An Example of Multimodal Biological Knowledge Representation

José Antonio VERA-RAMOS [a,1], Belén JUANES-CORTÉS [d],
Jesualdo Tomás FERNÁNDEZ-BREIS [d], Pascale GAUDET [g], Martin KUIPER [c],
Astrid LÆGREID [b], Colin LOGIE [f], María del Mar ROLDÁN-GARCÍA [e] and
Stefan SCHULZ [a]

[a] *Institute for Medical Informatics, Statistics and Documentation; Medical University of Graz, Austria*
[b] *Department of Clinical and Molecular Medicine; Norwegian University of Science and Technology, Norway*
[c] *Department of Biology; Norwegian University of Science and Technology, Norway*
[d] *Faculty of Computer Science; University of Murcia, Spain*
[e] *Faculty of Computer Science; University of Málaga, Spain*
[f] *Faculty of Science; Radboud Institute for Molecular Life Sciences, The Netherlands*
[g] *Swiss Institute of Bioinformatics, Switzerland*

Abstract. Biological knowledge evolves in a quick way whereas more people with other backgrounds (e.g., bioinformaticians, computer scientists, etc.) that help in biological research require detailed knowledge on biomolecular processes in order to understand the data they need to analyse. To solve this problem, in this paper we propose a multimodal knowledge representation using graphical diagrams representing biological knowledge, a natural language elucidation of the content of the graphical diagrams, linking the graphical elements to ontology instances; and a graph for visualising the ontology.

Keywords. Methodology, ontology, graphical diagrams, OWL, knowledge representation.

1. Introduction

Biological research produces highly diverse information types [1], and the overall volume of biological knowledge is rapidly increasing. This diversity and amount of data requires the semantic integration of information and knowledge, together with large datasets. Such integration and the application of advanced computing requires the cooperation of domain specialists with data scientists, bioinformaticians and computer scientists [2], who often lack basic knowledge of molecular biology, genomics and biochemistry. As a result, grasping sophisticated mechanisms (e.g., biochemical pathways or gene expression) requires new paths of knowledge standardisation, representation and visualisation. Such pieces of exchangeable and re-usable information about a specific

[1] Corresponding Author. jose.vera-ramos@medunigraz.at.

domain are also known as knowledge commons [3].

GREEKC [4] is a European network dedicated to the construction of high quality and interoperable knowledge commons covering the field of gene regulation. GREEKC aims to propose formally funded and interoperable Knowledge Representation (KR) models that can be easily employed and shared by all stakeholders of Life Science research. The educational and integrative aspects of the dissemination of these KR models suggest a multimodal approach, bringing together graphical representations, textual representations and formal-ontological representations. Such an approach should be standardised in terms of naming and definitions rooted in domain ontologies and languages (e.g., graph-based, logical (OWL), natural language) and requires a shared comprehension of the subject matter by all players in interdisciplinary teams because of the need to work with tightly interconnected data.

In order to satisfy these requirements, we propose a model that is ontology-based and follows a view on ontologies that emphasises them as artefacts for knowledge sharing within and across domains. Ontologies are therefore seen as formal descriptions of the characteristics of biological entities (e.g., molecules, organisms, cell components, processes, qualities, etc.) [5]. Our approach is inspired by principles formulated by the OBO Foundry [6], which recommends that domain ontologies be rooted in a foundational framework of basic categories and relationship types, which supports partitioning of domain ontologies (e.g., as done in the Gene Ontology [7][8]).

Bio-ontologies are normally restricted to T-Boxes, i.e., axiomatic descriptions of properties that universally hold for all particulars that instantiate a certain type (e.g., that all chromosomes are constituted by DNA). However, T-Boxes are neither sufficient in granularity and expressiveness nor appropriate to fulfil our educational goals. Traditionally, such information has been conveyed by texts and by graphical diagrams, albeit in a rather informal way. GREEKC proposes completing the picture by adding formal-ontological descriptions as a means to create and disseminate knowledge commons.

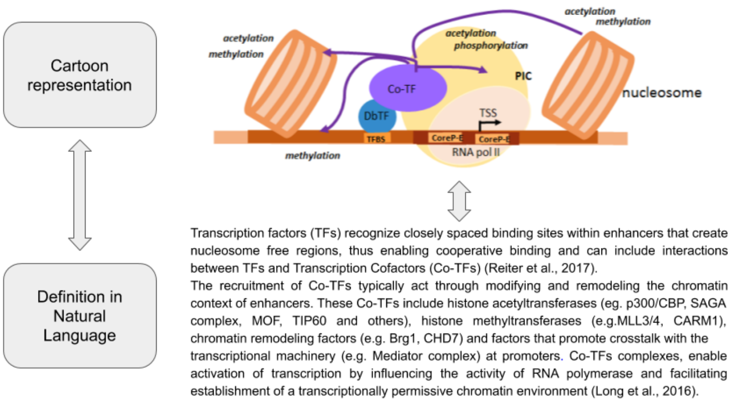
In this paper, we use the domain of gene regulation to describe an example of KR model that fulfils the above-mentioned requirements.

2. Methods

In order to build our model, we are assuming that a well-defined T-box exists, with universally agreed meanings (e.g., axioms) and sources such as domain ontologies connected to foundational ontologies as their building blocks. Once we have these blocks, the representation of prototypical examples, such as “transcription factor activity” from Gene Ontology is then expressed as A-box entities (prototypical instances) using: (i) elements of graphical diagrams, having appropriate labels and ideally having interactive functionality that links its graphical elements to the instances they represent in (ii) an ontology that provides universal descriptions in an OWL T-Box, instantiated by A-box entities and expressions that formally describe the processes depicted in the graphical diagram; (iii) a natural language elucidation of (i) (Figure 1); and (iv) a graph visualisation of (ii) (Figure 2).

