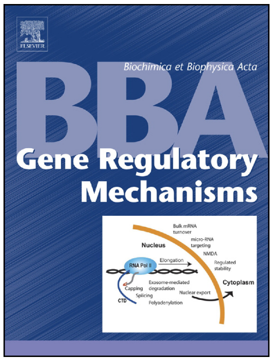


Sequence ontology terminology for gene regulation

David W. Sant, Michael Sinclair, Christopher J. Mungall, Stefan Schulz, Daniel Zerbino, Ruth C. Lovering, Colin Logie, Karen Eilbeck

PII: S1874-9399(21)00063-8

DOI: https://doi.org/10.1016/j.bbagrm.2021.194745

Reference: BBAGRM 194745

To appear in: *BBA - Gene Regulatory Mechanisms*

Received date: 30 January 2021

Revised date: 17 July 2021

Accepted date: 5 August 2021

Sequence Ontology terminology for gene regulation

Authors: David W Sant, Michael Sinclair, Christopher J Mungall, Stefan Schulz, Daniel Zerbino, Ruth C Lovering, Colin Logie, Karen Eilbeck

Addresses

DWS: Department of biomedical informatics, University of Utah, Salt Lake City, Utah, USA; Department of Biomedical Sciences, Noorda College of Osteopathic Medicine, Provo, Utah, USA. ORCID: 0000-0001-7372-9896. Email: david.sant@utah.edu

MS: Department of biomedical informatics, University of Utah, Salt Lake City, Utah, USA. ORCID: 0000-0002-3636-5920. Email: singla@gmail.com

CJM: Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory: Berkeley, CA, US. ORCID: 0000-0002-6601-2165. Email cjmungall@lbl.gov

SS: Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria. ORCID: 0000-0001-7222-3287 Email: stefan.schulz@medunigraz.at

DZ: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK, ORCID: 0000-0001-5350-3056. Email: zerbino@ebi.ac.uk

RCL: Functional Gene Annotation, Preclinical and Fundamental Science, UCL Institute of Cardiovascular Science, University College London, London, UK. ORCID: 0000-0002-9791-0064, email: r.lovering@ucl.ac.uk

CL: Radboud Institute for Molecular Life Sciences, Geert Grooteplein Zuid 28, 6525 GA Nijmegen, Netherlands. OrcID 0000-0002-8534-6582. Email C.Logie@ncmls.ru.nl

KE: Department of biomedical informatics, University of Utah, Salt Lake City, Utah, USA. ORCID:0000-0002-0831-6427, email keilbeck@genetics.utah.edu

Abstract

The Sequence Ontology (SO) is a structured, controlled vocabulary that provides terms and definitions for genomic annotation. The Gene Regulation Ensemble Effort for the Knowledge Commons (GREEKC) initiative has gathered input from many groups of researchers, including the SO, the Gene Ontology (GO), and gene regulation experts, with the goal of curating information about how gene expression is regulated at the molecular level. Here we discuss recent updates to the SO reflecting current knowledge. We have developed more accurate human-readable terms (also known as classes), including new definitions, and relationships related to the expression of genes. New findings continue to give us insight into the biology of gene regulation, including the order of events, and participants in those events. These updates to the SO support logical reasoning with the current understanding of gene expression regulation at the molecular level.

Introduction

With the rapid increase in genomic sequencing across a multitude of species came the need to automate the annotation of genetically encoded sequences. Defining the parts of a genome was key to the unification of the description of genomes across species. To address this issue, the Sequence Ontology was created by the Gene Ontology Consortium to be a structured controlled vocabulary for the definition of biological sequence features[1,2]. The SO is one of the original members of the OBO Foundry[3] (http://www.obofoundry.org/) and is interoperable with other ontologies such as the Gene Ontology[4] (GO) and Chemical Entities of Biological Interest (CHEBI)[5]. Terms (also known as classes in OWL) and relationships between them are added or updated as the team members become aware of new understanding or findings in the field.

