Published in final edited form as:
J Biomed Inform. 2011 February ; 44(1): 87–93. doi:10.1016/j.jbi.2010.03.002.

Evolution of the Sequence Ontology terms and relationships

Christopher J. Mungall[1], **Colin Batchelor**[2], and **Karen Eilbeck**[3]

[1] Lawrence Berkeley National Laboratory, Mail Stop 64R0121, Berkeley, CA 94720, USA

[2] Royal Society of Chemistry, Thomas Graham House, Cambridge, UK CB4 0WF

[3] Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA

Abstract

The Sequence Ontology is an established ontology, with a large user community, for the purpose of genomic annotation. We are reforming the ontology to provide better terms and relationships to describe the features of biological sequence, for both genomic and derived sequence. The SO is working within the guidelines of the OBO Foundry to provide interoperability between SO and the other related OBO ontologies. Here we report changes and improvements made to SO including new relationships to better define the mereological, spatial and temporal aspects of biological sequence.

Keywords

Sequence Ontology; biomedical ontology; genome annotation

1. Introduction

Genomic data was notorious for the multitude of file formats that expressed 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 (SO) [1] was initiated in 2003 to provide the terms, and relations that obtain between terms, to describe biological sequences. The main purpose was to unify the vocabulary used in genomic annotations, specifically genomic databases and flat file data exchange formats. The Sequence Ontology Project provides a forum for the genomic annotation community to discuss and agree on terminology to describe the biological sequence they manage, in the form of mailing lists, trackers, and workshops.

The purpose of annotating a genome is to find and record the parts of the genome that are biologically significant. In this way researchers can make sense of what would just be a very long string of letters. For example, after annotation, a researcher will be able to know which of the sequence variants fall in coding or non coding sequence and perform subsequent

Karen Eilbeck, Department of Human Genetics, Building 533, 15 N 2030 East, Salt Lake City, Utah 84108. keilbeck@genetics.utah.edu, Fax: 801 581 7796.

analyses accordingly. A genome annotation anchors knowledge about the genomic sequence and the sequence of molecules derived from the genome on to a linear representation of the replicon (chromosome, plasmid etc) using base pair coordinates to capture the position. A **sequence_feature** is a region or a boundary of sequence that can be located in coordinates on biological sequence, and SO was initially created as an ontology of these sequence feature types and their attributes.

The SO has a large user community of established model organism databases and newer 'emerging model organism' systems who use on the Generic Model Organism Database (GMOD) [2] suite of tools to annotate and disseminate their genetic information. GMOD is a group that provides an open source collection of tools for dealing with genomic data. 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 Generic Feature 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 [8,9]. There are other uses for SO such as natural language processing initiatives that use the SO terminology [10,11].

Genome annotations specify the coordinates of sequence features that are manifest in one or more of the kinds of molecule defined by the central dogma. For example, although an intron is manifest as an RNA molecule, the coordinates of the intron can be projected onto the genomic sequence. The term labels chosen for SO were those in use by the genome annotation community, thus "transcript", "intron" and so on were chosen as labels for the sequence feature types corresponding to genome regions encoding actual transcript and intron molecules. This polysemy does not cause problems when SO is used purely for genome annotation, but is potentially confusing when it is used in the context of other ontologies.

The current version of SO uses a subsumption hierarchy to describe the kinds of features and a meronomy to describe their part-whole structures. Sequence features were related by their genomic position. For example **polypeptide** (which referred to the sequence that corresponds to a polypeptide molecule) and **transcript** (which referred to the sequence that corresponds to an RNA molecule) were described only by genomic context, that is the region of the genome that encodes their sequence. This excluded the post-genomic topology of these features: how the topology of the features changes, as the sequence is expressed by different molecules.

The SO is one of the original members of the OBO Library, a collection of orthogonal, interoperable ontologies developed according to a shared set of principles. These later evolved into the OBO Foundry principles [12] which include a common syntax, a data-versioning system, collaborative development, and adherence to the same set of defined relationships [13]. The OBO Foundry ontology developers attempt to accurately represent biological reality. Membership in the OBO Foundry represents a commitment to adhere to common 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 organism, and also extends into the realm of experimental measurements, instrumentation and protocol. The OBO Foundry also partitions ontologies according to their relationship to time. Continuants endure through time, whereas occurrents, which include processes, unfold through time in stages. Anatomical entities such as cells and organs are continuants, as are molecules.

The SO is orthogonal to the neighbor ontologies within the OBO Foundry which represent molecular continuants. Chemical Entities of Biological Interest (ChEBI) is a dictionary of

small molecules [14]. The RNA Ontology [15] represents the secondary and tertiary motifs of RNA as well as describing the interactions between bases for base pairing and stacking. The Protein Ontology (PRO) defines the forms of proteins and the evolutionary relationships between protein families [16]. These ontologies are themselves orthogonal to ontologies of processes, such as the Biological Process (BP) and Molecular Function (MF) subsets of the Gene Ontology (GO) [17]. The GO BP ontology represents processes of relevance to SO, such as transcription, gene expression and splicing.

