BBA - Gene Regulatory Mechanisms 1864 (2021) 194766

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Formalization of gene regulation knowledge using ontologies and gene ontology causal activity models☆

Belén Juanes Cortés [a], José Antonio Vera-Ramos [b], Ruth C. Lovering [c], Pascale Gaudet [d], Astrid Laegreid [e], Colin Logie [i], Stefan Schulz [b], María del Mar Roldán-García [f,g,h], Martin Kuiper [j], Jesualdo Tomás Fernández-Breis [a,*]

[a] Departamento de Informatica y Sistemas, University of Murcia, CEIR Campus Mare Nostrum, IMIB-Arrixaca, Campus de Espinardo, 30100 Murcia, Spain
[b] Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerpl. 2, Graz, Austria
[c] Institute of Cardiovascular Science, Faculty of Pop Health Sciences, University College London, Rayne Building, 5 University Street, London WC1E 6JF, United Kingdom
[d] Swiss Institute of Bioinformatics, 1, rue Michel Servet, 1211 Geneva 4, Switzerland
[e] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Gastrosenteret, 431.03.046, Øya, Prinsesse Kristinas gate 1, Trondheim, Norway
[f] Departamento de Lenguajes y Ciencias de la Computación, University of Málaga,Bulevard Louis Pasteur 35, 29071 Málaga, Spain
[g] ITIS Software, University of Málaga, Calle Arquitecto Francisco Peñalosa s/n, 29071 Málaga,Spain
[h] Biomedical Research Institute of Málaga (IBIMA), University of Málaga, Calle Doctor Miguel Díaz Recio, 28, 29010 Málaga, Spain
[i] Faculty of Science, Radboud Institute for Molecular Life Sciences, Geert Grooteplein Zuid 28, 6525, GA, Nijmegen, the Netherlands
[j] Department of Biology, Norwegian University of Science and Technology, Realfagbygget, Høgskoleringen 5, 7034 Trondheim, Norway

ARTICLE INFO

Keywords:
Gene regulation
Bioinformatics
Knowledge representation
Ontologies
Gene ontology

ABSTRACT

Gene regulation computational research requires handling and integrating large amounts of heterogeneous data. The Gene Ontology has demonstrated that ontologies play a fundamental role in biological data interoperability and integration. Ontologies help to express data and knowledge in a machine processable way, which enables complex querying and advanced exploitation of distributed data. Contributing to improve data interoperability in gene regulation is a major objective of the GREEKC Consortium, which aims to develop a standardized gene regulation knowledge commons. GREEKC proposes the use of ontologies and semantic tools for developing interoperable gene regulation knowledge models, which should support data annotation. In this work, we study how such knowledge models can be generated from cartoons of gene regulation scenarios. The proposed method consists of generating descriptions in natural language of the cartoons; extracting the entities from the texts; finding those entities in existing ontologies to reuse as much content as possible, especially from well known and maintained ontologies such as the Gene Ontology, the Sequence Ontology, the Relations Ontology and ChEBI; and implementation of the knowledge models. The models have been implemented using Protégé, a general ontology editor, and Noctua, the tool developed by the Gene Ontology Consortium for the development of causal activity models to capture more comprehensive annotations of genes and link their activities in a causal framework for Gene Ontology Annotations. We applied the method to two gene regulation scenarios and illustrate how to apply the models generated to support the annotation of data from research articles.

Abbreviations: BFO, Basic Formal Ontology; ChEBI, Chemical Entities of Biological Interest Ontology; CoreP-E, core promoter element; coTF, co-factor transcription factor or chromatin modifier; dbTF, DNA-binding transcription factor; eRNA, enhancer RNA; GO, Gene Ontology; GO-CAM, Gene Ontology Causal Activity Model; GRAO, Gene Regulation Application Ontology; GREEKC, Gene Regulation Ensemble Effort for the Knowledge Commons; GTF, general transcription factor; OBO, Open Biomedical Ontology; OWL, Web Ontology Language; PIC, Pre-Initiation Complex; RDF, Resource Description Framework; RDFS, Resource Description Framework Schema; RNA pol II, RNA polymerase II; RO, Relations Ontology; SO, Sequence Ontology; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site.

☆ This article is part of a Special Issue entitled: Curation of the Gene Regulatory Knowledge Commons edited by Dr. Colin Logie, Dr. Wyeth Wasserman and Dr. Julio Collado.

* Corresponding author.
E-mail addresses: bjuanescortes@gmail.com (B. Juanes Cortés), jose.vera-ramos@medunigraz.at (J.A. Vera-Ramos), r.lovering@ucl.ac.uk (R.C. Lovering), pascale.gaudet@sib.swiss (P. Gaudet), astrid.lagreid@ntnu.no (A. Laegreid), c.logie@science.ru.nl (C. Logie), stefan.schulz@medunigraz.at (S. Schulz), mmar@lcc.uma.es (M.M. Roldán-García), martin.kuiper@ntnu.no (M. Kuiper), jfernand@um.es (J.T. Fernández-Breis).

https://doi.org/10.1016/j.bbagrm.2021.194766
Received 8 February 2021; Received in revised form 13 September 2021; Accepted 11 October 2021
Available online 25 October 2021

1. Introduction

Cellular mechanisms that control the transcription of genes at a given time and under particular intracellular conditions, as well as in response to signals from the cell environment, are central to the regulation of gene expression. The domain of transcription regulation comprises the transcription factors that interact with the genome and its transcriptional machinery, coordinated by a complex network of signal transduction molecules. It is therefore crucial to understand the mechanisms of cell control at the system level [1]. Inside the cell, cellular signaling cascades support the transmission of information from external stimuli (e.g. hormones) to distinct cellular responses (e.g. changes in gene expression) [2,3]. Gene expression is regulated at multiple levels starting from the accessibility of chromatin to the post-translational modification of proteins [4,5]. Most of this regulation is controlled by a network of highly interconnected proteins known as transcription regulators [3]. There is a large array of transcription regulators including general transcription factors, sequence-specific DNA binding transcription factors (dbTFs), various transcription co-factors and chromatin-modifying complexes [6]. DbTFs are one of those kinds of regulatory proteins. They bind short DNA sequences to regulate gene expression at the level of transcription by activating or blocking the recruitment of the transcriptional machinery to the transcription start site. Research in the field of gene regulation is of great relevance, because the alteration in gene regulation due to mutations in regulatory sequences, transcription factors, co-factors and chromatin regulators that interact with these regions are associated with the development of serious diseases, such as autoimmunity or cancer [7,8].

Gene regulation research takes advantage of the massive amounts of biological data produced by omics technologies. The representation of gene regulation data into a network that describes a biological system constitutes the first step for scientists to develop an understanding of the behavior of that system [9]. In systems biology, dynamic simulations with a model of a biological process serve as a means to validate the model's architecture and parameters, and to provide hypotheses for new experiments [9].

These computer models need to have access to all the knowledge about the components of the biological system and how they interact with each other. Unfortunately, as it happens in most biological domains [9,10,11,12], gene regulation data are often stored in disparate, specialized repositories. Hence, gene regulation research requires the integrated exploitation of different types of data, but current data resources lack interoperability. Effective data interoperability would enable biological knowledge discovery, which is more and more dependent on computational modeling and simulation.

Data interoperability has been addressed in biological research by applying knowledge representation technologies such as ontologies. For example, the Gene Ontology (GO) [13,14] supports the interoperability of functional annotations across databases and species. Ontologies can be defined as explicit specifications of the conceptualization of a domain [15], and they provide the specified concepts and relations between them that are possible in a particular domain. When supporting data annotation, they are mainly applied as controlled vocabularies, i.e. their labels (in Gene Ontology named GO terms) are used to characterize the role or location of an entity as described in a scientific publication.

GO annotations have been very useful in the last twenty years and have supported multi-species data analysis and functional inference methods, among other bioinformatics methods. However, GO annotations have the limitation of denoting associations between gene products and molecular functions, biological processes or cellular components only. GO includes terms for gene regulation processes, which permit the association of gene products with particular regulation activities. This contribution to the achievement of gene regulation data interoperability is limited by the fact that the conventional GO annotations do not specify the biological context and conditions in which the gene product participates in that process. Therefore, the knowledge model provided by GO to support data annotations needs to be enriched.

The Gene Ontology Consortium has proposed to enrich GO annotations by chaining them together in a knowledge graph by using the Gene Ontology Causal Activity Model (GO-CAM) [16]. GO-CAMs are based on annotations of the form <subject, predicate, object> and thus a model is a collection of triples connected to each other in which the broader biological context is described. This representation based on triples is characteristic of the Resource Description Framework (RDF)[1] and the A-Box segment in semantic languages such as the Web Ontology Language (OWL)[2]. GO-CAMs are activity centric annotations, whereas GO Annotations are gene product centric.

In this work, we hypothesize that GO-CAMs can be designed as templates for curators to annotate specific types of mechanisms involved in the regulation of transcription. By designing and testing several of these GO-CAM templates for practical use, we investigate whether causal activity modeling is an appropriate approach to formalize gene regulation knowledge.

The GREEKC Consortium aims at developing a Gene Regulation Knowledge Commons, focusing on the generation, curation and analysis of data and knowledge about gene regulation processes. GREEKC Working Group 1 (WG1) has been responsible for the development of ontologies for a sustainable knowledge framework for the interoperable representation of gene regulation data. The annotation of dbTFs, general transcription factors (GTFs) transcription co-factors, chromatin modifiers and regulatory RNAs, the genes that are under their regulation and the molecular mechanisms governing this regulation (including protein-DNA interaction, protein-protein interaction, protein/DNA modifications, RNA-mediated events, etc.), which was the domain of GREEKC WG2 (Curation and annotation of gene regulatory mechanisms), needs to be founded on well-defined standards and ontologies. A number of biological ontologies have been generated in the last years, and one of the driving principles of the GREEKC knowledge framework is the reuse of existing content, in order to reduce the development costs of the ontologies and increase their quality by including content that has already been tested. This allows for a consistent representation of a domain among all ontologies that reuse the same content, and increase the interoperability of data and applications through ontologies [17,18].