The terminology and definitions in the field of gene regulation are intricate. The Sequence Ontology covers the technical language elements necessary to denote the genomic regions involved in regulatory processes. Other aspects of gene regulation are covered by other ontologies such as the GO. The Gene Regulation Ensemble Effort for the Knowledge Commons (GREEKC) initiative was established in 2016 as the European branch of the Gene Regulation Consortium (GRECO) to enable multiple groups of researchers to determine how to accurately represent knowledge of the regulation of gene expression at the molecular level using several ontologies (http://www.greekc.org/, http://thegreco.org). The collaborative nature of this project has allowed for terms across databases to be updated concurrently, ensuring the interoperability of the ontologies. In this manuscript, we report updates related to gene regulation that have been made to SO as part of GREEKC. When discussing ontology terms in this manuscript, we will italicize and space with underscores to differentiate from the discussion of biological entities.

The Scope of Sequence Ontology

SO was initially developed by the Gene Ontology Consortium, although the scope of SO differed significantly from the Gene Ontology (GO). While GO describes the outward face of gene products - what they do and where they do it, the SO defines the internal parts of genes and genomes. Genomic annotations and gene models are the parts of genomes demarcated in coordinate space (e.g.; chromosome: start-end). They define where the exons, introns, and Transcription Start Sites (TSS) etc. begin and end. Representing the coordinates of features themselves is outside of SO's scope, but has been accomplished by other groups such as FALDO and Biolink[6,7]. Development of SO terminology is a result of curators' needs to adequately define the parts of their genomic annotations. The updates that have occurred concurrently with the GREEKC initiative are related to nucleotide sequences that are important for different aspects of gene regulation, not the proteins that produce actions at a molecular level. Here we describe the resultant terminology for describing the genomic features involved in gene regulation.

The Understanding of *Cis*-Regulatory Modules (CRM)

The correct spatial and temporal expression of genes is required for multicellular organisms to develop properly and maintain the different necessary cell types[8]. Many DNA elements act in tandem to regulate the expression of a specific gene or a set of genes, and these DNA elements are typically clustered into regions commonly referred to as *cis*-regulatory modules (CRMs). The SO definition for *CRM* (SO:0000727) is "A regulatory region where transcription factor binding sites are clustered to regulate various aspects of transcription activities. (CRMs can be located a few kilobases (kb) to hundreds of kb upstream of the basal promoter, in the coding sequence, within introns, or in the untranslated regions (UTR) sequences, and even on a different chromosome). A single gene can be regulated by multiple CRMs to give precise control of its spatial and temporal expression. CRMs function as nodes in a large, intertwined regulatory network." In short, a CRM is a region of DNA that contains multiple elements that regulate the expression of genes.

While CRMs contain multiple regulatory elements such as transcription factor binding sites, different CRMs have different functions for the regulation of transcription. These different CRMs include enhancers, silencers, locus control regions, and insulators (see Figure). Enhancers are CRMs that activate the expression of their target regardless of orientation and may be distant from the promoter region. Silencers are essentially the opposite of enhancers and function to suppress transcription. Insulators are CRMs that function to prevent another CRM from interacting with the promoter of a nearby gene when the insulator is located between two CRMs. Locus control regions are open chromatin (DNAse hypersensitive) regions of DNA that confer high-level, copy number dependent expression of a gene[9]. A CRM term that has recently been added to SO is *DNA_loop_anchor*, representing the ends of a DNA looping region. This DNA looping allows for areas of DNA that are very distant to remain in close proximity within the cell, allowing for CRMs to interact with distant genes[10].

As noted in the *CRM* definition, a single gene may be regulated by multiple CRMs and the regulation of expression of a single gene can be very complex. For example, a single gene may be active only when an enhancer region is active, which in turn inactivates a silencer and activates the promoter region of the gene[11,12].

