Towards the Prokaryotic Regulation Ontology: An Ontological Model to Infer Gene Regulation Physiology from Mechanisms in Bacteria

Citlalli Mejía-Almonte[a] and Julio Collado-Vides[b]
Center for Genomic Science, UNAM, Av. Universidad, Cuernavaca, Mexico

Keywords: Formal Ontology, Domain Ontology, Gene Regulation, Bacteria.

Abstract: Here we present a formal ontological model that explicitly represents regulatory interactions among the main objects involved in transcriptional regulation in bacteria. These formal relations allow the inference of gene regulation physiology from gene regulation mechanisms. The automatically instantiated classes can be used to assist in the mechanistic interpretation of gene expression experiments done at the physiological level, such as RNA-seq. This is the first step to develop a more comprehensive ontology focused on prokaryotic gene regulation. The ontology is available at https://github.com/prokaryotic-regulation-ontology

1 INTRODUCTION

Since the success shown by the Gene Ontology as a controlled vocabulary, bio-ontologies are increasingly important tools in bio-informatics. However, little has been explored regarding formal ontological representation in the domain of bacterial gene regulation. There are two granularity levels at which gene regulation can be studied. At the physiological level, transcript concentration or gene product activity is directly measured under some condition, normally adding or depleting certain chemicals to growth media (Burstein et al., 1965). At the mechanistic level, the effect of specific mutations on gene expression is studied to discover the precise regulators involved in some system (Ptashne, 1967). At this level, the most studied mechanisms are those of transcription initiation mediated by transcription factors. These proteins can adjust gene expression to environmental requirements using their two main functional domains: the effector-binding domain that senses the environmental signal and the DNA binding domain. Transcription factors bind to DNA in sites called transcription factor binding sites, thereby increasing or decreasing the activity of a promoter. Promoters are the DNA regions where transcription of transcription units (TUs) begins; TUs in turn contain one or more genes. Therefore, the expression of a TU is regulated by regulation of the promoter activity. Here, we develop an ontological model that can infer the physiology from mechanisms of gene regulation.

The result of transcriptome analysis are sets of genes that are either underexpressed or overexpressed under a given condition, including the addition of chemicals to growth media. The observation of underexpression corresponds to the observation of gene inhibition, whereas the observation of overexpression corresponds to the observation of gene induction. This means that transcriptome analysis gives us physiological insights, rather than mechanistic ones. The model presented here, automatically instantiates sets of genes that are induced or repressed by some molecule based on the mechanisms of induction or repression. The final terms will encode both the physiology and the mechanisms of gene regulation (see below). Thus, this ontology can help in the mechanistic interpretation of gene expression experiments that are done at the physiological level, such as transcriptome analysis.

No ontology explicitly states the aim of modeling gene regulation in the obo-foundry repository (Smith et al., 2007); whereas a search in bioportal (Noy et al., 2009; Noy et al., 2001) only retrieves the Gene Regulation Ontology (GRO) (Beisswanger et al., 2008). This ontology includes object properties to define *agents* and *patients* of regulation, but it focuses on the mechanistic description of gene regulation and it does not distinguishes the two granularity levels of gene regulation described in this paper. Thus, here we develop an ontology to represent both mechanisms and physiology of gene regulation, the later inferred from the former.

[a] https://orcid.org/0000-0002-0142-5591
[b] https://orcid.org/0000-0001-8780-7664

Mejía-Almonte, C. and Collado-Vides, J.
Towards the Prokaryotic Regulation Ontology: An Ontological Model to Infer Gene Regulation Physiology from Mechanisms in Bacteria.
DOI: 10.5220/0008387804950499
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 495-499
ISBN: 978-989-758-382-7

2 DEVELOPMENT PROCESS

We are using top-down ontology development approach (Noy et al., 2001). First, we included the most general and important entities involved in regulation of transcription initiation: transcription factor, transcription factor binding site, promoter, transcription unit, effector, etc. Second, we included the corresponding biological relations among them. Third, we created the classes that will be automatically instantiated: *TF bound to the DNA* and *regulated system*. Fourth, we formally defined these classes taking advantage of the biologically relations included in the second step (figure 2). Lastly, most specific terms have to be generated for each specific TF, TU, promoter, etc. along with their relations. The model will automatically classify these specific entities into the defined classes (see glycolate example). RegulonDB can be used to instantiate the ontology with knowledge about *Escherichia coli K-12* (Santos-Zavaleta et al., 2018).