In this work, we describe the research carried out in GREEK WG1 to facilitate data and knowledge curation in the domain of gene regulatory mechanisms of eukaryotic systems through data annotation templates. Those templates are based on knowledge models built using standardized resources. Hence, we describe in this paper a method for the development of gene regulation knowledge models, which should support data annotation. Given the high data heterogeneity and the lack of harmonized knowledge resources in the field, the starting point for this research was a set of graphical descriptions of typical gene regulation scenarios, which are referred to as cartoons in this article.

This study has the following research questions (RQ):

- RQ1. Can we define knowledge models based on GO-CAM that are well-suited for capturing gene regulatory data?
- RQ2. To what extent can content from existing ontologies be reused?
- RQ3. Is causal activity modeling an appropriate annotation approach for gene regulation?
- RQ4. How can knowledge models be applied to support data annotation?

In this paper, we describe a method for creating the knowledge models (RQ1), which is based on the conversion of the cartoons representing gene regulation scenarios, their description in natural language, the analysis of these descriptions in order to identify mentions of domain entities and relations, and the implementation of the models. The

[1] https://www.w3.org/RDF/
[2] https://www.w3.org/TR/owl-ref/

implementation step will require the reuse of entities from existing ontologies (RQ2). In this work, two gene regulation scenarios were chosen as use cases, for which knowledge models were implemented using two state of the art tools, namely, Noctua [19] and Protégé [20]. These scenarios permitted a proof-of-concept of the application of our method, through which prototypical instances of the knowledge models were developed. Different aspects of the models were evaluated, including their application to annotate data from gene regulation articles (RQ3, RQ4).

The goal of the method is to develop standardized knowledge models representing gene regulation scenarios. In this context, we can interpret standardized in two ways: (1) the knowledge is extracted from cartoons based on grounded knowledge and high quality scientific references; (2) the models are represented with classes and properties from reference bio-ontologies. The availability of a library of models would constitute a standardized representation of gene regulation knowledge that should be the reference for data curators and gene regulation data resources. We believe that the formalization and organization of the gene regulation knowledge will certainly contribute increasing the interoperability of gene regulation data resources.

2. Materials and methods

In this section we describe our method for creating knowledge models from cartoons of gene regulation scenarios (see Fig. 1). This method requires the availability of those cartoons, whose development will be described in Section 2.1. Section 2.2 will explain the construction of knowledge models from the cartoons, which consists of creating descriptions in natural language, extracting entities from existing ontologies and implementing the models. Finally, we will describe how the knowledge models were evaluated (Section 2.3).

2.1. Cartoons representing gene regulation scenarios

GREEKC experts have created a series of cartoons that represent interactions that may occur in selected gene regulation scenarios. The content of the cartoons reflect established molecular biology knowledge in agreement with up to date university text books and review articles in high quality journals. These cartoons are not based on any formalism, so they do not follow any formal grammar. The cartoons are the starting point for our methods. They describe the following two uses cases:

1. Regulation of transcription of a promoter. Fig. 2 shows the cartoon representing how coTFs can modify the structure of the chromatin to allow transcription machinery access to a gene region. Further details and the results on this use case are described in Section 3.1.
2. Regulation of transcription by regulators on an enhancer that loops to a promoter. Fig. 3 shows the cartoon representing how the proteins that regulate transcription, including dbTFs and transcriptional machinery, interact and bind to regulatory DNA regions in order to regulate transcription. In this case, the DNA strand forms a loop in order to facilitate transcription by bringing together a distal regulatory region and the promoter region of a gene. For further details and the results of this use case are described in Section 3.2.

2.2. Construction of knowledge models from cartoons

This section describes our method for creating knowledge models based on the cartoons. The resulting models will be expressed using knowledge representation languages, thus being machine-friendly and sharing formal principles. Our method consists on the following steps: (1) expressing the content of the cartoons in natural language (Section 2.2.1); (2) identifying textual elements that denote domain entities for which ontological representations exist (Section 2.2.2); and (3) implementation of the knowledge model (Section 2.2.3).

2.2.1. Description of the cartoons

The first step was to create natural language descriptions for each cartoon. For this purpose, bibliography in the gene regulation domain was consulted.

2.2.2. Extraction of the ontology entities

The objective of this step was to analyze the natural language descriptions and to identify domain ontology content referred to in the text, as a first step towards the creation of formal knowledge models. Reuse of existing ontology content instead of creating new ontology classes has been formulated as a major design principle for GREEKC. The first decision was therefore to select the ontologies whose content describes gene ontology use cases. The following ontologies were selected, because they represent regulation processes and activities, sequence features and biological entities:

- Gene Ontology (GO): GO describes the roles of genes and gene products in any organism. GO describes three biological aspects: biological process refers to major biological processes to which the gene or gene product contributes, whereas molecular function is defined as the biochemical activity of a gene product, and GO *cellular component* terms specify the place in the cell where a gene product is active [13,16]. It provides the most comprehensive resource currently available for computer- and human readable descriptions of functions of genes and gene products [13,14]. GO is structured using a formal ontology, by defining classes of gene functions (OWL classes), with OWL axioms describing their inter-relations [19].
- Sequence Ontology (SO): SO provides a standardized set of classes and relationships to describe biological sequences. It is focused on unifying the vocabulary used in genomic annotations and provides the structure and axioms needed for querying and automated reasoning over their contents, thus facilitating data exchange and comparative analyses of annotations [21,22].
- Relations Ontology (RO): RO offers a collection of binary relation types covering a variety of domains in the biological, biomedical and environmental sciences designed to support the integration of OBO Foundry ontologies [23]. RO provides biological relationship types such as “positively regulates” and reuses a small number of ontological relations from BFO [24], in particular, upper level classes and core relations such as “has part”.
- Chemical Entities of Biological Interest Ontology (ChEBI): ChEBI [25] focuses on small chemical compounds. Its main purpose is to provide a high quality ontology to promote the correct and consistent use of unambiguous biochemical knowledge. ChEBI consists of four sub-ontologies: molecular structure, which refers to molecular entities or parts which are classified according to structure; biological role, which classifies entities according to their role within a biological context; application, which classifies entities, if any, based on their human intended use; and subatomic particle for particles smaller than atoms.

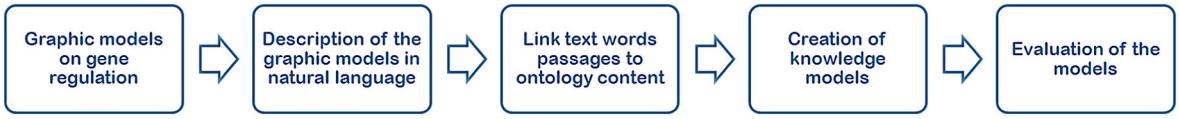

Fig. 1. Block diagram of the method applied in this work.

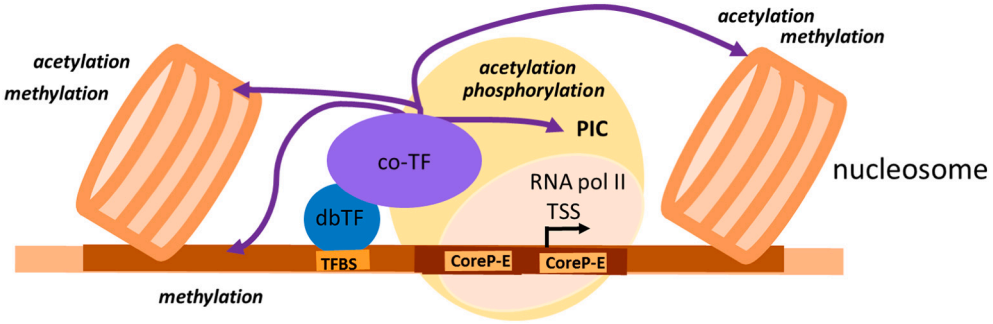

Fig. 2. Cartoon of transcription regulation of a promoter. The brown line represents DNA, with the promoter region in dark brown and the regulatory region in lighter brown. Nucleosomes are shown at the ends. The blue circle represents dbTFs at the TFBS and the purple circle denotes co-factors and chromatin modifiers (coTFs). Purple arrows indicate coTF activity on nucleosomes, DNA and the Pre-Initiation complex (PIC). The PIC and RNA polymerase II (RNA pol II) are shown as pale ovals at the core promoter element (CoreP-E) and the transcription start site (TSS). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