While databases of genes and proteins have been around for decades, the annotation of CRMs in most species aside from yeast and some particular bacteria has lagged, largely due to the difficulty of detecting them[13-16]. Just as gene expression is variable across cell types and conditions, CRMs may be active only in certain cell types and conditions. This would indicate that detecting all CRMs would require analysis using all cell types. Advancements in sequencing technologies have aided greatly in the detection of CRMs, especially the use of chromatin immunoprecipitation sequencing (ChIP-seq) to detect specific chromatin marks or the binding of specific proteins to DNA. For example, active enhancers are detected by the presence of histone 3 lysine 4 mono-methylation (H3K4me1) and acetylated lysine 27 of histone 3 (H3K27ac), while poised enhancers are repressed by trimethylated lysine at position 9 of histone 3 (H3K9me3) and/or trimethylated lysine at position 27 of histone 3 (H3K27me3)[17]. Silencers are typically marked by the binding of polycomb repressive complex 1 or 2 (PRC1/2) and H3K27me3[18]. It should be noted that many high-throughput experiments like ChIP-seq provide a starting point for understanding gene regulation, in this case with elucidation of transcription factor binding sites, additional experiments are required to prove the role in the regulation of a gene. The annotation of such genomic features must take into account the level of evidence that supports the role, such as predicted versus validated. This level of belief is not currently articulated in the ontology and therefore should be expressed in the annotation.

A promoter, like a CRM, is a *transcriptional_cis_regulatory_region* (see Figure). Some promoters are characterized by their expression pattern. Constitutive promoters are promoters that have continual transcription. Inducible promoters are those that can be induced for transcription by the presence of a factor. Cryptic promoters are promoters to a cryptic gene, which is a gene that is not transcribed under normal conditions and is not critical to normal cellular function. Bidirectional promoters are promoters that can initiate transcription in either direction[19].

Promoters for DNA template-dependent RNA polymerases have somewhat different structures within different types of organisms. Eukaryotic promoters include a TSS and serve as a region for the assembly of a pre-initiation complex (PIC), which is necessary for transcription of the gene. Prokaryotic promoters are regions of binding of a specific RNA polymerase (RNA pol) holoenzyme, which may lead to the transcription of multiple genes[20]. Prokaryotic promoters include bacterial RNA promoters. Viral promoters include Phage RNA polymerase promoters and they contain the TSS of the gene and will be bound by host machinery that varies with the host species. Eukaryotic promoters include RNA pol I, II, and III promoters, and in plants RNA pol IV and V promoters.

We recently introduced a new term *core_promoter_element* (SO:0002309), defined as “An element that exists within the promoter region of a gene. When multiple transcripts exist for a gene, the separate transcripts may have separate core_promoter_elements." The term *core_promoter_element* is further subdivided into subclasses *core_eukaryotic_promoter_element*, *core_prokaryotic_promoter_element* and *core_viral_promoter_element*. In SO the components of a *core_eukaryotic_promoter* are elements that are found within the promoter region of a eukaryotic gene, which indicates that they may be present in RNA pol I, II, or III promoters. These elements include well-known elements such as the *TATA_box* (SO:0000174) and *discontinuous_core_element* (DCE, SO:0001664).

While the parts of bacterial promoters have been described as child terms to bacterial_RNApol_promoter_sigma54_element (minus_12_signal and minus_24_signal) and bacterial_RNApol_promoter_sigma_70_element (minus_35_signal and minus_10_signal), recent work now suggests that these motifs are not considered an essential component of bacterial promoters[20]. This is due to recent changes in understanding about bacterial gene regulation whereby these motifs are not necessary in some instances and not sufficient in other instances to promote transcription. Therefore, the term core_prokarytotic_promoter_element currently does not contain motif sequence parts.

In summary, the restructuring of terms under the CRM branch in SO has allowed for a more accurate structuring of terms related to CRM. In particular, the term *core_promoter_element* has been created and the component parts of this region can now be annotated with specific core promoter element SO terms that include general transcription initiation factor binding sites that are distinct from sequence-specific DNA binding transcription factor binding sites (Gaudet *et al.* in preparation)[21].

