Evolution of the Sequence Ontology terms and relationships

Karen Eilbeck[1], Christopher J. Mungall[2]

[1]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA;
[2]Lawrence Berkeley National Laboratory, Mail Stop 64R0121, Berkeley, CA 94720, USA.

Abstract

The Sequence Ontology is undergoing reform to meet the standards of the OBO Foundry. Here we report some of the incremental changes and improvements made to SO. We also propose new relationships to better define the mereological, spatial and temporal aspects of biological sequence.

Introduction

The Sequence Ontology[1] was begun in 2003 as a means to provide the terms and relations that obtain between terms, to describe biological sequence. The main purpose being the unification of the vocabulary used in genomic annotations, specifically genomic databases and flat file data exchange formats. Genomic data has been notoriously unspecified with a multitude of file formats expressing the same kind of data in different ways. Each gene prediction algorithm for example, exported the gene models in either a different format from other groups, or when they used the same format, the terms often had slightly different meanings. Data integration between groups was therefore not straightforward. Likewise, validation of annotations relied on the programmers understanding the nuances of each kind of annotation and hard-coding their programs to match. The Sequence Ontology provides a forum for the genomic annotation community to discuss and agree on terminology to describe their biological sequence.

The SO was initially divided into aspects to describe the features of biological sequence and the attributes of these features. A sequence feature is a region or a boundary of sequence that can be located in coordinates on biological sequence. SO uses a subsumption hierarchy to describe the kinds of features and meronomy to describe their containment. Features were related by their genomic position. For example polypeptides and transcripts are described by genomic context. This excluded their post-genomic topology.

The SO has a large user community of established model organism databases and newer ‘emerging model organism’ systems who rely on the GMOD[2] suite of tools to annotate and disseminate their genetic information. GMOD schemas and exchange formats rely on the SO to type their features such as the Chado database[3], with its related XML formats and the tab delimited flat file exchange format GFF3[4]. Several GMOD tools use GFF3, for example GBrowse[5]. SO is also used by genome integration projects such as Flymine[6], modENCODE[7] and the BRC pathogen data repository[8]. There are other uses for SO such as natural language processing initiatives that use the SO terminology[9,10].

The SO is one of the original members of the OBO Foundry[11]. The OBO ontology developers agreed to a set of shared principles for formal ontology design, with the aim of achieving orthogonal, interoperable ontologies. There are 10 principles for OBO Foundry membership which include a common syntax, a data-versioning system, collaborative development, and adherence to the same set of defined relationships[12]. The OBO ontology developers attempt to accurately represent reality. Membership in the OBO Foundry represents a commitment to adhere to the ontology design principles and agree to reform where necessary. The OBO Foundry spans the biomedical domain in steps of granularity from the molecule to the population. It also encompasses the relations to time. Continuants endure through time, where as occurrents, which include processes, unfold through time in stages. The sequence features of SO are instantiated as molecules or parts of molecules.

The SO has orthogonal neighbor ontologies within the OBO Foundry, also describing molecular continuants. Chemical entities of Biological Interest (ChEBI) is a dictionary of small chemical compounds[13]. It does not describe molecules encoded by the genome such as transcripts and peptides. The RNA Ontology[14] describes the secondary and tertiary motifs of RNA as well as providing relationships between bases for base pairing and stacking. The Protein Ontology (PRO) defines the forms of proteins and the evolutionary relationships between protein families[15]. It is natural for these ontologies to interact and create inter-ontology terms in the form of cross products. To do so, the ontologies must all adhere to the same principles.

Coordinated reform of SO to OBO standards

The SO, like other pre-exisiting ontologies has begun to undergo reform to meet the OBO Foundry standards.

Textual Definitions

New terms are now defined using OBO Foundry guidelines for definitions. The existing terms in SO were initially either defined by a member of the developer community, or via a cross reference to a reputable source. This has lead to inconsistency between the definitions, and sometimes inconsistency between the definition and placement of the term. The OBO Foundry recommends that terms be defined with respect to the *is_a* parent, and the attributes that differentiate the term from its parent and sibling terms, called the differentiae. This practice forces a self check on the whether the position of the term in the ontology agrees with the defined meaning of the term. New definitions in SO must adhere to the "A is_a B that C's" principle. For example, the new term, **vector_replicon**, a subtype of **replicon**, has the following definition: *A replicon that has been modified to act as a vector for foreign sequence*. Existing terms are undergoing a refinement process.

Logical Definitions

In addition to providing text definitions, the SO includes over 100 'cross-product' definitions in genus/differentiae form[16]. A reasoner can then be used to place the terms in the correct place in the ontology. This is especially useful as it untangles the graph for editing purposes. The SO is released in two forms, either with the logical definitions, or fully classified for use without a reasoner.

