Gene Regulation Ontology (GRO) task at BioNLP Shared Task 2013
Task Proposal

Jung-jae Kim[1] , Xu Han[1] and Dietrich Rebholz-Schuhmann[2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Institute of Computational Linguistics, University of Zurich, Switzerland

Task Summary

Task name Gene Regulation Ontology (GRO)-based Corpus Annotation

Goal: To evaluate and improve ontological corpus annotation methods for the application to the GRO.

Motivation in biology: To increase the number of ontology concepts and relations extracted from and annotated on biomedical text.

Description of the entities and events Entities: genes, proteins, transcription factors, chemical entities, sequence regions, cells, cellular components, species, and others related to gene regulation. Events: biological processes related to gene regulation. Many GRO concepts are from Gene Ontology (GO), Sequence Ontology (SO), ChEBI, INOH Molecular Role (IMR), and NCBI taxonomy [1].

Concrete examples in BioNLP format Available in attached data.

Evaluation and criteria: GRO entities and events need to be predicted, while the GRO entities will be provided for those who like to focus on events. Evaluation is mention-based and might follow the BioNLP ST'09 and '11 criteria.

Task corpus: PubMed abstracts selected by relevance to gene regulation at the transcription level in human. 200 abstracts are already annotated [2], and ongoing annotation effort aims for 100 more annotated documents, expected to involve around 15,000 entity and event annotations in total.

Existing corpus: The first release of the GRO corpus contains 200 PubMed abstracts and approximately 10,395 annotations of entities and events and will be used as the training data. Note that the training data do not contain any entity annotations that are not involved in event annotations, while the dataset provided to participants for the evaluation purpose will contain all possible entity annotations.

Feasibility of task: The GRO task is similar to the BioNLP ST'09 and the BioNLP ST'11 GENIA task in that it also aims at automatically annotating ontology concepts and relations of events on text. The main challenge in this task is the bigger size of the underlying ontology (i.e. GRO) than that of the GENIA ontology. More than 200 concepts and 10 relations from the GRO are used for our task, while the previous GENIA task involved 10 concepts and 6 relations.

Annotation process: The existing corpus was created by a PhD student with biology background and will be validated by a PhD-level biologist. The two will annotate the test data independently, while we will measure the inter-annotator agreement.

Tentative schedule for annotated corpus delivery: We plan to complete the validation of the training data and the annotation of the test data by the end of December 2012.

Contents

1. Introduction 1
2. Representation and Task Setting 1
3. Document Selection 2
4. Entity Annotation 2
5 Event annotation 4
    5.1 Annotation overview 4
    5.2 Event argument role 5
    5.3 Annotation process 5
6 Relation annotation 5
    6.1 Annotation overview 6

1. Introduction

This document presents the Gene Regulation Ontology (GRO) task proposed for the BioNLP Shared Task 2013.

Information seeking with search engines in the biomedical domain is getting more difficult because of the rapidly increasing volume of biomedical research publications. This difficulty partly results from the limitations of keyword search including the lack of dealing with semantic relations between keywords. To address it, we should move towards semantic search, adapting Semantic Web technologies to the biomedical literature. Semantic Web may be facilitated with formal semantic representation of the documents with ontology concepts and relations. The ontological annotation was initiated in the biomedical domain by the GENIA corpus (Kim *et al*., 2008), and the tasks of the BioNLP Shared Task 2011 aimed at automatically populating such ontological annotations. However, the tasks dealt only with a small number of ontology concepts (17 concepts in total), considering thousands of concepts defined in standard biomedical ontologies (e.g. Gene Ontology, anatomy ontologies). The goal of this task we propose is to confirm if text mining techniques can be scaled up to cover hundreds of concepts.