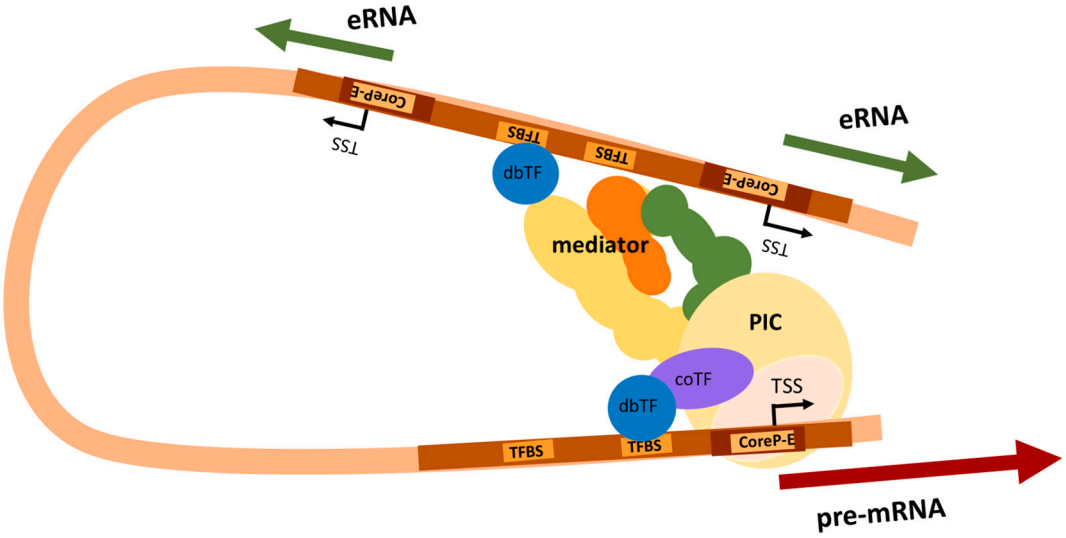

Fig. 3. Cartoon of gene regulation by regulators describing an enhancer that loops to a promoter. The light brown line represents the DNA strand, medium brown represents the regulatory region, with the promoter region in dark brown. The TFBSs and Core P-E are indicated with black arrows marking the TSS. Blue circles represent dbTFs and purple circles co-TFs. The subunits of the mediator complex are depicted in yellow, orange and green. The PIC and RNA polymerase II are represented by pale ovals at the coreP-E and the TSS. Red and green arrows indicate RNA transcription: pre-mRNA, enhancer RNA (eRNA). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- The Gene Regulation Application Ontology (GRAO): GRAO is a GREEKC Consortium development which describes knowledge relevant to the domain of gene regulation. GRAO is an upper-level schema that links the parts of the above-described ontologies relevant for gene regulation. GRAO includes specific cellular processes (such as the binding of dbTFs to DNA binding sites) and the elements involved in those processes (dbTFs and other proteins), but those elements should be reused from existing ontologies. Fig. 4 shows the major classes and relationships that are part of GRAO. This shows the use of prefixes such as GO or SO, which means that those terms are reused from the corresponding ontologies.

The text fragments referring to the potential domain entities (classes and relation) were marked. With the support of ontology browsers such as OntoBee [26] and annotation facilities offered by BioPortal [27], those fragments were first searched in GO, SO, RO, and ChEBI. These search facilities are able to handle synonyms and close matches, which helped to find the appropriate ontology entities. In case the terms were not found in any of these ontologies, they were created as local terms in our GRAO ontology. However, since the aim was to keep the local content of GRAO as small as possible, those concepts and properties were examined. When it was identified that a term should be included in one of the reused ontologies, a new term request was submitted to the corresponding ontology development team.

2.2.3. *Implementation of the knowledge model*

Once the terms and relations involved in the gene regulation scenario had been identified, the next step was to implement the knowledge model. In this work the models were initially implemented using Noctua [28], a web-based, collaborative GO annotation editor, and Protégé [20], an ontology editor. In both cases the same modeling approach was applied to create models based on instances (in OWL terms: to create A-box models).

This approach revealed that Noctua modeling was more appropriate for the goals of the GREEKC Consortium, so in this section we explain how the models were created using this tool. Section 4 provides a justification and discussion of this decision and the major differences between both approaches.

Noctua is a tool developed by the GO Consortium to extend conventional GO annotations, which associate gene products with processes, functions or components, but leave out information about the relevant context or conditions. Thus, the GO Consortium defined the GO-CAM, which chains GO annotations together, and allow for additional contextual information with the application of additional

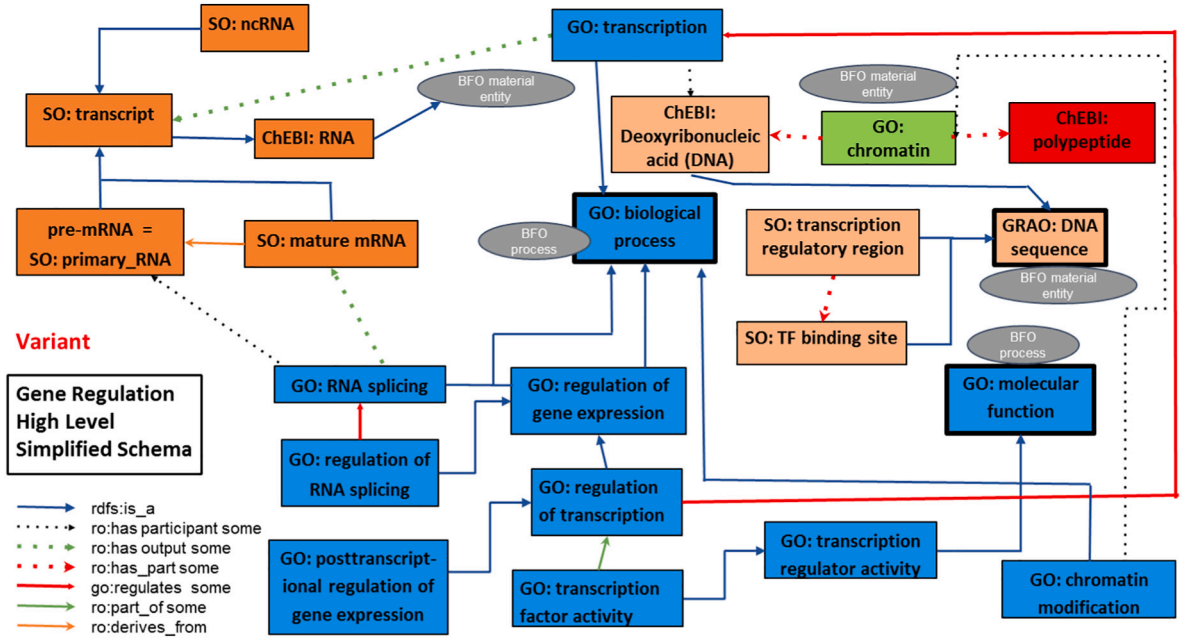

Fig. 4. Gene Regulation High Level Simplified Schema. The use of the prefixes GO, SO, RO, ChEBI and GRAO indicates that those classes and relationships are reused from the corresponding ontologies. Source: https://github.com/GREEKC/GRAO

ontologies. Thus, GO-CAMs improve the expressiveness of the annotations and present a more accurate representation of a biological system rather than a biological component [16]. A GO-CAM combines conventional GO annotations to produce a network of annotations or model [19]. Each term is an individual, that is, an instance of an ontology class. This corresponds to the fact that the knowledge acquired in the annotation process is not universal in the sense that it does not make statements about all members of a class. Instead it is about individual entities observed in individual experiments described in individual publications. To infer universal laws of nature from particular observations and to represent them in ontologies (in OWL terms: to create T-box models) is a separate process in scientific communities beyond this discussion [29,30]. For instance, a scientific author's assertion 'x regulates y' (with x being an instance of X and y an instance of Y) as a result of a particular experiment must not be over-interpreted as a universal statement such as "all instances of X regulate some instance of Y".

The GO-CAM specification generated by the GO Consortium[3] provides an abstract data model and guidelines for the design of GO-CAM graphs. These guidelines specify how activities have to be modeled, the types of relations that can be used and the domains and ranges for them. This allows for checking the correctness and validating the GO-CAMs created. Noctua allows for creating GO-CAMs as OWL A-boxes, that is, based on instances. Noctua provides a graphical editor for creating the GO-CAMs and provides a knowledge base, the Noctua Entity ontology, which includes classes from ontologies such as GO, SO or ChEBI. The relations in these models are object properties from the RO. Each term added to the graphical editor is displayed in a box, called *activity unit* [28], which represents an individual with a unique identifier. GO-CAMs can also include content from databases such as UniProtKB. Furthermore, Noctua permits the export of GO-CAM graphs in OWL, so the model can be imported and edited using tools such as Protégé or text editors. Noctua also offers the possibility for real-time reasoning to check model consistency.

An example of a GO-CAM developed with Noctua is shown in Fig. 5. This GO-CAM describes the molecular activity of *mATF4*, *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228), which is *part_of* (BFO:0000050) the process *positive regulation of the transcription* (GO:0045944) of its input (or target), *mBglap*. This activity *occurs_in* (BFO:0000066) the *chromatin* (GO:0000785) and *has_part* (BFO:0000051) *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978), which also *occurs_in* (BFO:0000066) the *chromatin*, is *enabled by* (RO:0002333) *mATF4*, and *has input* a DNA motif (SO:0000713). In this case, all the mentions of *mATF4* and *mBglap* refer to the same instance of each molecule.

The implementation of these models in Noctua required a search for the corresponding entities in the relevant knowledge bases, selecting them in Noctua and creating the corresponding activity units and relations. In the Noctua editor, the GO MF terms were searched using their identifier or keywords, for example GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific. Next, the relationships between MF and other GO elements, such as cellular components (CC) and biological processes (BP), were selected.

2.3. Evaluation of the knowledge models

The evaluation of the developed models was carried out by applying the following steps: (1) technical evaluation; and (2) application of the models for supporting data annotation. Regarding the technical evaluation, the content of the models was checked for completeness and consistency. First, the terms included in the GO-CAMs were manually compared with the descriptive texts associated with the cartoons. The work was reviewed during group meetings by experts involved in the GREEKC WG1. In a first meeting, the accuracy of the content of the texts in natural language was checked. In a second meeting, the knowledge models were reviewed to determine whether the gene regulation scenarios represented in the cartoons were properly represented. Once our experts agreed on the GO-CAMs, their consistency was tested using the reasoner offered by Noctua. Then, they were implemented from scratch using Protégé. The consistency of the models implemented in Protégé was tested using the reasoners offered by the tool. For simplicity, we have reused the URI of the terms, so not all the logical axioms available in the reused OWL ontologies have been imported.

