

# Gene Regulation Ontology (GRO): Design Principles and Use Cases

Elena BEISSWANGER<sup>a1</sup>, Vivian LEE<sup>b</sup>, Jung-Jae KIM<sup>b</sup>, Dietrich REBHOLZ-SCHUH-MANN<sup>b</sup>, Andrea SPLENDIANI<sup>c</sup>, Olivier DAMERON<sup>c</sup>, Stefan SCHULZ<sup>d</sup>, Udo HAHN<sup>a</sup>

<sup>a</sup> *Jena University Language and Information Engineering (JULIE) Lab, Jena, DE*

<sup>b</sup> *European Bioinformatics Institute, Hinxton, Cambridge, UK*

<sup>c</sup> *Laboratoire d'Informatique Médicale, Université de Rennes 1, Rennes, FR*

<sup>d</sup> *Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, DE*

**Abstract.** The Gene Regulation Ontology (GRO) is designed as a novel approach to model complex events that are part of the gene regulatory processes. We introduce the design requirements for such a conceptual model and discuss terminological resources suitable to base its construction on. The ontology defines gene regulation events in terms of ontological classes and imposes constraints on them by specifying the participants involved. The logical structure of the ontology is intended to meet the needs of advanced information extraction and text mining systems which target the identification of event representations in scientific literature. The GRO has just been submitted to the OBO library and is currently under review. It is available at [http:// www.ebi.ac.uk/Rehholz-srv/GRO/GRO.html](http://www.ebi.ac.uk/Rehholz-srv/GRO/GRO.html)

**Keywords.** Bio-ontologies, Knowledge bases, Terminology-vocabulary

## Introduction

In this paper we introduce an ontology, which deals with the conceptual structures underlying gene regulation. It constitutes the ontological backbone of an information extraction and text mining system developed within the framework of the BOOTStrep project,<sup>2</sup> whose goal is to automatically harvest unstructured information from natural language documents and to structure this information in a biological fact data base.

The processing requirements for information extraction are usually more sophisticated than those for document retrieval or classification tasks. Hence, formal foundations and a high level of expressivity are needed for the representation of domain knowledge in terms of an ontology. These considerations preclude, to a large extent, the direct re-use of existing biomedical vocabularies or terminological resources

---

<sup>1</sup>Corresponding author: Elena Beisswanger, Jena University Language & Information Engineering (JULIE) Lab, Fürstengraben 30, 07743 Jena, Germany; E-mail: [elena.beisswanger@uni-jena.de](mailto:elena.beisswanger@uni-jena.de)

<sup>2</sup><http://www.bootstrep.org>

as available from the *Unified Medical Language System* (UMLS),<sup>3</sup> since they are neither formal nor expressive enough. These resources, however, still have to be checked for relevant terminology in the field of gene regulation that subsequently could be enriched with new logical interconnections. In a similar way, the *Open Biomedical Ontologies* (OBO) library<sup>4</sup> has to be screened, a collection of publicly available biomedical ontologies that has some intersections with the medically focused UMLS.

After a brief introduction into the domain of gene regulation, we will review the resources we used in setting up the *Gene Regulation Ontology* (GRO) (Section 3.1), and then survey the structure of GRO in terms of its basic entity types and relations, closely following the recommendations issued by the OBO Foundry [1].

## 1. Gene regulation - a brief refresher

Gene regulation, the regulation of gene expression, characterizes the cellular mechanisms that control the amount of gene products of individual genes synthesized at a particular time and under particular extra- and intracellular conditions. Most known mechanisms regulate the expression of protein coding genes. Gene expression falls into two major phases, *viz.* transcription and translation. During transcription, proteins called transcription factors (TF) bind to specific binding sites of a gene. This process starts or stops the transcription of the gene into an RNA, the intermediate product of gene expression, by a polymerase enzyme. In the second part of gene expression, the RNA is translated into a protein. Regulatory processes occur on all of the different steps of gene expression, from transcription to post-translational protein modification. They enable the cell to adapt to different conditions controlling its structure and function. In many cases abnormal regulation of gene expression causes diseases. A prominent example is the induction of cancer in cells, in which abnormal regulation of gene expression plays a crucial role. Our work will particularly focus on the regulation occurring on the transcriptional level of gene expression.