The task we propose here is to automatically annotate biomedical documents with the Gene Regulation Ontology (GRO)[1] [1]. GRO is a conceptual model of gene regulation and includes 507 concepts, which are cross-linked to such standard ontologies as Gene Ontology and Sequence Ontology. It has two top-level categories, Continuant and Occurrent, where the Occurrent branch has concepts for processes that are related to the regulation of gene expression (e.g. Transcription, Translation), and the Continuant branch has concepts for physical entities that are involved in those processes (e.g. Gene, TranscriptionFactor, Cell). It also has semantic relations that link the instances of the concepts (e.g. hasAgent, locatedIn). In short, the proposed task is to identify the semantics of text related to the gene regulation domain, where the elements of the textual semantics are the concepts and relations of GRO.

2. Representation and Task Setting

The BioNLP Shared Task (ST) 2013 Gene Regulation Ontology (GRO) task is an event extraction task following the representation and task setting of the ST'09 and ST'01 main tasks.

The representation involves three primary categories of annotation: the entity annotation (either physical entity or non-physical entity), event annotation and

[1] http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html

relation annotation. Mentions of entities in text are annotated with both contiguous and discontinuous spans that are assigned specific entity concepts (e.g. Gene, Protein, CellularComponent). References to biological processes are annotated using the event representation, in which the event annotation is associated with the specific contiguous span in text (called event trigger) that explicitly suggests the biological process. In addition, such a span is assigned with the proper event type that is pre-defined in ontology. If the participants (entities or other events) of the event can be explicitly identified with specific spans in text, the relation between the participant and the event trigger is also annotated with the proper relation type (e.g. hasAgent, hasPatient). The relation annotation is to annotate relations between entities (e.g. hasPart, fromSpecies) though without event triggers. Figure 1 illustrates some of the annotations.

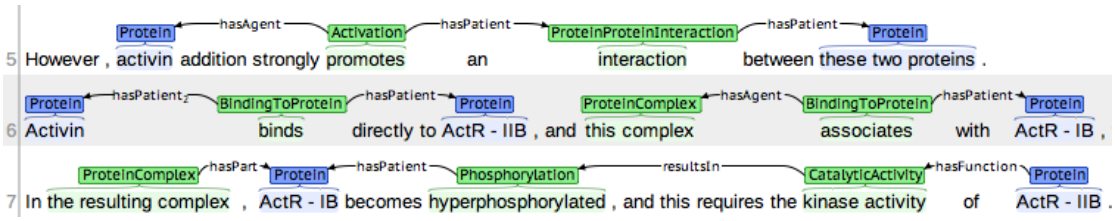

Figure 1. Example annotations for the GRO task

3. Document Selection

The corpus texts are selected based on the relevance to gene regulation in human. Specifically, we first obtained a list of human transcription factors (TFs) and then used PubMed to collect a set of candidate documents. The GRO task documents are then selected as a random subset of 200 documents from the collection. We will further select 100 documents and annotate them for test purposes.

4. Entity Annotation

This section introduces the entity annotation applied in the GRO task.

The entity annotation marks mentions of entity types defined under the Continuant branch of the GRO. An entity mention annotation is identified by its start and end offsets (contiguous span) in text or by such two pairs (discontinuous span) and is assigned an entity type from the GRO. Figure 2 depicts the top levels of the Continuant branch of the GRO. We used the BRAT for the annotation [3].