Topologically Defined Regions (TDRs) and Topologically Associated Domains (TADs)

In order for DNA elements to be active and contribute toward gene transcription, the elements must be in regions of open euchromatin[22]. Some CRMs, such as enhancers, regulate genes located over 100 kb away. An open stretch of DNA 100 kb in length would account for more than 30 ⌠M of distance[10]. If all euchromatin existed as free-flowing DNA, it would be highly unlikely that the enhancer region would ever interact with the promoter region of a gene to increase transcription. This is why DNA within the cell remains in chromatin loops. The ends of the loops are held in close proximity, promoting physical interaction of the elements on either end of the loop. Furthermore, all the DNA within such chromatin loops appears to self-associate efficiently[23]. The entire region between these interacting ends has therefore been called a *topologically_associated_domain* (TADs, SO:0002304). Several technologies have emerged over recent years to allow for the identification of TADs, including chromatin conformation capture (3C) and related technologies 4-C, 5-C, GCC, Hi-C, ChIA-PET and GAM[24]. An area where self-interaction occurs more frequently than expected by chance is known as a *topologically_defined_region* (TDRs, SO:0001412).

TADs are flanked by a *topologically_associated_domain_boundary* (TAD boundary) (SO:0002305) on both sides. The DNA inside a TAD can form a *DNA_loop* (SO:0002307). The term *DNA_loop* refers to the phenomenon of loop formation of DNA. Importantly, loops are molecular *conformations*, and as such they are continuants, i.e. static entities, opposed to intrinsically related looping *processes* displayed by a chromosome in a cell at one specific time are occurrents, since processes are always occurrents (https://en.wikipedia.org/wiki/Basic_Formal_Ontology)[25]. The TAD and the TAD boundary are continuants that respectively refer to the regions of DNA that self-associate frequently and to the regions across which chromatin loops occur infrequently. A DNA loop anchor will usually occur at the TAD boundary where the ends of the loop are held in close proximity, but a majority of loop anchors actually reside inside TADs[10]. During interphase, the DNA loop anchors are CCCTC-binding factor (CTCF) binding sites. Several studies have investigated the binding of CTCF in different tissues to determine the endpoints of DNA loops and help decipher TADs[10].

While the concept of regions of self-interaction of DNA for gene regulation has been established for some time[23,26-29], these new updates to SO allow for an accurate representation of the current understanding of TADs and the related concepts of TAD boundary and insulator elements. This new terminology enables these regions to be annotated in databases and knowledgebases whereby the precise biological condition and cell type can be captured. The hierarchical structure of TAD-related SO terms and their relationship to CRM is shown in the Figure.

Discussion

The GREEKC initiative has provided scientists and creators of biological ontologies an opportunity to collectively discuss how to accurately represent terms and relationships pertaining to gene regulation. Many terms have been either added to SO or updated in SO to allow it to better represent the current understanding of gene regulation. Specifically, several terms have been updated under the branch of *cis-regulatory module* (*CRM*). A new term, *core_promoter_element*, has been created to annotate elements that exist within the promoter region of a gene. The term *topologically_associated_domain* (TAD) has been added along with terms describing parts of TADs. Since insulators harbor CTCF sites and since CTCF sites form loop anchors, the addition of the TAD boundary term should allow different data types and analysis approaches that focus on either chromosome looping, enhancer insulation or topological segregation to be accurately annotated in an experimental entity-oriented fashion so as to permit objective discovery of epigenetic patterns and mechanisms of gene regulation in humans and other eukaryotes for biomedical research in particular.

These updates to SO have been conducted in parallel with updates to other biological ontologies, including the Gene Ontology (Gaudet *et al.* in preparation). The concurrent updates have allowed the ontologies to be interoperable, which will allow for the most accurate representation of complex concepts. For example, these updates to SO can already and will soon be used by reference annotations such as the Ensembl Regulatory Build[30], which annotates CRMs across genomes based on available public epigenomic data. Although already using SO, this new annotation will thus express more precisely the nature of the elements. These CRMs can be associated with transcription factor binding events (e.g. through motif analysis or ChIP-Seq), and therefore to upstream genes. In future, using cis-regulatory evidence (e.g. eQTLs or Hi-C), these CRMs will further be attached to their downstream target genes. Therefore, these sequence elements will constitute links between GO annotations.