We are following the OBO-foundry principles. For this, we are taking advantage of the OBO tools ROBOT (Overton et al., 2015) and the Ontology Development Kit (https://github.com/INCATools/ontology-development-kit). The first one is mainly used to extract terms and modules from existing ontologies, while the later is designed for standardized ontology documentation and release of OBO ontologies, taking care of quality control issues. We are using the Basic Formal Ontology as upper-level ontology. So far, we have reused terms from six OBO-foundry ontologies: CHEBI, GO, MSO, NCIT, OGG, and SO (Ashburner et al., 2000; de Matos et al., 2010; Mungall et al., 2011; Sioutos et al., 2007; He et al., 2014) The creation of new classes and axioms was done using Protégé version 5.5. (Musen et al., 2015)

3 MODEL DESCRIPTION

In this paper, classes are written in italics and object properties are written in bold face. Hierarchy is represented as indentation of bulleted lists.

3.1 An n-ary Relation to Represent the Central Transcriptional Regulatory Interaction

Figure 1 depicts the main elements involved in transcriptional regulation along with the relations that exist among them. These were ontologically represented as follows. *Transcription factor* (TF), *TF binding site* (TFBS), *effector*, and *functional conformation* classes were created. Then, an n-ary relation design pattern was used to link these four elements (Noy and Rector, 2004). **TF bound to TFBS** class was created with four properties: **has binding transcription factor**, **has target TFBS**, **is realized in functional conformation**, and **has effector** (Figure 1).

3.2 A Property Chain to Infer Regulation from Anatomy

Figure 1 also depicts how the two key relations that distinguish physiology from mechanisms of transcriptional regulation were ontologically represented. The mechanistic level describes the direct effect that a TF bound to a TFBS has over its cognate promoter, while the physiological level describes the effect that the environmental condition (in our current model represented by the effector molecule) has over the expression of genes in a transcription unit. *Promoter* and *transcription unit* classes were created. Then transcription unit was related with promoter using the property **is transcribed from**, whereas promoter was related to the class *TF bound to TFBS* with the property **has activity regulated by**. The **has expression regulated by** property was created along with the following rule chain expressed in functional syntax (Figure 1) (Hitzler et al., 2009):

```
SubObjectPropertyOf(
   ObjectPropertyChain( :is transcribed from
   :has activity regulated by )
   :has expression regulated by
 )
```

This rule chain represents the fact that if a TU is transcribed from a promoter, and this promoter has its activity regulated by a TF bound to a TFBS, then this TF bound to a TFBS regulates the expression of the TU.

3.3 Automatic Classification of Regulated Systems

At the physiological level, there are only two possibilities: induction or inhibition of gene expression. At the mechanistic level, there are four possibilities. Transcription factors bind to their cognate TFBSs and regulate transcription only when they are in functional conformation. Induction can be achieved by activation when the binding of the effector activates a transcription factor that increases the expression of a TU (active conformation of TF is holo), or by derepression when the binding of the effector deactivates a transcription factor that decreases the expression of a TU (active conformation of TF is apo). Inhibition can be achieved by repression when the bind-