Regarding the support for data annotation, our models formalize the knowledge associated with particular gene regulation scenarios. Consequently, they can be considered knowledge templates for

[3] http://wiki.geneontology.org/index.php/Noctua

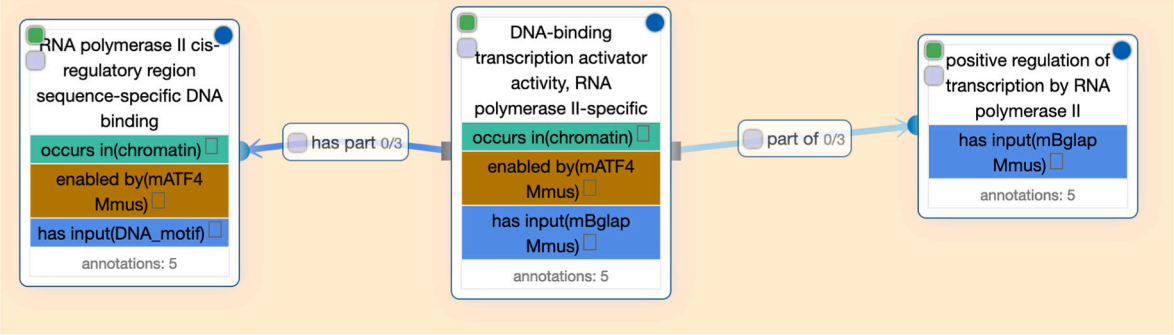

Fig. 5. An example of GO-CAM graph created using Noctua that describes the positive regulation of transcription. The instances of molecular activities and biological processes are shown in boxes and the arrows represent the relationships between the instances.

annotating data related to those scenarios. This component of the evaluation aimed to study whether those templates are really applicable to data annotation. For this purpose, we applied the GO-CAMs developed to annotate the gene regulation scenarios described by selected research articles. Four domain experts evaluated the suitability of the GO-CAMs for representing the data presented in the articles. One expert analyzed all the articles, evaluated the applicability of the GO-CAMs developed and assigned each article to one or more GO-CAMs, depending on the content of the article. Each research article was then analyzed by another domain expert. Hence, three experts were asked to provide one of the following answers for the articles:

1. The GO-CAM is not appropriate for annotating this article.
2. The article does not provide all the data required by the GO-CAM.
3. The GO-CAM permits to capture all the data of the article.
4. The article has relevant data which is not represented in the GO-CAM.

The experts could also add additional information in free text. The results were analyzed by GO-CAM template and globally, that is, all together.

3. Results

In this section the presentation of the results are organized by use case. For each use case, two results are described: the description in natural language and the prototypes of GO-CAMs generated. The description in natural language will serve to provide the bridge between the cartoons of the gene regulation scenarios, which were shown in Section 2, and the GO-CAMs obtained. In fact, as mentioned earlier in this paper, the models were created by analyzing the content of the text descriptions.

3.1. Use case 1: transcription regulation of a promoter

3.1.1. Description in natural language

The gene regulation scenario shown in Fig. 2 can be described in text as: DNA-binding transcription factors (dbTFs) recognize short DNA sequences termed TF binding sites within enhancers relatively free of nucleosomes, thus enabling cooperative binding that can include interactions between dbTFs and coTFs [31]. The coTFs recruited by the dbTFs typically act on gene regulation through modifying and remodeling the chromatin context of enhancers. Such coTFs include histone acetyltransferases (eg.p300/CBP, SAGA complex, MOF, TIP60 and others), histone methyltransferases (e.g.MLL3/4, CARM1), ATP-dependent chromatin remodeling factors (e.g. Brg1, CHD7) and factors that promote crosstalk with the transcriptional machinery (e.g. mediator complex) at promoters. coTFs complexes enable activation or repression of transcription by influencing the activity of RNA polymerase and facilitating the establishment of a transcriptionally permissive or restrictive chromatin environment [32].

3.1.2. The knowledge models

The analysis of this gene regulation scenario revealed that two different situations could happen for this regulation activity. A transcriptional coregulation activity could be involved in the process, but this is not always the case. Given that regulation can be either positive or negative, four models were created. For simplicity, only the GO-CAMs for the positive regulation are shown in Figs. 6 and 7. The other models can be found at https://github.com/jesualdotomasfernandezbreis/greekc.

Fig. 6 shows the model for the positive regulation of transcription by a dbTF and where transcriptional coregulation activity is not represented. The main process of this model is the *positive regulation of transcription by RNA polymerase II* (GO:0045944) which *has input* (RO:0002233) the regulated *gene* (SO:0000704) product. The *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228) is *part of* (BFO:0000050) the regulation of transcription process, which *occurs in* (BFO:0000066) *chromatin* (GO:0000785), and is *enabled by* (RO:0002333) a dbTF, represented by the *chebi information biomacromolecule* (CHEBI:33695), and *has input* (RO:0002233) the same regulated gene product. This molecular activity *has part* (BFO:0000051) a *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (activity) (GO:0000978), which *occurs in chromatin*, and is *enabled by* (RO:0002333) the same instance of dbTF and *has input* a *DNA_motif* (SO:0000713). The dbTF and gene product used would depend on the species curated. For example, UniProt IDs are used for humans, whereas for mouse Mouse Genome Informatics IDs would be included. Note that inputs may be dbTF motifs such as E box, or promoter coordinates, for example, for human genes, using Ensembl genomic coordinates.

The model for the analogous negative regulation would have the same structure but the activities involved would be *negative regulation of transcription by RNA Polymerase II* (GO:0000122) and *DNA-binding transcription repressor activity, RNA polymerase II-specific* (GO:0001227).

Fig. 7 describes the model for the positive regulation of transcription where the activity of a coTF is represented. The main process of this model is the *positive regulation of transcription by RNA Polymerase II* which *has input* the regulated *gene* (SO:0000704) product. The *transcription coactivator activity* (GO:0003713) is *part of* the main activity, *occurs in* the *chromatin*, is *enabled by* a coTF, represented by the *chebi information biomacromolecule* (CHEBI:33695) and *has input* the same regulated gene product. The *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228) is also part of the main process, *occurs in* the *chromatin*, is *enabled by* a dbTF, represented by the *chebi information biomacromolecule* (CHEBI:33695), and *has input* the same regulated gene product. This molecular activity *has part* (BFO:0000051) *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978), which *occurs in* the *chromatin*, is *enabled by* (RO:0002333) the same

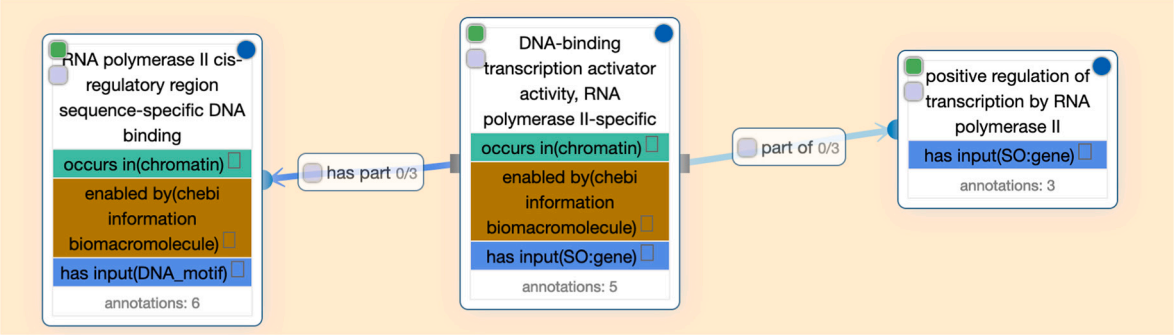

Fig. 6. The GO-CAM representing positive transcription regulation by a promoter when no transcriptional coregulation activity is involved. The main process of this model is the *positive regulation of transcription by RNA polymerase II* (GO:0045944) which *has input* (RO:0002233) the regulated *gene* (SO:0000704) product. The *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228) is *part of* (BFO:0000050) the regulation of transcription process, which *occurs in* (BFO:0000066) the *chromatin* (GO:0000785), and is *enabled by* (RO:0002333) a dbTF, represented by the *chebi information biomacromolecule* (CHEBI:33695), and *has input* (RO:0002233) the same regulated gene product. This molecular activity *has part* (BFO:0000051) a *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978), which *occurs in* the *chromatin*, is *enabled by* (RO:0002333) the same instance of dbTF and *has input* a *DNA_motif* (SO:0000713). The dbTF and gene product IDs used would depend on the species curated, with human UniProt IDs are used whereas for mouse Mouse Genome Informatics IDs would be included. Note that inputs may be dbTF motifs such as E box, or promoter coordinates, for example, for human genes, using Ensembl genomic coordinates.