Acknowledgments

DWS (trainee) and MS (trainee) supported by the National Library of Medicine through training grant T15LM007124. RCL supported by the Alzheimer's Research UK (ARUK-NAS2017A-1), and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The GO Consortium is funded by the National Human Genome Research Institute (US National Institutes of Health), grant number HG002273. The Gene Regulation Ensemble Effort for the Knowledge Commons was funded by GREEKC COST Action CA15205.

References

1. 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.* 2005;6(5):R44.
2. C.J. Mungall, C. Batchelor, K. Eilbeck. Evolution of the Sequence Ontology terms and relationships. *J Biomed Inform.* 2011;44(1):87-93.
3. B. Smith, M. Asnburner, C. Rosse, J. Bard, W. Bug. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007;25(11):1251-1255.
4. Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* 2001;11(8):1425-1433.
5. K. Degtyarenko K, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 2008;36:D344-350.
6. J.T. Bolleman, C.J. Mungall, F. Strozzi, J. Baran, M. Dumontier, R.J.P. Bonnal, R. Buels, R. Hoehndorf, T. Fujisawa, T. Katayama, P.J.A. Cook. FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *J Biomed Semantics.* 2016;7:39.
7. K. Verspoor, H. Shatkay, L. Hirschman, C. Blanschke, A. Valencia. Summary of the BioLINK SIG 2013 meeting at ISMB/ECCB 2013. *Bioinformatics*. 2013;31(2):297-298.

8. D.M. Jeziorska, K.W. Jordan, K.W. Vance. A systems biology approach to understanding cis-regulatory module function. *Semin Cell Dev Biol.* 2009;20(7):856-862.
9. Q. Li, K.R. Peterson, X. Fang, G. Stamatoyannopoulos. Locus control regions. *Blood*. 2002;100(9):3077-3086.
10. L. Nanni, S. Ceri, C. Logie. Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biol.* 2020;21(1):197.
11. L.T. Huong, M. Kobayashi, M. Nakata, G. Shioi, H. Miyachi, T. Honjo, H. Nagaoka. In vivo analysis of Aicda gene regulations: a critical balance between upstream enhancers and intronic silencers governs appropriate expression. *PLoS One*. 2013;8(4):e61433.
12. P. Kolovos, T.A. Knoch, F.G. Grosveld, P.R. Cook, A. Papantonis. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin*. 2012;5:1.
13. S. Lisser and H. Margalit. Compilation of E. coli mRNA promoter sequences. *Nucleic Acids Res.* 1993;21(7):1507-1516.
14. R. Hershberg, G, Bejerano, A. Santos-Zavaleta, H. Margalit. PromEC: An updated database of Escherichia coli mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.* 2001;29(1):277.
15. J. Zhu and M.Q. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*. 1999;15(7-8):607-611.
16. H. Li, J. Hou, L. Bai, C. Hu, P. Tong, Y. Kang, X. Zhao, Z. Shao. Genmoe-wide analysis of core promoter structures in Schizosaccharomyuces pombe with DeepCAGE. *RNA Biol*. 2015;12(5):525-537.
17. G.E. Zenter, P.J. Tesar, P.C. Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular function. *Genome Res.* 2011;21(8):1273-1283.
18. A. Laugesen, J.W. Hojfeldt, K. Helin. Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer. *Cold Spring Harb Perspect. Med.* 2016;6(9):a026575.