In order to best divide work between curators of neighboring ontologies, and to ensure that SO can reuse material from these ontologies and *vice versa*, the ontologies must all adhere to the same principles. In this paper we will describe how we have been developing the Sequence Ontology in two respects, first to promote interoperability and second to provide a solid framework to describe how sequences change over the course of genomic and post-genomic processes. The rest of the paper is structured as follows: in Section 2 we describe the OBO Foundry standards we have been adopting. In Section 3 we describe new relations for post-genome topology and in Section 4 we describe the relation of SO to neighboring ontologies.

2. Coordinated reform of SO to OBO standards

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

2.1 Conformation to an upper Ontology

Upper ontologies such as Basic Formal Ontology (BFO) [18] provide a formal structure upon which to base domain ontologies. BFO provides a hierarchy of upper-level abstract classes. Classes in domain-specific ontologies can be defined as sub-classes of appropriate abstract classes and inherit their properties. This allows the multiple independently developed ontologies of the OBO Foundry to be linked together. The development of SO preceded the adoption of BFO by the OBO Foundry, so it was necessary to align SO to BFO post-hoc. In order to do this, a fundamental question must be answered: what kind of entity is a sequence feature? This is not a trivial question and suggested answers have ranged from: molecules or molecule regions, the physical pattern of electrons in a computer or purely abstract mathematical forms. None of these solutions was biologically satisfying. Our position is that biological sequences exist independently of our abstractions or computational representations, but are not identical with the molecules themselves. Multiple molecules can have the same sequence, and a sequence feature exists so long as there is a molecule with that sequence. This can be seen as analogous to the distinction between the physical content of a book, and the words written in that book.

BFO divides continuants into **independent continuants** and **dependent continuants**. The former include physical objects such as molecules, and the latter include entities such physical qualities, shapes and functions. The relation that links these is called *inheres_in*, and we say that for example my temperature *inheres_in* me, or that I am the bearer of my temperature. Dependent continuants are broken down into **specifically dependent continuants** (SDCs) and **generically dependent continuants** (GDCs). What differentiates these is the number of bearers – a SDC has a single bearer, and ceases to exist when that bearer ceases to exist (thus the shape of a particular apple disappears after the apple is eaten). A GDC can have multiple bearers, and can continue to exist when bearers cease to exist, so long as there is at least one bearer. A given genomic sequence may be borne by a DNA molecule, an RNA molecule, a polypeptide chain, or indeed by other molecules or systems that are not products of the replication machinery of the cell, for example the set of instructions that drive a solid-phase nucleic acid synthesis device. For this reason we take

biological sequences to be GDCs (Fig 1). One of the consequences of this decision is that genes such as the gene denoted by the NCBI Gene ID 6469 (human Shh) are *individuals* rather than *types*.

The other SO root classes have also been aligned to BFO, as shown in Figure 1. We take **sequence_collection**, which is a non-contiguous set of sequences, and **sequence_variant**, such as a mutation, to be the same sort of thing as a sequence_feature, and hence a GDC. For the moment we are treating **sequence_attribute** as an intrinsic property of the molecule that bears the sequence, hence in BFO terms a quality, but this is under review. Lastly, the **sequence_variant_effect**, for example a structural change or a change in transcription, need not necessarily happen so we treat it as a disposition.

2.2 Definitions

We now define new terms according to the OBO Foundry guidelines for definitions. Initially the terms in SO were either defined by a member of the developer community, or taken directly from a reputable website or textbook, giving the ISBN or the URL as the cross-reference. This has led to inconsistency between the definitions, and sometimes inconsistency between the definition and placement of the term within the ontology. This especially led to confusion over the kind of entity described by a feature, whether it was a molecule or a sequence, as there was not conformity in the definitions. For example, **mRNA** was defined as: *Messenger RNA is the intermediate molecule between DNA and protein. It includes UTR and coding sequences. It does not contain introns.* This has been updated to *'Messenger RNA sequence is a mature transcript sequence, a portion of which is coding. It may include UTR but not intron sequence'*. 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*. We are actively refining existing terms.

The SO was the first ontology in the OBO library to augment free text definitions aimed at humans, with computable necessary and sufficient 'cross-product' definitions. SO has 100 of these definitions in genus/differentiae form [19]. The genus is the broader category to which the term belongs, and the differentiae are the properties that other members of the genus do not have.