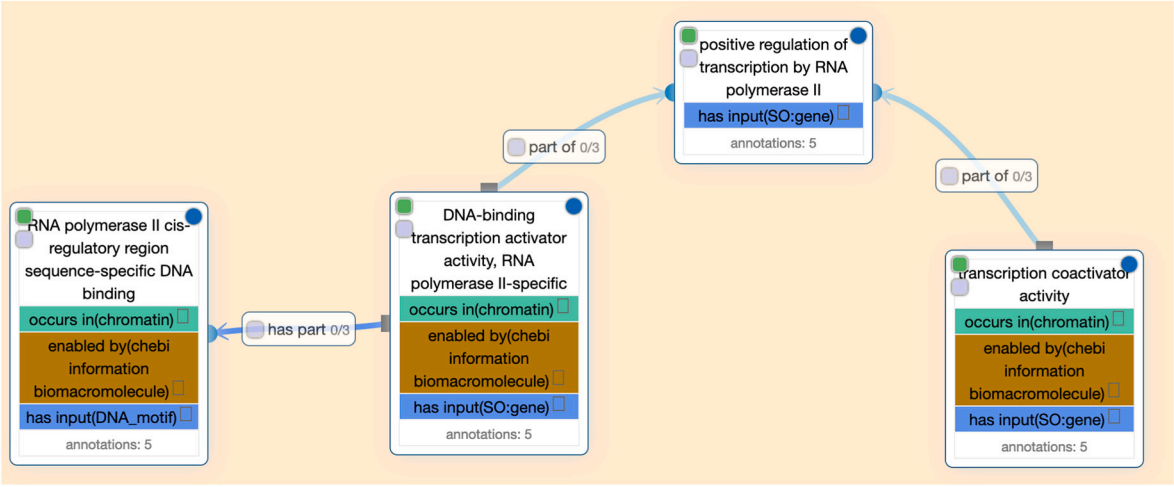

Fig. 7. The GO-CAM representing positive transcription regulation by a promoter when transcriptional coregulation activity is involved. The main process of this model is the *positive regulation of transcription by RNA polymerase II* which *has input* the regulated *gene* (SO:0000704) product. The *transcription coactivator activity* (GO:0003713) is *part of* the main activity, *occurs in* the *chromatin*, is *enabled by* a coTF, represented by the *chebi information biomacromolecule* (CHEBI:33695) and *has input* the same regulated gene product. The *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228) is also part of the main process, *occurs in* the *chromatin*, is *enabled by* a dbTF, represented by the *chebi information biomacromolecule* (CHEBI:33695), and *has input* the same regulated gene product. This molecular activity *has part* (BFO:0000051)*RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978), which *occurs in* the *chromatin*, is *enabled by* (RO:0002333) the same instance of dbTF, and *has input* a *DNA_motif* (SO:0000713). In this case, the two instances of *chebi information biomacromolecule* represent the dbTF and the coTF. Again, inputs may be dbTF or coTF motifs, or promoter coordinates. This model implies that both the dbTF and the co-TF are required for the transcription process to take place.

instance of dbTF, and *has input* a *DNA_motif* (SO:0000713). In this case, the two instances of *chebi information biomacromolecule* represent the dbTF and the coTF. Again, inputs may be dbTF or coTF motifs, or promoter coordinates. Causality/directionality links can be made between the dbTF and coTF activities. The model on Fig. 7 implies that both the dbTF and the coTF are required for the transcription process to take place.

The model for the analogous negative regulation would have the same structure but the activities involved would be *negative regulation of transcription by RNA Polymerase II* (GO:0000122), *transcription corepressor activity* (GO:0003714) and *DNA-binding transcription repressor activity, RNA polymerase II-specific* (GO:0001227).

Table 1 shows the classes and relations that have been reused from GO, SO, ChEBI and RO to create the models of this use case.

3.2. Use case 2: gene regulation by regulators on an enhancer that loops to a promoter

3.2.1. Description in natural language

This is the text obtained for the gene regulation scenario described in Fig. 3: Sequence-specific DNA binding transcription factors (dbTFs) bind to transcription factor binding sites (TFBS) within a gene regulatory region (promoter or enhancer). The DNA-dbTF complex recruits transcription cofactors (coTFs), which bind to different mediator subunits [33]. The mediator complex is crucial for enhancer-promoter loops [34]. The Pre-Initiation Complex (PIC) includes RNA polymerase II and assembles at the transcription start site (TSS) where it initiates transcription of the pre-mRNA [35].

Table 1
Classes and relations reused in the first use case, their URI, source ontology and models in which they appear: 1 is positive regulation of transcription by a promoter without coregulator activity; 2 is negative regulation of transcription by a promoter without coregulator activity; 3 is positive regulation of transcription by a promoter with coregulator activity; and 4 is negative regulation of transcription by a promoter with coregulator activity.

Term	ID	Ontology	Models
RNA polymerase II cis-regulatory region sequence-specific DNA binding	GO:0000978	Gene Ontology (GO)	1,2
DNA-binding transcription activator activity, RNA polymerase II-specific	GO:0001228	Gene Ontology (GO)	1,3
DNA-binding transcription repressor activity, RNA polymerase II-specific	GO:0001227	Gene Ontology (GO)	2,4
transcription coactivator activity	GO:0003713	Gene Ontology (GO)	3
transcription corepressor activity	GO:0003714	Gene Ontology (GO)	4
positive regulation of transcription by RNA polymerase II	GO:0045944	Gene Ontology (GO)	1,3
negative regulation of transcription by RNA Polymerase II	GO:0000122	Gene Ontology (GO)	2,4
chromatin	GO:0000785	Gene Ontology (GO)	1–4
information biomacromolecule	CHEBI:33695	ChEBI	1–4
gene	SO:0000704	Sequence Ontology (SO)	1–4
DNA_motif	SO:0000713	Sequence Ontology (SO)	1–4
has part	BFO:0000051	Relations Ontology (RO)	1–4
enabled by	RO:0002333	Relations Ontology (RO)	1–4
has input	RO:0002233	Relations Ontology (RO)	1–4
occurs in	BFO:0000066	Relations Ontology (RO)	1–4
part of	BFO:0000050	Relations Ontology (RO)	1–4

3.2.2. The knowledge model

Fig. 8 shows the model created for this use case. The main process in this model is the *positive regulation of transcription by RNA polymerase II* (GO:0045944), which *has input* a *gene* (SO:0000704) product. Three activities are *part of* (BFO:0000050) this process, namely, *DNA-binding transcription activator activity, RNA polymerase II-specific* (GO:0001228), the *transcription co-activator activity* (GO:0003713), and *promoter-enhancer loop anchoring activity* (GO:0140585).

The *DNA-binding transcription activator activity occurs in* (BFO:0000066) the *chromatin*(GO:0000785), is *enabled by* (RO:0002333) a dbTF represented by *chebi information biomacromolecule* (CHEBI:33695), and *has input* (RO:0002233) a gene (SO:0000704) product. Besides, this activity *has part* (BFO:0000051) the *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978) which *occurs in* the *chromatin*, is *enabled by* the same dbTF as the previous activity, and *has input* the same gene product. The *transcription co-activator activity, RNA polymerase II-specific occurs in* the *chromatin*, and is *enabled by* the coTF represented by *chebi information biomacromolecule*. The *promoter-enhancer loop anchoring activity occurs in* the *chromatin*, and is *enabled by chebi information biomacromolecule*.

It should be noted that in this model, the two gene products refer to the same gene product, whereas two transcription regulators are involved: the dbTF participates in the activities *DNA-binding transcription activator activity* and *RNA polymerase II cis-regulatory region sequence-specific DNA binding*, and the coTF participates in *transcription co-activator activity*.

Table 2 shows the classes and properties that have been reused from GO, SO, ChEBI and RO to create the models of this use case.

3.3. Evaluation

Regarding the technical evaluation, the reviews of the models during the expert meetings served to ensure that the models were covering the content included in the textual description. The evaluation forms and results are available as spreadsheets at https://github.com/jesualdotomasfernandezbreis/greekc. The application of reasoners to the models indicated no inconsistencies or errors. Regarding data annotation support, we selected a series of gene regulation research articles for the models generated for the two use cases:

- Transcription regulation of a promoter (model 1): [36,37,38,39].
- Transcription regulation of a promoter (model 2): [36].
- Transcription regulation of a promoter (model 3): [40,41,42].
- Transcription regulation of a promoter (model 4): [41]
- Gene regulation by promoter-enhancer looping: [43,39,44,45,46,47,48].

These are the evaluation results by GO-CAM model:

- Transcription regulation of a promoter (model 1): The model permits to capture all the data of the four articles. Fig. 5 is the GO-CAM corresponding to this model created for (Xiao et al, 2005) [37], where ATF4 is the dbTF and mBglap is the gene product.
- Transcription regulation of a promoter (model 2): The model was found not appropriate for the associated article. The same expert considered model 1 appropriate for this article.
- Transcription regulation of a promoter (model 3): The model permits the capture of all the data presented in one article, but two articles were missing some of the data represented in the model.
- Transcription regulation of a promoter (model 4): The article does not contain all the data represented in the model.
- Gene regulation by regulators on an enhancer that loops to a promoter: The model permits the capture of all the data presented in 4 out of 6 articles, whereas one article contained more data that the model could capture, and one article did not provide all the data included in the model. Fig. 9 describes the GO-CAM created for the data presented in Kim et al, 2009 [44]. In this article, GATA1 is the dbTF, SMARCA4 is the coTF, and HBB is the regulated gene product.

Globally speaking (see Table 3), in 60% of the cases the model permits the capture all the data presented in a selected article; there was data missing in 27% of the articles; in 6.7% of cases the article had more data than the GO-CAM, and the same percentage was obtained for the GO-CAM not being appropriate for the article.

4. Discussion

4.1. Research findings

Expressing gene regulation knowledge in a machine-processable manner is required for enabling advanced methodologies for gene regulation information and knowledge exploitation, including answering complex queries or testing biological hypotheses. In this work, we have contributed to the efforts of the GREEKC Consortium by investigating methods for facilitating the use of standardized knowledge to support data annotation processes aiming at building interoperable gene regulation content. More concretely, we have focused on studying how gene regulation knowledge could be formalized with the support of ontologies. Ontologies such as GO or SO have been successfully applied in the last twenty years for biological knowledge representation, management and interoperability, so they have constituted the ontological backbone of this effort.