19. W. Wei, V. Pelechano, A.I. Jarvelin, L.M. Steinmetz. Functional consequences of bidirectional promoters. *Trends Genet.* 2011;27(7):267-276.
20. C. Mejia-Almonte, S.J.W. Busby, J.T. Wade, J. van Helden, A.P. Arkin, G.D. Stormo, K. Eilbeck, B.O. Palsson, J.E. Galagan, J. Collado-Vides. Redefining fundamental concepts of transcription initiation in bacteria. *Nat Rev Genet*. 2020.
21. R. Lovering, P. Gaudet, M.L. Acencio, A. Ignatchenko, A. Jolma, O. Fornes, M. Kuiper, I.V. Kulakovskiy, A. Lægrid, M.J. Martin, C. Logie. A GO catalogue of human DNA-binding transcription factors. *bioRxiv* 2020; doi:10.1101/2020.10.28.359232.
22. M. Falk, Y. Feodorova, N. Naumova, M. Imakaev, B.R. Lajoie, H. Leonhardt, B. Joffe, J. Dekker, G. Fudenberg, I. Solovei, L.A. Mirny. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*. 2019;570(7761):395-399.
23. J.R. Dixon, S Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485:376–80.
24. R.A. Beagrie, A. Scialdone, M. Schueler, D.C.A. Kraemer, M. Chotalia, *et al.* Complex multi-enhancer contacts captured by Genome Architecture Mapping (GAM). *Nature*. 2017;543(7646):519-524.
25. A. Galton. On generically dependent entities. *Applied Ontology*. 2014; 9(2):129-153.
26. E.P. Nora, B.R. Lajoie, E.G. Schulz, L. Giorgetti, I Okamoto, *et al*. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381-385.
27. E. Alipour E, J.F. Marko. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* 2012;40:11202–12.
28. A.L. Sanborn, S.S.P. Rao, S.C. Huang, N.C. Durand, M.H. Huntley, A.I. Jewett, *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA.* 2015;112:E6456–65.

29. K. Nasmyth. Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. *Annu Rev Genet.* 2001;35:673–745.

30. D.R. Zerbino, N. Johnson, T. Juetteman, D. Sheppard, S.P Wilder, *et al.* Ensembl regulation resources. *Database (Oxford)* 2016;bav119.

Figure.

Dendrogram showing the relationships between SO terms related to gene regulation discussed in this manuscript. Black arrows represent 'is_a' relationships, red arrows represent 'part_of' relationships and green arrows represent 'overlaps' relationships.

Author Statement

David W Sant: Writing – original draft, data curation, visualization, software. **Michael Sinclair**: Data curation **Christopher J Mungall**: Writing – review and editing, data curation, software. **Stefan Schulz**: Writing – review and editing, data curation. **Daniel Zerbino**: Writing – review and editing, data curation. **Ruth C Lovering**: Writing – review and editing, data curation. **Colin Logie**: Writing – original draft, data curation, visualization. **Karen Eilbeck**: Writing – original draft, supervision, resources, funding acquisition.

Conflict of Interest

November 5, 2020
Lennart Nilsson
Editor
Biochimica et Biophyisica Acta

The authors of this manuscript declare that they have no competing interests related to this manuscript.

Sincerely,

David W. Sant, Ph.D.
Postdoctoral Fellow
University of Utah
Department of Biomedical Informatics
Salt Lake City, UT 84108

