verbs and verb nominalizations. Rectangles represent individuals (biological entities), hexagons represent relations, and ovals represent classes

Classes (original GENIA categories)	47
Object properties	142
Individuals (biological entities)	34,842
Relations between individuals	7,174
Lexical taxonomy (*stemsFrom*)	10,386
Individuals linked to CUI (directly / indirectly)	14,700 (6,851 / 7,849)

Table 1: xGENIA ontology statistics

Conclusion:
The xGENIA ontology together with the GENIA corpus provide a sophisticated benchmark for researchers who design and test applications in the field of biological information extraction. The overall statistics of xGENIA are presented in Table 1. The xGENIA ontology is an open project and we will continue improving it in the future. Each new release will be labelled by a unique version number and will be given the description of changes and additions.

References:
[01] J. D. Kim *et al.*, *Bioinformatics*, 19:180 (2003) [PMID:12855455]

[02] G. Zhou, *Int J Med Inform*, 75:456 (2006) [PMID: 16112894]
[03] E. SanJuan *et al.*, *Comp Speech & Lang*, 19:524 (2005)
[04] M. Abulaish & L. Dey,. *Data & Knowledge Eng*, doi:10.1016/j.datak.2006.06.007 (2006)
[05] N. Daraselia *et al*., *Bioinformatics*, 20:604 (2004) [PMID: 15033866]
[06] http://www.w3.org/TR/owl-features/
[07] http://www.nlm.nih.gov/research/umls/umlsdoc.html

Edited by P.Kangueane
Citation: R. Rak *et al.*, Bioinformation 1(9): 360-362 (2007)