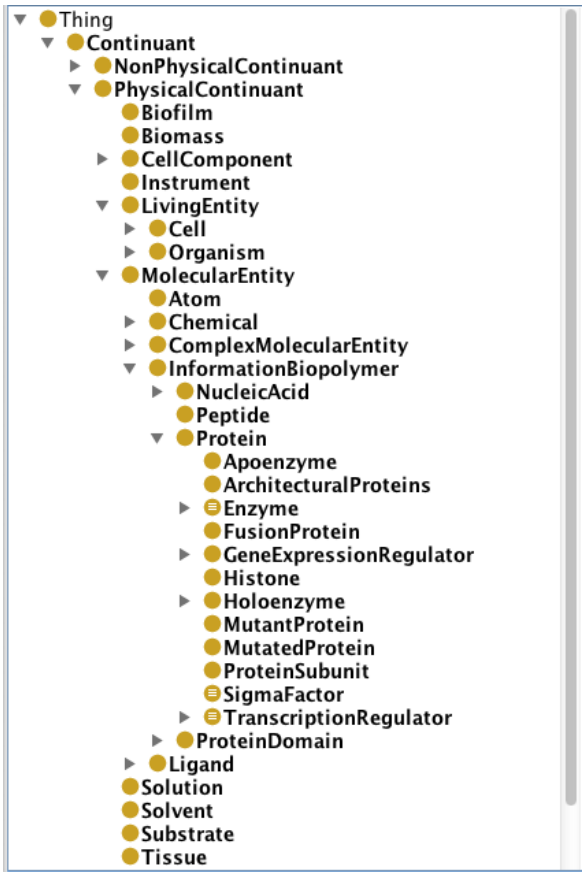

Figure 2. A screenshot of the Continuant branch of the GRO

A mention should refer to an instance or a group of instances of a GRO concept. A mention may be one of 1) the name of an instance, 2) the name of a concept, 3) a synonym of the names, and 4) an expression that refers to an instance or a group of instances. First, we annotated all instance names unless they are parts of expressions of other entities. Second, we did not annotate concept names if they are not used as arguments of events or relations, assuming that they can be easily recognized by string matching. Third, we annotated entity-referring expressions (e.g. “the protein”) also only when they are arguments of some events. Note that we have not annotated the co-references between entity names and entity-referring expressions.

It is important to annotate only the expression that has correspondence with an ontology concept, rejecting other expressions that have little or no correspondence with the concepts defined in the GRO. Also it is worth noting that no instances of the concepts are included in the ontology.

In some cases, a mention consists of two discontinuous spans. In Figure 3, “U4 … snRNA” is an example discontinuous mention, which sometimes appears in such structures as coordination. However, we do not allow a discontinuous span with more than two segments or with two segments from different sentences.

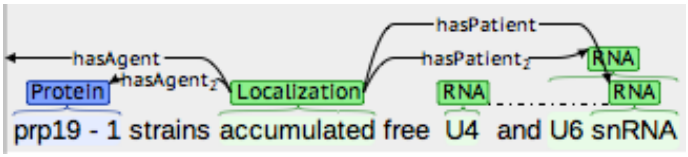

Figure 3. An example discontinuous mention

In the training dataset, the entities that are not involved in any relations will not be annotated, but the test dataset will have all possible entity annotations, and the participants are supposed to find only valid relations among them.

5 Event annotation

This section introduces the event annotation applied in the GRO task.

5.1 Annotation overview

The event annotation marks references to biological processes occurring on the intracellular level. Many concepts for those biological processes are defined under the Occurrent branch of the GRO. Table 1 shows the definitions of such example concepts.

Event type	Definition
affecting	process that affects the living units at the molecular level, including maintenance, modification, producing and regulation
biological process	process pertinent to the functioning of integrated living units
decrease	process of becoming smaller, less numerous, less important, or less likely
experimental intervention	process that has intervention to the living units through biological experiment
increase	process of becoming larger, more numerous, more important, or more likely
intra organismal process	biological processes occurring inside the organism
localization	process for entity that relates to its transportation or maintenance in a specific location
molecular process	process that is carried out at the molecular level
mutation	permanent change in the DNA sequence of a gene that alters its genetic message
organismal process	process occurring at the organismal level that pertinent to the function of the organism
physical interaction	action occurring as two or more objects have a physical effect upon each other
regulatory process	process that modulates the frequency, rate or extent of a biological process
response process	process of changing state or activity of a living unit resulting from a stimulus

Table 1. Example Occurrent concepts and their definitions

5.2 Event argument role

An event consists of a mention referring to one of the Occurrent concepts and one or more relations with other mentions, where the relations are defined in the GRO. The following two relations are exclusively, yet optionally, used for events:

hasAgent: the entity/event that deliberately performs the event. For example, in the text "P1 inhibits P2 expression", the entity of "P1" is the agent of the event "inhibits"; therefore, in the structured representation of events, the event "inhibits" **hasAgent** "P1".

hasPatient: the entity/event that undergoes the event and changes its states. In the above example, note that there is another event mention "expression", and "P2" is the patient of the event; therefore, the event "expression" **hasPatient** "P2".