This has followed a multimodal approach for the representation of biological knowledge, which was initially described in Vera-Ramos et al., 2019 [49]. Gene regulation scenarios represented as cartoons

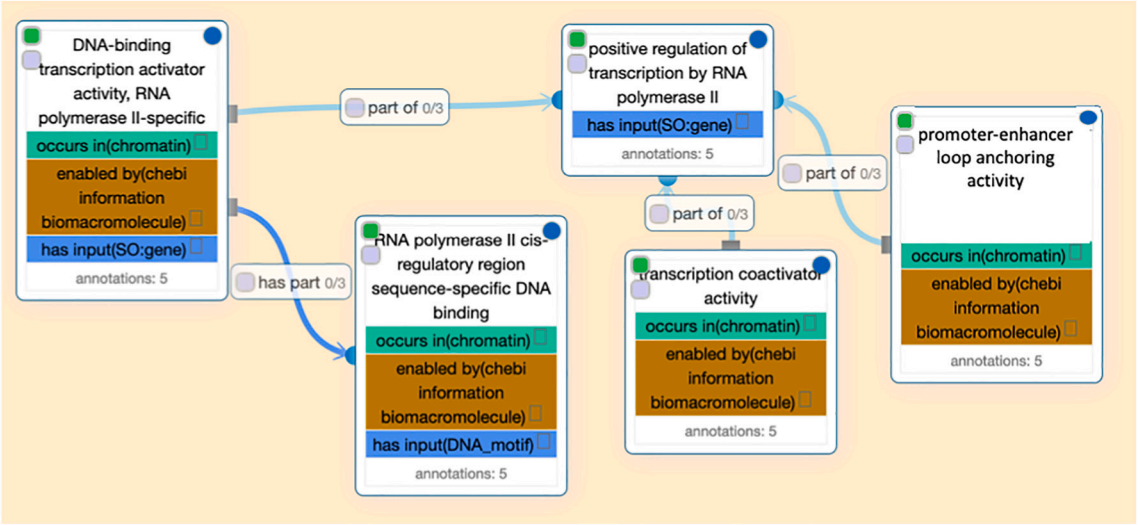

Fig. 8. GO-CAM representing gene regulation by regulators on an enhancer that loops to a promoter. The main process in this model is the *positive regulation of transcription by RNA polymerase II* (GO:0045944), which *has input* a *gene* (SO:0000704) product. Three activities are *part of* (BFO:0000050) this process, namely, *DNA-binding transcription activator activity* (GO:0001216), the *transcription co-activator activity, RNA polymerase II-specific* (GO:0001228), and *promoter-enhancer loop anchoring activity* (GO:0140585). The *DNA-binding transcription activator activity*, *occurs in* (BFO:0000066) the *chromatin*(GO:0000785), is *enabled by* (RO:0002333) a dbTF represented by *chebi information biomacromolecule* (CHEBI:33695), and *has input* (RO:0002233) a gene (SO:0000704) product. Besides, this activity *has part* (BFO:0000051) the *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (GO:0000978) which *occurs in* the *chromatin*, is *enabled by* the same dbTF as the previous activity, and *has input* the same gene product. The *transcription co-activator activity, RNA polymerase II-specific occurs in* the *chromatin*, and is *enabled by* the coTF represented by *chebi information biomacromolecule*. The *promoter-enhancer loop anchoring activity occurs in* the *chromatin*, and is *enabled by chebi information biomacromolecule*. It should be noted that in this model, the two gene products refer to the same gene product, whereas two transcription regulators are involved: the dbTF participates in the activities *DNA-binding transcription activator activity* and *RNA polymerase II cis-regulatory region sequence-specific DNA binding*, and the coTF participates in *transcription co-activator activity*.

Table 2
All the terms and relations reused in the second use case, their URI and source ontology.

Term	ID	Ontology
RNA polymerase II cis-regulatory region sequence-specific DNA binding	GO:0000978	Gene Ontology (GO)
DNA-binding transcription activator activity, RNA polymerase II-specific	GO:0001228	Gene Ontology (GO)
transcription coactivator activity	GO:0003713	Gene Ontology (GO)
promoter-enhancer loop anchoring activity	GO:0140585	Gene Ontology (GO)
positive regulation of transcription by RNA polymerase II	GO:0045944	Gene Ontology (GO)
chromatin	GO:0000785	Gene Ontology (GO)
information biomacromolecule	CHEBI:33695	ChEBI
gene	SO:0000704	Sequence Ontology (SO)
DNA_motif	SO:0000713	Sequence Ontology (SO)
has part	BFO:0000051	Relations Ontology (RO)
enabled by	RO:0002333	Relations Ontology (RO)
has input	RO:0002233	Relations Ontology (RO)
occurs in	BFO:0000066	Relations Ontology (RO)
part of	BFO:0000050	Relations Ontology (RO)

were the starting point for our approach, because cartoons enable an effective representation of the diversity and complexity of the gene regulation domain. The cartoons were developed by domain experts and they were analyzed together with ontology experts to generate a description in natural language. Such description was processed and interpreted by ontology experts to select and connect terms from existing ontologies in order to formally represent this meaning as knowledge models (RQ1).

One hypothesis was that reusing content from the major biological ontologies such as GO, SO, RO or ChEBI is a way of representing gene regulation knowledge in sufficient depth (RQ2). Our analysis of the completeness and correctness of the content and structure of those ontologies regarding gene regulation knowledge led to the identification of missing terms which were created as local terms in the Gene Regulation Application Ontology, GRAO. Their inclusion into GO and SO was requested, facilitated by members of the GO and SO consortia who participated in GREEKC and with whom new modeling approaches for gene regulation knowledge were shared and discussed. This allowed the alignment of our work and modeling approaches with the on-going actions in those consortia, which should contribute to the sustainability of the work. Examples of local terms in GRAO can be found in the Protégé models developed in this work.

The models created in this work and their application to support the annotation of research articles were evaluated by domain expert members of the GREEKC Consortium (RQ4). The results (see Table 3) show that the models were useful for annotating the content of the articles in 93% of the cases, so the models can be considered useful for data annotation. This result is positive but must be conservatively interpreted due to the limited sample size and because the articles were selected by one expert who considered that they were related to the use cases, therefore getting this high percentage was an expected result.

27% of the articles did not contain all the data required by the model. This implies that those articles do not provide a complete description of all regulation activity aspects represented in the model. This may occur because scientists have not been able to decipher completely those activities or because the authors chose not to include (experimental) investigations of those activities in the article. For the former case, the use of GO-CAMs as templates should specify if all data are compulsory. For the latter case, the use of GO-CAMs as templates should guide authors in data entry and help to prevent omissions in those cases in which

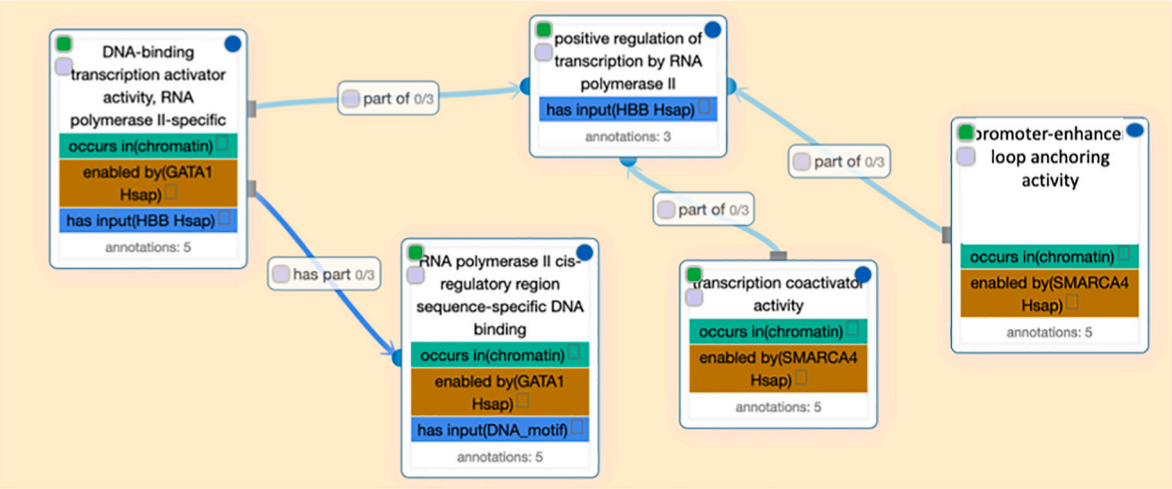

Fig. 9. The use of a GO-CAM template to curate experimental data. The GO-CAM describing gene regulation by regulators on an enhancer that loops to a promoter was used as a template to curate the data presented in Kim et al., 2009 [44].

Table 3
Summary of the results obtained in the evaluation carried out by the expert. Each row corresponds to a knowledge model and each column corresponds to one possible answer: A1: The GO-CAM is not appropriate for annotating this article; A2: The article does not provide all the data required by the GO-CAM; A3: The GO-CAM permits to capture all the data of the article; and A4: The article has relevant data which is not represented in the GO-CAM.

Model/Answer	A1	A2	A3	A4
Transcription regulation of a promoter (model 1)			4	
Transcription regulation of a promoter (model 2)	1			
Transcription regulation of a promoter (model 3)		2	1	
Transcription regulation of a promoter (model 4)		1		
Gene regulation by promoter-enhancer looping		1	1	1
	6.67%	26.67%	60%	6.67%

scientists know all the entities involved in the regulation activity. The use of the templates would not only help achieving consistency by professional data curators but also promote community curation efforts like the existing one for *S. pombe* [50].

One article contained more information than the model. In this case, it was due to the fact that two dbTFs were involved in the regulation of transcription and the corresponding model only provided the option to include one. This illustrates the need for creating additional models for these situations or to include cardinality constraints to GO-CAMs. Finally, one model was found not appropriate for the article because the proteins do not bind directly. This is a consequence of the complexity of the gene regulation domain, which will require a wider range of templates to cover the entire domain.

4.2. Noctua versus Protégé

In this work, gene regulation knowledge models were implemented using the state-of-the-art ontology editors Noctua and Protégé. Most researchers involved in this work had experience in knowledge representation based on ontologies and most of them had previous experience with Protégé but not with Noctua. As a result of this work, we considered that Noctua provides an environment which is more appropriate for creating these gene regulation knowledge models than Protégé, for a number of reasons that will be discussed next. Besides, we considered that GO-CAMs can be made that are appropriate for capturing gene regulation data (RQ3).