Parthood Relations

The SO must adhere to the principles of OBO Relations Ontology (RO). The RO provides a set of defined formal type level and instance level relations. The list of relations may be extended by individual ontologies as required. The class level relations follow the "ALL_SOME" rule[17]. This rule is necessary to improve the ability to reason over data that uses the ontology. In practice, making these changes to SO has required the addition of the '*has_part'* relation to the ontology. Prior to this change the SO stated that:

TATA_box *part_of* **RNApol_II_promoter** and
TATA_box *part_of* **RNApol_III_promoter**.

This was incorrect as all **TATA_boxes** are not part of both kinds of promoter. The ontology now states that **RNApol_III_promoter** *has_part* **TATA_box**.

The *integral_part_of* relation and its inverse have been added to clarify the occasions when the part and the whole must both exist.

Temporal relations and spatial interval relations

There are several kinds of relation that are needed to describe the complex nature of biological sequence. Mereological relations are needed to describe containment. Spatial relations are needed to relate the positional information about features. Each transformation of sequence requires a temporal relation. Finally as SO is part of a larger suite of ontologies, it will need relations with which to make cross products and refer to other ontologies. We propose to extend SO with the relations outlined in Table 1.

Biological sequence is predominantly instantiated in three kinds of polymeric molecule: DNA, RNA and polypeptide, although man-made polymers such as PNA do exist. The SO will represent the transformation of sequence from one kind of molecule to another using the temporal relations shown in Table 1. A **gene**, manifest in DNA *transcribed_into* the **primary_transcript**, which is expressed as RNA. A **polypeptide** sequence is a *translation_of* the **CDS** sequence. **Transcript** molecules also undergo processing such as splicing and editing which remove or add additional sequences. The relations *processed_from* and *processed_into* relate the primary transcript to its mature processed form.

It is important to understand how the proposed changes will affect the annotation community who already use the terms and relations of SO in their pipelines and processes. This will effect how the changes are released. The terminology used to type the features already in use will not change. The GFF3 format will be unaffected as it lists the feature types and the parent term of a given relation. It does not name the relation – this is maintained in the ontology.

Developers will need to be given notice of new relationships and structures however, as this may have adverse effects of pipelines and programs.

The proposed changes to the SO relationships and structure can be found on the SO website at following address:
http://www.sequenceontology.org/resources/proposed_relationships.html

Conclusions

The updates to the SO, based on OBO Foundry recommendations have strengthened the ontology as a tool for reasoning. The treatment of definitions enforces a tight regulation on the position of a new term in the ontology and synchronizes the textual

	Name	Definition	example
Mereological	part_of	X part_of Y if X is a subregion of Y.	**amino_acid** *part_of* **polypeptide**
	has_part	Inverse of part_of	**operon** *has_part* **gene**
	integral_part_of	X integral_part_of Y if and only if: X part_of Y and Y has_part X	**exon** integral_part_of **transcript**
	has_integral part	X has_integral_part Y if and only if: X has_part Y and Y part_of X	**mRNA** *has_integral_part* **CDS**
Temporal	transcribed_from	X is transcribed_from Y if X is synthesized from template Y.	**primary_transcript** *transcribed_from* **gene**
	transcribed_to	Inverse of transcribed_from	**gene** *transcribed_to* **primary_transcript**
	translation_of	X is translation of Y if X is translated by ribosome to create Y.	**Polypeptide** *translation_of* **CDS**
	translates_to	Inverse of translation _of	**codon** *translates_to* **amino_acid**
	processed_from	Inverse of processed_into	**miRNA** *processed_from* **miRNA_primary_transcript**
	processed_into	X is processed_into Y if a region X is modified to create Y.	**miRNA_primary_transcript** *processed into* **miRNA**
Spatial Interval	contained_by	X contained_by Y iff X starts after start of Y and X ends before end of Y	**intein** *contained_by* **immature_peptide_region**
	contains	Inverse of contained_by	**Pre-miRNA** *contains* **miRNA_loop**
	overlaps	X overlaps Y iff there exists some Z such that Z contained_by X and Z contained_by Y	**coding_exon** *overlaps* **CDS**
	maximally_overlaps	A maximally_overlaps X and Y iff all parts of A (including A itself) overlap both X and Y	**non_coding_region_of_exon** *maximally overlaps* the intersection of **exon** and **UTR**
	connects_on	X connects_on Y,Z,R iff whenever X is on a R, X is adjacent_to a Y and adjacent_to a Z	**splice_junction** *connects_on* **exon**, **exon mature_transcript**
	disconnected_from	X is disconnected_from Y iff it is not the case that X overlaps Y	**intron** *disconnected_from* **exon** {on **transcript**}
	adjacent_to	X adjacent to Y if and only if: X and Y share a boundary but do not overlap	**UTR** *adjacent_to* **CDS**
	started_by	X is started by Y, if Y is part_of X and X and Y share a 5 prime boundary.	**CDS** *started_by* **start_codon**
	finished_by	X is finished by Y if Y is part_of X and X and Y share a 3 prime boundary	**CDS** *finished_by* **stop_codon**
	starts	X starts Y is X is part of Y and X and Y share a 5 prime boundary.	**start_codon** *starts* **CDS**
	finishes	X finishes Y if X is part_of Y and X and Y share a 3' boundary.	**stop_codon** *finishes* **CDS**
	is_consecutive_sequence_of	R is_consecutive_sequence_of U if and only if every instance of R is equivalent to a collection of instances of U u1,u2,...,un such that no pair ux uy is overlapping, and for all ux, ux is adjacent_to ux-1 and ux+1, with the exception of the initial and terminal u1 and un (which may be identical).	**region** *is_consecutive_sequence_of* **base** **processed_transcript** *is_consective_sequence_of* **exon**
Cross ontology	site_of	A is a site of B if A is the sequence_feature of a molecule where a GO:biological process B occurs.	**CDS** *site_of* **RNA polymerase activity**
	output_of	A is an output_of B if A is a sequence_feature of a molecule that is produced by GO:biological process B.	**primary_transcript** *output_of* **transcription**
	regulates_expression_of	A regulates expression of B if A is a regulatory region that controls the expression of B, where B is a gene.	**regulatory_region** *regulates_expression_of* **gene**