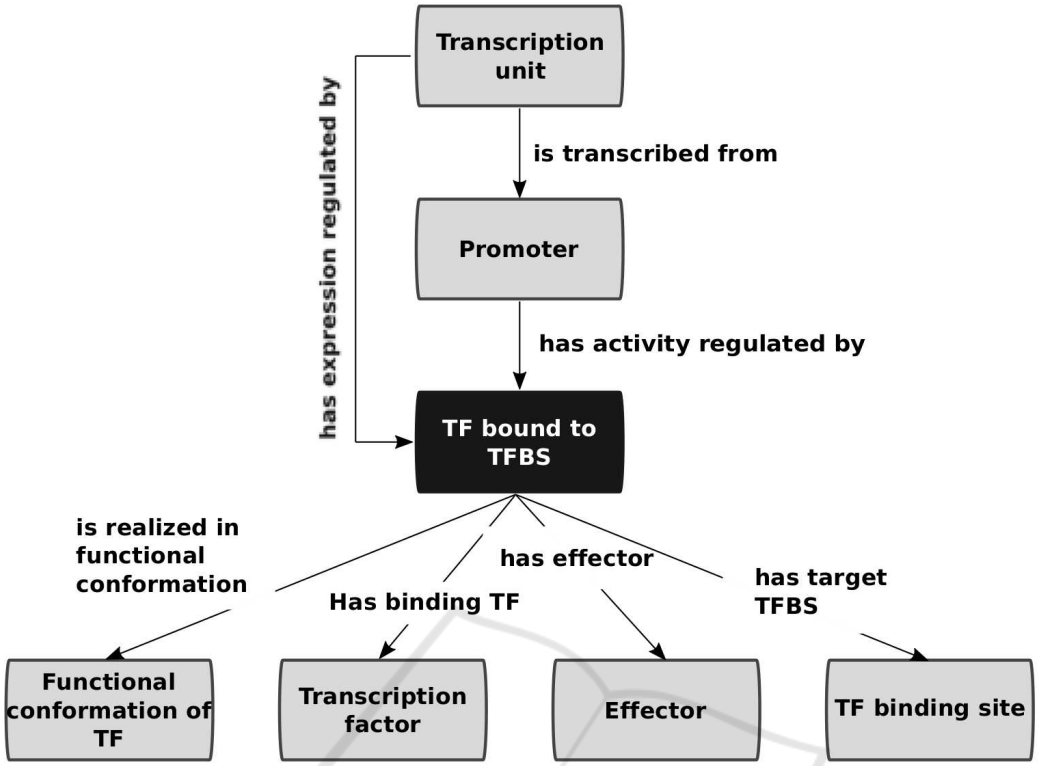

Figure 1: An n-ary relation and a property chain to represent the central regulatory interaction.

ing of the effector activates a transcription factor that decreases the expression of a TU (active conformation is holo), or by de-activation when the binding of the effector deactivates a transcription factor that increases the expression of a TU (active conformation is apo) (Balderas-Martínez et al., 2013). All of these cases describe the physiological response to the appearance of the effector. The disappearance of the effector reverses the response. We will treat these cases later.

Therefore, to automatically classify TUs that are induced or inhibited by an effector we have have created the following subclasses of *TF bound to TFBS* (Figure 2). Equivalent class axioms are shown.

- *TF bound to TFBS in apo conformation* **is realized in functional conformation** some *apo functional conformation of TF*
  - *TF-glycolate active in apo* **has effector** some *glycolate*
- *TF bound to TFBS in holo conformation* **is realized in functional conformation** some *holo functional conformation of TF*
  - *TF-glycolate active in holo* **has effector** some *glycolate*

The classes *inducible system* and *inhibitable system* were created with the following subclasses. Equivalent class axioms are shown.

- *System induced by activation* **has expression increased by** some *transcription factor bound to TFBS in holo conformation*
  - *System induced by activation by glycolate* **has expression increased by** some *TF-glycolate active in holo*
- *System induced by derepression* **has expression decreased by** some *transcription factor bound to TFBS in apo conformation*
  - *System induced by derepression by glycolate* **has expression decreased by** some *TF-glycolate active in apo*
- *System inhibited by repression* **has expression decreased by** some *transcription factor bound to TFBS in holo conformation*
  - *System inhibited by repression by glycolate* **has expression decreased by** some *TF-glycolate active in holo*