Nevertheless, we believe that both approaches could be useful for the objectives of GREEKC. Both tools allow for the creation of ontology-based knowledge models, which can be aligned with major biological ontologies and that, therefore, could provide the knowledge level required to support data representation and interoperability for complex molecular mechanisms of gene regulation.

In terms of semantics, both have in common the possibility of creating content by applying OWL semantics. In both tools, IRIs are used to identify each term included in the model. As mentioned, Noctua allows for creating models at the level of OWL's A-Box, and Protégé at both the A-Box and T-Box level.

GO-CAMs are models based on instances (A-Box models) instead of on classes (T-Box ones), so it could be thought that models based on classes can be used as templates in a more natural way, since the instantiation would consist on creating the instances of the corresponding classes. However, we are creating OWL models, which provides specific meaning and logical implications to instances and classes. OWL-based reasoning has different implications for modeling based on classes and based on instances. Axioms on OWL classes, so-called T-box axioms, are always universal statements on all instances of a class. Class-level statements hold universally, that is, for every instance of the class and in every context. For example, stating that a sequence is the promoter of a transcription activity would mean that that sequence would play the promoter role in every possible context for that transcription activity, which may not be true. Consequently, wrong inferences can be made by the reasoner. Modeling such role at the individual level does not have such unintended effect and allows for a more precise description of the biological context in which particular regulation activities occur. Gene regulation knowledge models are intended to describe scenarios in which specific genes or proteins participate (referring to the specific instances used in the experiments described in the article to be annotated), they do not intend to model scenarios in a universal way, since regulation activities happen in a particular context. Therefore, modeling based on instances is the appropriate option. Although there is some degree of universality in our models, we consider them as prototypical instances.

Noctua is more focused on the most specific terms that are part of the model, assuming that, given that they are created as instances of ontology classes, they are automatically linked to them. Protégé is more oriented to the construction of ontology T-boxes, i.e. the ontologies themselves and not their instances, where hierarchical organization and axiomatic structure are fundamental. From a usability point of view, A-

Box modeling is more user friendly in Noctua, which automatically creates and names every instance. Besides, Noctua's graphical interface is likely to be more user-friendly than Protégé GUI for domain, non-ontology experts. For example, the GO-CAMs are shown in a graphical way similar to how pathways are represented. The representation of the classes in Protégé is in the form of a hierarchy. In order to visualize the model graphically, it is necessary to use Protégé plug-ins or other tools.

One limitation of Noctua is that it has been developed for creating GO-CAM graphs, so it may not be useful for modeling other domains, in contrast, Protégé is generic and can be applied to a variety of domains. Another limitation of Noctua is that the user must select the content of the graph models from the Noctua Entity Ontology. In our use cases, we were not able to include some terms in our GO-CAMs because they were not available in the Noctua Entity Ontology, but this has to be regarded as tooling limitation that is easy to solve. On the other hand, Protégé supports the unrestricted creation of all kinds of OWL entities (classes, properties, individuals), subclass and equivalence statements, logical restriction and metadata annotations. Protégé facilitates the reuse of content from ontologies imported by the user. Protégé would also permit to include content from external databases such as UniProt or DbTF, which should be represented as OWL classes, since distinct instances of the same entity (e.g., protein) could be required for a GO-CAM. Hence, in case some terms cannot be included using Noctua, the GO-CAM can be exported in OWL format and modified using Protégé.

The GO-CAMs were presented in Section 2. As mentioned there, we also implemented those models using Protégé. Tables 4–8 (Supplementary Material) describe the content of the models obtained with Protégé. We can see that the Protégé models have more entities than the GO-CAMs. This is due to the fact that Protégé content is hierarchically organized and, therefore, we did not include only the most specific terms required for the use case, but also parent terms such as *Biological Process* or *Cellular Component* to provide a more detailed context for the model. It should be noted that, for simplicity, the Protégé models developed for the uses cases do not import the reused ontologies such as GO or SO, only the terms needed to build the models were included. In order to exploit the logical axioms of those ontologies, they should be explicitly imported. In the case of the GO-CAMs, their export in OWL format would import all the reused ontologies, which would make the semantic context available to the machine.

4.3. Practical implications

The work reflected in this manuscript is a proof-of-concept implementation of a method for the creation of knowledge models to support data curation processes. The tools used in this project are expected to be used by the developers of the knowledge models. Most tasks have been manually carried out in this study, because our main objective was to propose design patterns and evaluate their feasibility. A set of resources has been generated for each gene regulation scenario (cartoon, natural language specification, and knowledge models implemented in Noctua and Protégé), which could support data curators in different ways. For example, the cartoons could help data curators to select the right gene regulation scenario. Each set of resources should be offered as a research object [51]. We would expect curators use knowledge models that can be retrieved from the GO-CAM database[4] managed by the GO-CAM consortium, which should be included in the data curation tools in a way as transparent as possible for data curators. For example, annotation forms of the knowledge bases should be based on or mapped to the knowledge models and the curator should only fill the variable parts of the gene regulation scenario selected.In addition to this, since GO-CAMs permit defining the provenance for each activity unit, the content of one GO-CAM can be associated with multiple articles.

[4] http://geneontology.org/go-cam

4.4. Limitations and further work

The automation of the creation of GO-CAMs and the definition of reusable subunits of these models (those that specify a dbTF, coTF, gene or gene product) can be considered outside of the scope of this paper, but it should be the focus of future work. It would be desirable to have a language for the specification of the cartoons, since this would permit the development of methods for the extraction of information from them. This would allow for the automatic generation of the textual description or the direct extraction of entities to be searched in existing ontologies. Nevertheless, developing such a language seems difficult due to the intrinsic complexity of biology. The application of natural language processing tools for the automatic extraction of entities from the text descriptions should reduce the manual effort. Algorithms developed by the GREEKC working group on text mining could allow for identifying mentions of entities relevant for the gene regulation domain. Testing their effectiveness against general purpose text mining approaches would also be of interest. Finally, in this work, we have manually used search facilities offered by BioPortal or Ontobee for retrieving ontology content associated with the mentions of entities in the text. These and other APIs and alignment algorithms [52,53] could be integrated into an automatic pipeline.

We have mentioned that we want to use the knowledge models as templates for guiding data annotation. Nevertheless, GO-CAMs have not yet been designed to represent templates, and what has been illustrated here is the need for template models to include variables such as gene product. It might be interesting to extend the GO-CAM specification to permit the creation of templates. For examples, languages such as OPPL [54] have been applied for creating knowledge patterns for enriching the Gene Ontology [55].

Another limitation of this work is the number of gene regulation publications used for the validation of the models generated and the number of gene regulation scenarios itself. An automatic recognition of entities and their relations mentioned in domain texts would also facilitate the application of the methods to a large number of publication in order to instantiate more models, which would allow for validation on a larger scale. This would be a sensible next step prior to inclusion of the models into data annotation and curation pipelines.

5. Conclusions

In this work we have shown how gene regulation knowledge models can be generated for gene regulation scenarios by reusing existing ontologies and state-of-the-art semantic tools. Mostly, the methods applied have been manual, and their automation will be addressed in the future. We have been able to identify that some gene regulation knowledge is still missing in reference ontologies such as GO and SO. We have found that the two tools used in this work, Noctua and Protégé permit the creation of useful knowledge models.

Data statement

All data generated through is work is publicly accessible at https://github.com/jesualdotomasfernandezbreis/greekc.

CRediT authorship contribution statement

Belén Juanes Cortés: Investigation, Methodology, Software, Writing – original draft. **José Antonio Vera-Ramos:** Methodology, Software, Writing – review & editing. **Ruth C. Lovering:** Methodology, Validation, Writing – review & editing. **Pascale Gaudet:** Methodology, Validation, Writing – review & editing. **Astrid Laegreid:** Conceptualization, Validation, Writing – review & editing. **Colin Logie:** Validation, Writing – review & editing. **Stefan Schulz:** Conceptualization, Methodology, Writing – review & editing. **María del Mar Roldán-García:** Conceptualization, Methodology, Writing – review & editing. **Martin**

Kuiper: Conceptualization, Methodology, Funding acquisition, Supervision. **Jesualdo Tomás Fernández-Breis:** Conceptualization, Methodology, Investigation, Writing – original draft.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Acknowledgements

We thank the support of the Noctua development team, and the scientific exchange with the members of the GO, SO and GREEKC consortia.

Funding

This publication is based upon work from COST Action CA15205: GREEKC, supported by COST (European Cooperation in Science and Technology), and received funding from the Ministerio de Economía, Industria y Competitividad and the European Regional Development Fund [grant numbers TIN2017-86049-R, TIN2017-85949-C2-1-R]). RCL has been supported by Alzheimer's Research UK grant (ARUK-NAS2017A-1) and the National Institute for Health Research, University College London Hospitals, Biomedical Research Centre.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2021.194766.