## 2. The Gene Regulation Ontology (GRO)

GRO has been created as a conceptual model for the domain of gene regulation following best practice design principles as recommended by the OBO Foundry [1]: It is publicly available, uses a *commonly shared syntax*, *viz. the Web Ontology Language (OWL)*,<sup>5</sup> and has a *clearly specified content*. It covers processes occurring on the intracellular level (such as the binding of TFs to DNA binding sites), and physical entities that are involved in these processes (such as genes and TFs) in terms of ontology classes interlinked by semantic relations. No instances of classes are included in the GRO. As an ontology supporting natural language processing (NLP) applications (see Section 4), the GRO is intended to represent *common knowledge* about the domain and focuses on the relations between types in the domain, rather than representing overly fine-grained classes as can be found in ontologies created for data base annotation purposes, such as the *Gene Ontology* (GO) [2].

---

<sup>3</sup><http://www.nlm.nih.gov/pubs/factsheets/umls.html>.

<sup>4</sup><http://www.bioontology.org/repositories.html#obo>

<sup>5</sup><http://www.w3.org/TR/owl-features/>

## 2.1. Design and Construction of the GRO

The basic structure of the GRO was developed manually based on textbook knowledge and considering the terminology already available within the UMLS resources. To populate GRO and interlink it with existing ontological resources the OBO ontologies were screened by two biologists for entries related to gene regulation. These were subsequently extracted and integrated into the GRO, while keeping the references to the source terminologies. In addition, information was taken from the transcription factor database TransFac [3]. Table 1 lists all selected sources and the kind of information we derived. To complete the conceptual representation of gene regulation, in the next step, 150 Medline<sup>6</sup> abstracts (selected by a MeSH<sup>7</sup> query and additional selection criteria) were analyzed with regard to potentially new GRO terms.

The GRO relies on a taxonomic backbone based on *is-a* (subclass) relations between ontology classes. In addition, it provides a set of semantic relation types following and extending the OBO *Relation Ontology* (RO) [4]. The RO has recently become a *de facto* standard for ontology relations in the biomedical domain and using the RO relations will improve the interoperability between different ontologies.

**Table 1.** Conceptual resources for the population of the GRO

| Resource with URL  | Relevant Information   |
|--|--|
| Gene Ontology (GO) <a href="http://geneontology.org/">http://geneontology.org/</a>             | molecular functions, biological processes, cellular components |
| Sequence Ontology (SO) <a href="http://sequenceontology.org/">http://sequenceontology.org/</a> | sequence regions and attributes of sequence regions            |
| ChEBI <a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a>                    | chemical entities  |
| INOH Molecule Role (IMR) <a href="http://www.inoh.org/">http://www.inoh.org/</a>               | transcription factors and their functional domains             |
| NCBI taxonomy, <a href="http://130.14.29.110/Taxonomy/">http://130.14.29.110/Taxonomy/</a>     | eukaryotes, prokaryotes  |
| TransFac <a href="http://www.gene-regulation.com/">http://www.gene-regulation.com/</a>         | transcription factors, domains of transcription factors        |

The GRO is continuously evolving. The current version (as of February 20, 2008) comprises 433 classes related by 396 semantic relations of 8 different types, exclusive the taxonomic *is-a* relation and inverse / reciprocal relations (see Section 3.2).

## 2.2. GRO Classes and Relations

Basically the GRO consists of two branches. While the **continuant** branch describes entities ‘which persist through time’, the **occurrent** branch describes process entities, such as **transcription**, **gene expression**, and various regulatory processes. A continuant can either be ‘physical’, i.e., it has a spatial dimension (such as **gene**, **regulatory sequence**, and **protein**), or ‘non-physical’ (such as **protein function**), whereas occurrents are always non-physical.