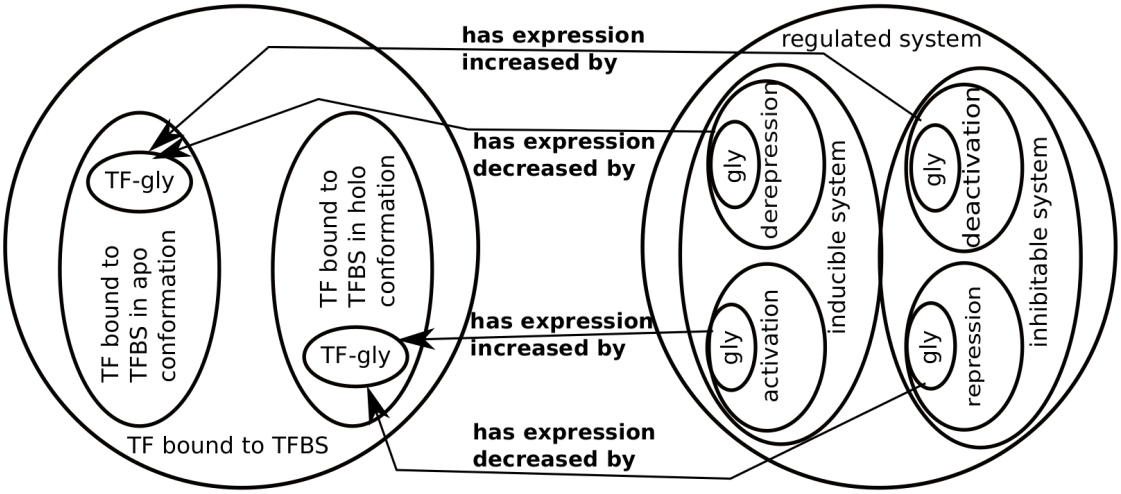

Figure 2: Defined classes to infer physiology from mechanisms. The outer circles represent the most general classes and inner ovals more specific classes. On the left, the hierarchy of the molecular complex TF-TFBS-effector classes is shown. In the text, the most specific classes are named *TF-glycolate active in holo* and *TF-glycolate active in apo*; in the figure, the terms were shortened as TF-gly due to space issues. These classes are automatically instantiated due to the n-ary relation shown in Figure 1. On the right, the hierarchy of effector-induced or effector-repressed systems are shown. The terms were shortened due to space issues: activation is short for *system induced by activation*, derepression is short for *system induced by derepression*, deactivation is short for *system inhibited by deactivation*, and repression is short for *system inhibited by repression*, whereas gly is short for *system induced by activation by glycolate*, *system induced by derepression by glycolate*, *system inhibited by deactivation by glycolate*, and *system inhibited by repression by glycolate*, depending on the superclass. These classes can be automatically instantiated due to the property chain shown in Figure 1.

- *System inhibited by deactivation* **has expression increased by** some *transcription factor bound to TFBS in apo conformation*
  - *System inhibited by deactivation by glycolate* **has expression increased by** some *TF-glycolate active in apo*

In this listing of formal definitions, we included only examples of classes defined by the specific effector glycolate. The final ontology have to be extended to include classes for all known effectors. We plan to do this extension using *E. coli* information retrieved from RegulonDB.

4 CONCLUSIONS

An ontological model that can automatically classify transcription units as effector-dependent repressible or inducible systems was developed. This adds a layer of formal knowledge to the mechanistic representation of bacterial gene regulation included in databases like RegulonDB.

ACKNOWLEDGEMENTS

C.M.A. is a Ph.D. student from the Programa de Doctorado en Ciencias Biomedicas, Universidad Nacional Autonoma de Mexico, receives fellowship 576333 from CONACYT and received financial aid from Programa de Apoyos para Estudios de Posgrado (PAEP) for this conference. JCV acknowledges support by UNAM and by NIH-NIGMS grant RO1-GM110597.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Balderas-Martínez, Y. I., Savageau, M., Salgado, H., Pérez-Rueda, E., Morett, E., and Collado-Vides, J. (2013). Transcription factors in escherichia coli prefer the holo conformation. *PloS one*, 8(6):e65723.

Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U., et al. (2008). Gene regulation ontology (gro): design principles and use cases. In *MIE*, pages 9–14.

Burstein, C., Cohn, M., Kepes, A., and Monod, J. (1965). Role du lactose et de ses produits metaboliques dans

l'induction de l'operon lactose chez escherichia coli. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis*, 95(4):634–639.

de Matos, P., Dekker, A., Ennis, M., Hastings, J., Haug, K., Turner, S., and Steinbeck, C. (2010). Chebi: a chemistry ontology and database. *Journal of cheminformatics*, 2(1):P6.

He, Y., Liu, Y., and Zhao, B. (2014). Ogg: a biological ontology for representing genes and genomes in specific organisms. In *ICBO*, pages 13–20. Citeseer.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). Owl 2 web ontology language primer. *W3C recommendation*, 27(1):123.

Mungall, C. J., Batchelor, C., and Eilbeck, K. (2011). Evolution of the sequence ontology terms and relationships. *Journal of biomedical informatics*, 44(1):87–93.

Musen, M. A. et al. (2015). The protégé project: a look back and a look forward. *AI matters*, 1(4):4.

Noy, N. and Rector, A. (2004). Defining n-ary relations on the semantic web: Use with individuals. *W3C Working Draft*, 21:102.

Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173.

Overton, J. A., Dietze, H., Essaid, S., Osumi-Sutherland, D., and Mungall, C. J. (2015). Robot: A command-line tool for ontology development. In *ICBO*.

Ptashne, M. (1967). Specific binding of the λ phage repressor to λ dna. *Nature*, 214(5085):232.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P., et al. (2018). Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. *Nucleic acids research*, 47(D1):D212–D220.

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L., and Wright, L. W. (2007). Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251.