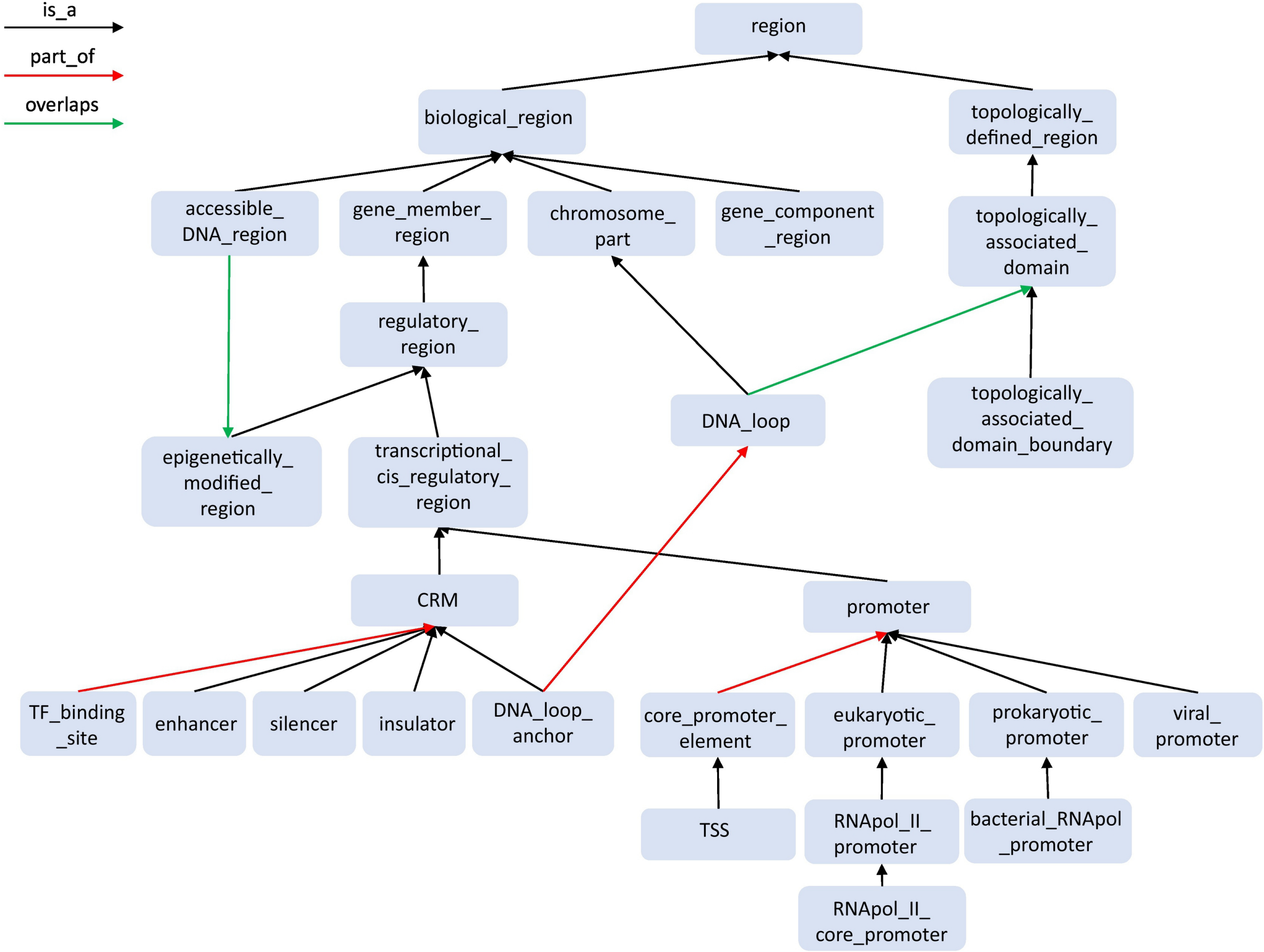
Highlights

Sequence Ontology has updated terminology related gene expression for GREEKC Project.

Cis-regulatory modules contain multiple elements that regulate gene expression.

Locus control regions confer high-level, copy number dependent expression of a gene.

Topologically associated domains promote physical proximity to regulate transcription.



is_a
part_of
overlaps
region
biological_region
topologically_defined_region
accessible_DNA_region
gene_member_region
chromosome_part
gene_component_region
topologically_associated_domain
regulatory_region
DNA_loop
topologically_associated_domain_boundary
epigenetically_modified_region
transcriptional_cis_regulatory_region
CRM
promoter
TF_binding_site
enhancer
silencer
insulator
DNA_loop_anchor
core_promoter_element
eukaryotic_promoter
prokaryotic_promoter
viral_promoter
TSS
RNApol_II_promoter
bacterial_RNApol_promoter
RNApol_II_core_promoter

Figure 1