A feature that distinguishes the GRO from most other biomedical ontologies is that its classes are highly interlinked by various manually encoded relations (reciprocal relations are given in brackets, hereafter). The relation *part-of* (*has-part*) is used to

<sup>6</sup><http://www.nlm.nih.gov/pubs/factsheets/medline.html>

<sup>7</sup><http://www.nlm.nih.gov/pubs/factsheets/mesh.html>

relate spatial or temporal parts to the whole, such as the class **protein domain** to the class **protein** or the class **transcription initiation** to the class **transcription**. Wholes and their parts must belong to the same ontological category, i.e., a continuant can only have continuants as parts and an occurrent can only have occurrents as parts. The relation *from-species* relates species-specific classes to the species they refer to, such as **bacterial RNA polymerase** refers to **bacterium**. Also, continuants and occurrents are related to processes in which they are involved by the relation *participates-in* (*has-participant*), or by one of the two sub-relations *agent-of* (*has-agent*) or *patient-of* (*has-patient*), respectively. Consider, e.g. the event **regulation of transcription** – in the GRO it is defined as a subclass of a **regulation of gene expression** with the following restrictions: (i) **regulation of transcription** *has-agent* **transcription regulator**, and (ii) **regulation of transcription** *has-patient* **transcription**. Finally, the relation *encodes* (*encoded-in*) relates genes to proteins, *function-of* (*has-function*) links functions to their bearers, *has-quality* specifies qualities inherent in particular entities, and *results-in* (*results-from*) identifies the outcome of a process. To further exploit the potential of formal ontologies, domain and range constraints, as well as the algebraic properties (such as transitivity, reflexivity) of the relations need to be defined in the GRO in a pending revision step.

### 2.3. Modeling Decisions and Implementation

Exactly as the GO and the *Sequence Ontology* (SO), which are both members of the OBO, the GRO is designed in a species-independent manner. However, since major differences between eukaryotic and prokaryotic gene regulation exist, they have to be reflected by the ontological representation. To allow for species-specific classes in GRO we introduced the relation *from-species*. An advantage of this approach is that it does not require a separate ontology for every particular species. In contrast to the GO, which is basically used by human curators, the GRO is designed for NLP applications (Section 4), which require, e.g., to distinguish between functions and function bearers. Thus, the GRO holds both, functions as described in the GO molecular function branch, and classes representing function bearers, linked by *has-function* relations.

The GRO is implemented in OWL-DL, the description logic variant of OWL. OWL offers mechanisms to define classes and to relate them according to their properties. The formal definition of a typical GRO class holds a Uniform Resource Identifier (URI), a human readable class name provided by the predefined OWL property `rdfs:label`, a textual definition explaining the meaning of the class, at least one *is-a* relation, and references to similar classes in external resources such as the OBO ontologies. In addition, OWL class restrictions which involve further relations (e.g., *has-agent*, *has-patient*) are used to logically constrain the meaning of a class. OWL supports negation as well as existential and universal constraints.

Classes which do not share any instances are disjoint. To enable more restrictive consistency checks in GRO, we exploited OWL's support to explicitly express class disjointness and added corresponding statements to the GRO wherever this was necessary. We marked, e.g., the classes **DNA** and **RNA** as disjoint because an existing nucleotide sequence, by nature, cannot belong to both classes, whereas the classes **cell** and **organism** are not disjoint because they share instances, such as bacterial cells. There are some limitations of OWL-DL that we had to deal with when implementing the ontology. For example, n-ary relations are not supported, but usually they can be expressed by a set of binary relations. Further, neither statements such as “some

process X *has-agent* some protein Y”, nor uncertainty, nor exceptions can be expressed in OWL. In the BOOTStrep project, we decided to move this kind of information into a separate biological fact data base that is linked to the GRO.

### 3. Applications of the Gene Regulation Ontology

Bio-ontologies are designed to meet the needs of predefined use cases (e.g., annotation of genetic sequences). GRO is intended to be used as the conceptual backbone for automatic document analysis, information extraction and text mining tasks, in particular. This imposes specific requirements on GRO which are different from those for document retrieval or classification tasks (typical applications of the UMLS). They also differ from those for functional annotations in bio databases, the common application area for many of the OBO ontologies.

#### 3.1. Vocabulary for the Semantic Annotation of Scientific Documents

In the BOOTStrep project, a proper subset of the GRO terms is used as annotation vocabulary for the semantic annotation of biological documents. Corpora annotated by such semantic meta data are a prerequisite for supervised machine learning algorithms. In BOOTStrep, semantic annotations are carried out on several levels. The bottom level is the annotation of entities, i.e., the assignment of labels from a controlled vocabulary to the participants of gene regulation, such as TFs and genes. The vocabulary is taken from the **continuant** branch of the GRO. The next level of annotation relates to regulatory processes (event annotation), a much more complex task, which requires entity annotations as a prerequisite. Terms from the **occurrent** branch of the GRO are used as vocabulary and participation relations specified for these terms are exploited to constrain the assignment of semantic roles. There is some evidence here that annotation tasks like the one outlined above might profit from an ontology-based annotation vocabulary [5].