For (i) we proposed a pre-existing prepared set of diagrams, i.e., graphical depictions of biological processes (Figure 1), supplied by the Norwegian GREEKC partner Astrid

Lægreid. In order to implement (ii) and (iv) we were using Noctua [9], a web-based tool used for the collaborative annotation of the activities that can be attributed to proteins in biological processes, based on A-box assertions. This tool produces so-called GO-CAM models, expressed as triplets (subject - predicate - object). Every model is a collection of triplets that describes broader biological processes.



Transcription factors (TFs) recognize closely spaced binding sites within enhancers that create nucleosome free regions, thus enabling cooperative binding and can include interactions between TFs and Transcription Cofactors (Co-TFs) (Reiter et al., 2017).
The recruitment of Co-TFs typically act through modifying and remodeling the chromatin context of enhancers. These Co-TFs include histone acetyltransferases (eg. p300/CBP, SAGA complex, MOF, TIP60 and others), histone methyltransferases (e.g.MLL3/4, CARM1), chromatin remodeling factors (e.g. Brg1, CHD7) and factors that promote crosstalk with the transcriptional machinery (e.g. Mediator complex) at promoters. Co-TFs complexes, enable activation of transcription by influencing the activity of RNA polymerase and facilitating establishment of a transcriptionally permissive chromatin environment (Long et al., 2016).

Figure 1. Example of a cartoon representation of epigenetic regulation of a promoter together linked to its definition written in Natural Language [10][11].

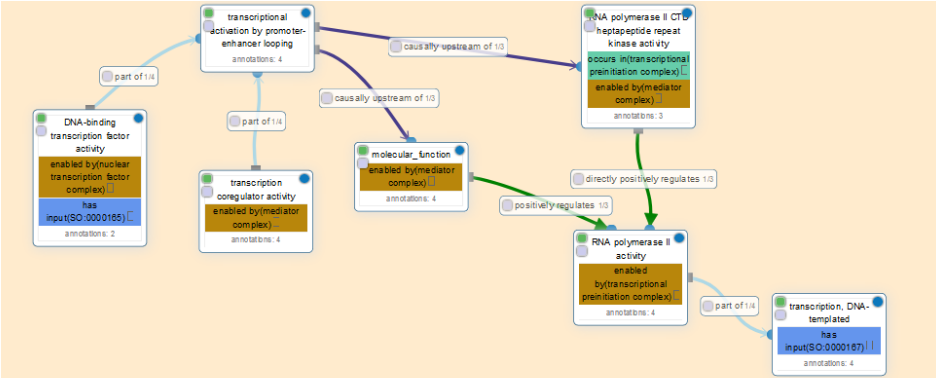

transcriptional activation by promoter-enhancer looping
annotations: 4
causally upstream of 1/3
causally upstream of 1/3
part of 1/4
part of 1/4
RNA polymerase II CTD heptapeptide repeat kinase activity
occurs in(transcriptional preinitiation complex)
enabled by(mediator complex)
annotations: 3
DNA-binding transcription factor activity
enabled by(nuclear transcription factor complex)
has input(SO:0000165)
annotations: 2
transcription coregulator activity
enabled by(mediator complex)
annotations: 4
molecular_function
enabled by(mediator complex)
annotations: 4
directly positively regulates 1/3
positively regulates 1/3
RNA polymerase II activity
enabled by(transcriptional preinitiation complex)
annotations: 4
part of 1/4
transcription, DNA-templated
has input(SO:0000167)
annotations: 4

Figure 2. Example of a graph visualisation of “transcription factor activity” using the tool Noctua.

Our Noctua models rely on the Gene Ontology (GO) as domain ontology. They are centred on a particular molecular activity class from the GO Molecular Function (MF) ontology, represented as a prototypical OWL instance. MF annotations are connected to provide the context in which that particular molecular function occurs. All connections within a GO-CAM model are relations as OWL object properties from the OBO Relations Ontology. GO-CAM models can be created using the graphic interface of the Noctua website. The first step to create the triplets was to analyse each statement in the textual description in order to extract the suitable GO terms by searched identifiers and keywords. Next, the relations between MF and other GO elements (precisely, instances of GO classes, particularly from the cellular component (CC) and biological process (BP) ontologies were added. Finally, instances from Sequence Ontology (SO) [12] classes were added and related to the MF instances.

3. Conclusion and Outlook

We proposed a way to build an educational knowledge representation artefact that helps people working with biological data to understand the interconnected nature of biological molecules and processes. This support is even more relevant to people that lack basic biological knowledge. This constitutes a work in progress and its advance can be monitored on the GREEKC website. Therefore, the next step is implementing such a KR model since it will be of great value for the community.

References

[1] Robert A Wallace, Gerald P Sanders, and Robert J Ferl. *Biology, the science of life*. HarperCollins New York, 1996.

[2] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[3] Charlotte Hess and Elinor Ostrom. Introduction: An overview of the knowledge commons. 2007.

[4] GREEKC. `http://greekc.org/`. Accessed: 14-6-2019.

[5] Peter D Karp. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16 (3):269–285, 2000.

[6] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251, 2007.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[8] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2018.

[9] Noctua. `http://noctua.berkeleybop.org/`. Accessed: 31-5-2019.

[10] Jeremy F Reiter and Michel R Leroux. Genes and molecular pathways underpinning ciliopathies. *Nature Reviews Molecular Cell Biology*, 18(9):533, 2017.

[11] Hannah K Long, Sara L Prescott, and Joanna Wysocka. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, 2016.

[12] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44, 2005.