References

[1] A. Venkatesan S. Tripathi A. S. de Galdeano W. Blondé A. Lægreid V. Mironov M. Kuiper , Finding gene regulatory network candidates using the gene expression knowledge base, BMC Bioinforma. 15 (1).
[2] A.H. Brivanlou, J.E. Darnell, Signal transduction and the control of gene expression, Science 295 (5556) (2002) 813–818.
[3] V.M. Weake, J.L. Workman, Inducible gene expression: diverse regulatory mechanisms, Nat. Rev. Genet. 11 (6) (2010) 426–437.
[4] P. Cramer, Organization and regulation of gene transcription, Nature 573 (7772) (2019) 45–54.
[5] L. Serebreni, A. Stark, Curr. Opin. Cell Biol. 70 (2021) 58–66.
[6] V. Perissi, K. Jepsen, C.K. Glass, M.G. Rosenfeld, Deconstructing repression: evolving models of co-repressor action, Nat. Rev. Genet. 11 (2) (2010) 109–123.
[7] T.I. Lee, R.A. Young, Transcriptional regulation and its misregulation in disease, Cell 152 (6) (2013) 1237–1251.
[8] J.L. Payne, F. Khalid, A. Wagner, Rna-mediated gene regulation is less evolvable than transcriptional regulation, Proc. Natl. Acad. Sci. 115 (15) (2018) E3481–E3490.
[9] E. Antezana, M. Egaña, W. Blondé, A. Illarramendi, I. Bilbao, B. De Baets, R. Stevens, V. Mironov, M. Kuiper, The cell cycle ontology: an application ontology for the representation and integrated analysis of the cell cycle process, Genome Biol. 10 (5) (2009) R58.
[10] O. Bodenreider, R. Stevens, Bio-ontologies: current trends and future directions, Brief. Bioinform. 7 (3) (2006) 256–274.
[11] D.L. Rubin, N.H. Shah, N.F. Noy, Biomedical ontologies: a functional perspective, Brief. Bioinform. 9 (1) (2008) 75–90.
[12] E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, M. Kuiper, Biogateway: a semantic systems biology tool for the life sciences 10 (S10) (2009) S11.
[13] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.
[14] The gene ontology resource: enriching a gold mine, Nucleic Acids Res. 49 (D1) (2021) D325–D334.
[15] T.R. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.
[16] GO Consortium, Expansion of the gene ontology knowledgebase and resources, Nucleic Acids Res. 45 (D1) (2017) D331–D338.
[17] D. Lonsdale, D.W. Embley, Y. Ding, L. Xu, M. Hepp, Reusing ontologies and language components for ontology generation, Data Knowl. Eng. 69 (4) (2010) 318–330.
[18] M.R. Kamdar, T. Tudorache, M.A. Musen, A systematic analysis of term reuse and term overlap across biomedical ontologies 8 (6) (2017) 853–871.
[19] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic Acids Res. 47 (D1) (2019) D330–D338.
[20] M.A. Musen, The protégé project: a look back and a look forward, AI Matters 1 (4) (2015) 4–12.
[21] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The sequence ontology: a tool for the unification of genome annotations, Genome Biol. 6 (5) (2005) R44.
[22] C.J. Mungall, C. Batchelor, K. Eilbeck, Evolution of the sequence ontology terms and relationships, J. Biomed. Inform. 44 (1) (2011) 87–93.
[23] Smith B., Kumar A., Bittner T., Basic formal ontology for bioinformatics. IFOMIS Reports, PhilArchive copy 1. https://philarchive.org/archive/KUMIRv1.
[24] R. Arp, B. Smith, A.D. Spear, Building Ontologies With Basic Formal Ontology, Mit Press, 2015.
[25] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, Nucleic Acids Res. 36 (suppl_1) (2007) D344–D350.
[26] E. Ong, Z. Xiang, B. Zhao, Y. Liu, Y. Lin, J. Zheng, C. Mungall, M. Courtot, A. Ruttenberg, Y. He, Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration, Nucleic Acids Res. 45 (D1) (2017) D347–D352.
[27] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic Acids Res. 39 (suppl 2) (2011) W541–W545.
[28] P.D. Thomas, D.P. Hill, H. Mi, D. Osumi-Sutherland, K. Van Auken, S. Carbon, J. P. Balhoff, L.-P. Albou, B. Good, P. Gaudet, S.E. Lewis, C.J. Mungall, Gene ontology causal activity modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems, Nat. Genet. 51 (10) (2019) 1429–1433.
[29] B. Smith, The logic of biological classification and the foundations of biomedical ontology, in: Invited Papers From the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain, 2003, pp. 19–25.
[30] B. Sterner, J. Witteveen, N. Franz, Coordinating dissent as an alternative to consensus classification: insights from systematics for bio-ontologies, Hist. Philos. Life Sci. 42 (1) (2020) 1–25.
[31] F. Reiter, S. Wienerroither, A. Stark, Combinatorial function of transcription factors and cofactors 43 (2017) 73–81.
[32] H.K. Long, S.L. Prescott, J. Wysocka, Ever-changing landscapes: transcriptional enhancers in development and evolution, Cell 167 (5) (2016) 1170–1187.
[33] Z.C. Poss, C.C. Ebmeier, D.J. Taatjes, The mediator complex and transcription regulation, Crit. Rev. Biochem. Mol. Biol. 48 (6) (2013) 575–608.
[34] B.L. Allen, D.J. Taatjes, The mediator complex: a central integrator of transcription, Nat. Rev. Mol. Cell Biol. 16 (3) (2015) 155–166.
[35] M.A. Zabidi, A. Stark, Regulatory enhancer–core-promoter communication via transcription factors and cofactors, Trends Genet. 32 (12) (2016) 801–814.
[36] H.-J. Kim, J.-Y. Kim, Y.-Y. Park, H.-S. Choi, Synergistic activation of the human orphan nuclear receptor shp gene promoter by basic helix–loop–helix protein e2a and orphan nuclear receptor sf-1, Nucleic Acids Res. 31 (23) (2003) 6860–6872.
[37] Xiao G., Jiang D., Ge C., Zhao Z., Lai Y., Boules H., Phimphilai M., Yang X., Karsenty G., Franceschi R. T., Cooperative interactions between activating transcription factor 4 and Runx2/Cbfa1 stimulate osteoblast-specific osteocalcin gene expression, Journal of Biological Chemistry 280 (35).
[38] V. Firlej, B. Bocquet, X. Desbiens, Y. De Launoit, A. Chotteau-Lelièvre, Pea3 transcription factor cooperates with usf-1 in regulation of the murine bax transcription without binding to an ets-binding site, J. Biol. Chem. 280 (2) (2005) 887–898.
[39] H. Jing, C.R. Vakoc, L. Ying, S. Mandat, H. Wang, X. Zheng, G.A. Blobel, Mol. Cell 29 (2) (2008) 232–242.
[40] Zhang F., Boothby M., T helper type 1-specific Brg1 recruitment and remodeling of nucleosomes positioned at the IFN- promoter are Stat4 dependent, Journal of Experimental Medicine 203 (6), doi:10.1084/jem.20060066.
[41] C. Chen, E.A. Rowell, R.M. Thomas, W.W. Hancock, A.D. Wells, Transcriptional regulation by foxp3 is associated with direct promoter occupancy and modulation of histone acetylation, J. Biol. Chem. 281 (48) (2006) 36828–36834.
[42] J.-W. Lee, D.-M. Kim, J.-W. Jang, T.-G. Park, S.-H. Song, Y.-S. Lee, X.-Z. Chi, I. Y. Park, J.-W. Hyun, Y. Ito, et al., Runx3 regulates cell cycle-dependent chromatin dynamics by functioning as a pioneer factor of the restriction-point, Nat. Commun. 10 (1) (2019) 1–17.
[43] Tang H., Sharp P. A., Transcriptional regulation of the murine 3' IgH enhancer by OCT-2, Immunity 11 (5), doi:10.1016/S1074-7613(00)80127-2.
[44] S.-I. Kim, S.J. Bultman, C.M. Kiefer, A. Dean, E.H. Bresnick, Brg1 requirement for long-range interaction of a locus control region with a downstream promoter, Proc. Natl. Acad. Sci. 106 (7) (2009) 2259–2264.
[45] F. Gong, L. Sun, Z. Wang, J. Shi, W. Li, S. Wang, X. Han, Y. Sun, The bcl2 gene is regulated by a special at-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-utr, Nucleic Acids Res. 39 (11) (2011) 4640–4652.
[46] X. Ren, R. Siegel, U. Kim, R.G. Roeder, Direct interactions of oca-b and tfii-i regulate immunoglobulin heavy-chain gene transcription by facilitating enhancer-promoter communication, Mol. Cell 42 (3) (2011) 342–355.
[47] J. Chaumeil, J.A. Skok, The role of ctcf in regulating v (d) j recombination, Curr. Opin. Immunol. 24 (2) (2012) 153–159.
[48] I. Krivega, A. Dean, Enhancer and promoter interactions—long distance calls 22 (2) (2012) 79–85.
[49] J.A. Vera-Ramos, B. Juanes-Cortés, J.T. Fernández-Breis, P. Gaudet, M. Kuiper, A. Lægreid, C. Logie, M.d.M Roldán-García, S. Schulz, An example of multimodal biological knowledge representation, in, JOWO (2019).

[50] A. Lock, M.A. Harris, K. Rutherford, J. Hayles, V. Wood, Community curation in pombase: enabling fission yeast experts to provide detailed, standardized, sharable annotation from research publications, Database 2020 (2020).
[51] S. Bechhofer, D.De Roure, M. Gamble, C. Goble, I. Buchan, Research objects: towards exchange and reuse of digital knowledge, Nat. Preced. (2010) 1, https://doi.org/10.1038/npre.2010.4626.1.
[52] M. Quesada-Martínez, E. Mikroyannidi, J.T. Fernández-Breis, R. Stevens, Approaching the axiomatic enrichment of the gene ontology from a lexical perspective, Artif. Intell. Med. 65 (1) (2015) 35–48.
[53] Zobolas J., Touré V., Kuiper M., Vercruysse S., UniBioDicts: Unified access to Biological Dictionaries, Bioinformatics 37 (1) 143–144.
[54] M.E. Aranguren, R. Stevens, E. Antezana, Transforming the axiomisation of ontologies: the ontology pre-processor language, Nat. Preced. (2009) 1, https://doi.org/10.1038/npre.2009.4006.1.
[55] J.T. Fernandez-Breis, L. Iannone, I. Palmisano, A.L. Rector, R. Stevens, Enriching the gene ontology via the dissection of labels using the ontology pre-processor language, in: in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2010, pp. 59–73.