#### 3.2. Basis for SWRL Rules to Derive Biological Knowledge

The reasoning facilities that come with the OWL language family are needed for additional processing tasks like the analysis of complex representations of events that require the identification of participants involved and their relations. We are currently investigating into a system that extracts mentions of events involving gene regulations from the scientific literature by using the GRO. The system first recognizes entities such as gene and protein names and labels them with classes in the **continuant** branch of the GRO, particularly with the two classes **gene** and **transcription regulator**. The entity recognition utilizes UniProtKB<sup>8</sup> and RegulonDB<sup>9</sup> as sources for names of the two classes. The system then identifies instances of the classes in the **occurrent** branch by matching linguistic patterns of keywords of the classes (e.g. ‘regulate’ for **regulatory process**). It finally deduces complex information by employing rules written in SWRL<sup>10</sup> thus capturing sophisticated forms of biological knowledge. For

---

<sup>8</sup><http://www.ebi.ac.uk/uniprot/>

<sup>9</sup><http://regulondb.ccg.unam.mx/>

<sup>10</sup><http://www.w3.org/Submission/SWRL/>

example, the system may identify two instances of the classes **regulation of gene expression** and **binding of transcription factor to DNA**, which share instances of **transcription regulator** and **gene** as their agent and patient, respectively. It will then make the rule-based deduction that the instance of **regulation of gene expression** also belongs to the class **regulation of transcription** which is a subclass of **regulation of gene expression**. We are developing such rules, which should be OWL-DL safe [6], based on the classes and relations provided by the ontology. The GRO is necessary to implement these rules, since such inferences cannot be achieved without a logically sound ontology

#### 4. Conclusions and Outlook

We have introduced the *Gene Regulation Ontology* GRO, a conceptual model for the regulation of gene expression. It covers both, gene regulatory processes occurring on the intracellular level and molecular entities participating in these processes. The GRO was created within the BOOTStrep project aiming at the integration of biomedical knowledge at a very large scale. In particular the GRO serves as controlled vocabulary for semantic annotation of biomedical texts which forms the basis for knowledge-intensive NLP tasks, such as rule-based high quality information extraction.

In the next period of work the participation relations provided by the ontology will be refined to enable an even more fine-grained process modeling. Further on, the expressivity of the ontology will be increased by defining domain and range constraints of the already existing relations, as well as their algebraic properties (such as transitivity, reflexivity). We expect that besides using the GRO in support of NLP tasks it might become relevant for ongoing biological research, too, raising interest and demands for its use in the classification of genes and transcription factors according to the processes they are involved in. In this way, the GRO would complement the Gene Ontology by formally describing what a particular GO term ‘means’ in terms of specifying relations between participants and the processes in which they are involved.

#### Acknowledgments

The work presented here is part of the BOOTStrep project, a Specific Targeted Research Project (STREP) of the 6th Framework Program of the European Union (FP6 - 028099).

#### References

- [1] B. Smith, M. Ashburner, C. Rosse, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
- [2] The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(1):258–261, 2004.
- [3] P. Stegmaier, A. E. Kel, E. Wingender. Systematic DNA-Binding Domain Classification of Transcription Factors. *Genome Informatics*, 15(2): 276–286, 2004.
- [4] B. Smith, W. Ceusters, B. Klagges, et al. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46 (1:15), 2005.
- [5] A. Kawazoe, L. Jin, M. Shigematsu, et al. The development of a schema for the annotation of terms in the BioCaster disease detec./trac. system. Proceedings of the KR-MED 2006 Workshop, Baltimore, MD, USA, Nov. 2006, pp.77-85
- [6] B. Motik, U. Sattler, R. Studer. Query answering for OWL-DL with rules. Proceedings of the 3rd International Semantic Web Conference (ISWC 2004), Hiroshima, Japan, November 2004, pp. 549-563.