However, one event type that is worth special notes is the binding. If the binding is between DNA and protein, we use the concept BindingOfProteinToDNA, discarding the concept BindingOfDNAToProtein and assuming that the protein is the agent of the binding event and DNA is the patient. If the binding is between two entities of the same type (e.g. two proteins), we use such a concept as BindingToProtein, where both entities are related to the binding event via hasPatient, as shown in Figure 4. We allow only for such events as BindingToProtein to have multiple arguments of the same type. Other events have only one argument of a type, as we split events with multiple arguments of a type, which are mostly related to coordination structures, into those with a single argument of a type.

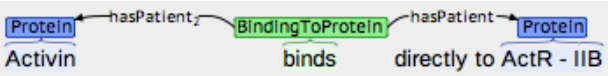

Figure 4. An example annotation of BindingToProtein

5.3 Annotation process

We first find a mention that refers to one of the Occurrent concepts (i.e. event trigger), and we link it to its arguments via relations defined in the GRO. Note that we do not allow cross-sentence annotations.

6 Relation annotation

This section introduces the relation annotation applied in the GRO task.

6.1 Annotation overview

Similar to the event annotation, the relation annotation also assumes the entity annotation as its input. However, in the relation annotation, we usually do not identify the trigger words that explicit express the relation between the entities. The relation annotation is thus to represent relations between arguments without trigger words. Currently, we consider the following seven types of relation:

encodes: describes the biologically encoding relation between entities. For example, the "gene" **encodes** "protein".

hasFunction: defines the relation between an entity and its function. For instance, the "enzyme" **hasFunction** "catalytic activity".

locatedIn the location where an entity/event exists or occurs.

precedes the sequential order between the events.

hasPart: the containment or involvement relation between two entities/events. For example, we can have the representation of "proteincomplex" **hasPart** "protein".

resultsIn: the cause-and-result relation between two entities/events. For instance, "transcription" **resultsIn** "RNA".

fromSpecies: the fromSpecies relation relates species-specific entity/event to the species they refer to, such as "bacterial RNA polymerase" refers to "bacterium", e.g. "RNA polymerase" **fromSpecies** "bacterium".

startsIn: a spatial relation between an event and an entity where the event starts from the place referred to by the entity. E.g. "localization" **startsIn** "Cytoplasm"

endIn: a spatial relation between an event and an entity where the event ends at the place referred to by the entity. E.g. "localization" **endsIn** "Nucleus"

Note that the "bacterium" of the example **fromSpecies** relation above acts as a modifier of the "RNA polymerase". The **encodes** relation can be also used for such modification, e.g. "mRNA" **encodes** "pAT 133" in the example "pAT 133 mRNA". We annotated such relations for modification when the modifee is used as an argument of an event, as exemplified in Figure 5.

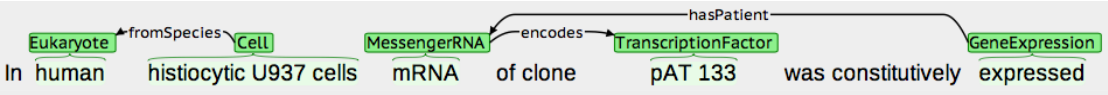

Figure 5. Example relation annotation for the GRO task

References

[1] E. Beisswanger, V. Lee, J.-J. Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz, and U. Hahn, "Gene Regulation Ontology (GRO): design principles and use cases," *Stud Health Technol Inform*, vol. 136, pp. 9–14, 2008.

[2] Jung-Jae Kim, Xu Han and WatsonWei Khong Chua. Annotation of biomedical text with Gene Regulation Ontology:Towards Semantic Web for biomedical literature. Proceedings of LBM 2011, pp.63–70.

[3] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. ichi Tsujii, "brat: a Web-based Tool for NLP-Assisted Text Annotation," *EACL*. The Association for Computer Linguistics, pp. 102–107, 2012.