Table 1. *New relations proposed for SO. Definitions are for instance level relations, examples are for class-level relations, which follow from the instance-level definition in the standard all-some pattern.*

definition within the subsumption hierarchy. The process of updating all of the definitions is ongoing. Stricter adherence to the OBO Relations Ontology is making SO interoperable with the other OBO ontologies. The SO uses a reasoner to maintain the is_a parents of cross product terms. This aids ontology maintenance and can be used as a model for other OBO ontologies.

The application of sequence features that span the range of the molecular biology central dogma, rather than simply the position of the genomic region that encodes the molecule, is a subtle but important step forward. It allows the topological relations at each stage from genome to transcript or peptide to be catalogued. It roots the SO within OBO making cross products between the sibling ontologies possible.

The addition of a suite of mereological, topological and temporal relations will dramatically enhance the ability to use the SO as a tool for computational reasoning. Each of the new defined relationships adds another avenue for analysis. This is especially important for the validation of sequence annotations using SO.

Acknowledgments
This work is supported by the NHGRI, via the Gene Ontology Consortium, HG004341.

References

1. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: The Sequence Ontology: a tool for the unification of genome annotations. Genome biology 2005, 6(5):R44.
2. GMOD [www.gmod.org]
3. Mungall CJ, Emmert DB: A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics (Oxford, England) 2007, 23(13):i337-346.
4. GFF3 [www.sequenceontology.org/gff3.shtml]
5. Donlin MJ: Using the Generic Genome Browser (GBrowse). Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al 2007, Chapter 9:Unit 9 9.
6. Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, McLaren P, North P et al: FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome biology 2007, 8(7):R129.
7. modENCODE [www.modencode.org]
8. BRC [http://www.brc-central.org/cgi-bin/brc-central/brc_central.cgi]
9. Andreas Vlachos CG, Ian Lewin, Ted Briscoe Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles. In: Pacific Symposium on Biocomputing. vol. 11; 2006: 100-111.
10. RSC [http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp]
11. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ et al: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 2007, 25(11):1251-1255.
12. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: Relations in biomedical ontologies. Genome biology 2005, 6(5):R46.
13. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research 2008, 36(Database issue):D344-350.
14. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE et al: The RNA Ontology Consortium: an open invitation to the RNA community. RNA (New York, NY 2006, 12(4):533-541.
15. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B, Wu CH: Framework for a protein ontology. BMC bioinformatics 2007, 8 Suppl 9:S1.
16. [http://wiki.geneontology.org/index.php/SO:Composite_Terms]
17. Smith B, Kumar A, Ceusters W, Rosse C: On carcinomas and other pathological entities. Comparative and functional genomics 2005, 6(7-8):379-387.
18. ALLEN JF: Maintaining Knowledge about Temporal Intervals. Communications of the ACM 1983, 26(11):832-843.
19. OBO format 1.3 [http://www.geneontology.org/GO.format.obo-1_3.shtml