To achieve these computable definitions, **sequence_feature** terms are defined with **sequence_attribute** terms, using a new relationship *has_quality*[1]. Previous to the creation of cross product terms, a complex term such as **engineered_foreign_transposible_element_gene** would have several manually edited *is_a* parents: **transposable_element_gene, engineered_foreign_gene**, and **engineered_foreign_region**. These multiple parents cause problems for the ontology developer and for visualization and reasoning software. The developer must manually check for other *is_a* parents percolating further up the graph. The graph itself becomes difficult to navigate. With the addition of the cross-product relations, the definition becomes computationally visible. The term **engineered_foreign_transposible_element_gene** now has a single *is_a* parent: **transposable_element_gene** and two qualities: **engineered** and

[1]This is under review – according to BFO, the quality inheres in the independent continuant, so we will likely need a relation that chains the sequence feature to the molecule to the quality.

foreign. 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.

2.3 Parthood Relations

In order for reasoners to be able to draw correct inferences about the entities in an ontology, the class level relations must be of the all–some, all–only or all all–types, of which "all–only" is the weakest. This is one of the reasons for the Foundry principle that ontologies should reuse relations from the OBO Relations Ontology (RO), which provides a set of defined formal type level and instance level relations, typically of the all–some form [13]. The list of relations may be extended by individual ontologies as required. In practice, making these changes to SO has required the addition of the '*has_part*' relation to the ontology. For example, the ontology states that **overlapping_EST_set** *has_part* **EST**. If this relation was reversed and the ontology stated that **EST** *part_of* ***overlapping_EST_set*** it would have serious implications for software that use reasoning to validate sequence annotations. This would imply that all EST sequence annotations were part of a region composed of more than one EST, and therefore single EST's would incorrectly cause a validation error.

We have added the *integral_part_of* relation and its inverse. X *integral_part_of* Y iff every X *part_of* some Y and every Y *has_part* some X. This covers the cases where the existence of the part implies the existence of the whole and *vice versa*.

3. 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. These relations are based on Allen's interval logic [20]. Each transformation of sequence requires a temporal relation. We propose to extend SO with the relations outlined in Table 1.

Biological sequences inhere in three kinds of polymeric molecule that are produced by the cell's replication machinery: DNA, RNA and polypeptide. There are also man-made polymers that can bear sequences, such as PNA [21]. The SO will represent the transformation of sequence from one kind of molecule to another using the temporal relations shown in Table 1. A **primary_transcript**, which is expressed as RNA, is *transcribed_from a* **gene**. A **polypeptide** sequence is a *ribosomal_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.

The actual names of relations are under review – for example, we may decide to use sequence-specific relations such as *upstream_adjacent_to* in place of the *starts* relation, as it may be desirable to reserve *starts* as a temporal relation between processes.

4. Relation to neighboring OBO ontologies

4.1 SO and GO

Ontology term reuse is a vital part of the OBO Foundry project [12]. SO and its neighboring ontologies are shown in Table 2.

The Gene Ontology is reforming itself in line with OBO Foundry principles by adding cross-product definitions of its classes where possible [19]. However, for those terms that

involve DNA, RNA and polypeptides, this alignment is hampered by SO describing those sequences that inhere in molecules rather than the molecules themselves. Most biologically relevant molecules belong in ChEBI; however, the scope of ChEBI explicitly excludes molecules that are specified by the genome. This gap is now filled by Sequence Ontology:Molecules (SOM), an ontology of molecules of genomic origin. This separate ontology that represents the molecules and molecular parts that correspond to SO terms such as **exon**, **intron**, **transposon** and so forth, will provide a bridge to neighboring ontologies in the form of cross product generation. ChEBI will continue to provide the molecular units from which genomic molecules are constructed, such as nucleotide residues. A further distinction from a purely structure-based interpretation of the ChEBI ontology is that the circumstances of a molecule are important.[2] For example, an intron molecule is necessarily the result of a splicing process—an atom-for-atom identical molecule in a comet would not be an intron—hence in BFO terms they are defined classes rather than universals. The classes in SOM are cross-referenced to the Sequence Ontology via their logical definitions. In some cases the SO term (the sequence, for example an intron) takes logical precedence, hence the SOM term will be defined in terms of SO, while in other cases the SOM term (the molecule, for example a transfer RNA) comes first. Figure 3 illustrates the difference between SO and SOM. Note that the ontology structure is not always completely isomorphic – a transcript feature such as an exon or intron can be a subsequence of the genomic sequence or the transcript sequence, but this is not true of the corresponding molecule. Equally, not every class in SO has a SOM counterpart, and *vice versa*.

4.2 GO and SO

Conversely, there were terms in SO that described what are really processes, such as **rolling circle replication** and **theta replication**. As such, these terms have been obsoleted in SO and donated to the Gene Ontology.

4.3 SO and the Ontology for Biomedical Investigations

The Sequence Ontology has always contained terms for annotating sequence regions according to how they were obtained and how much and how well they have been sequenced. As such there is overlap with recent work on the Ontology for Biomedical Investigations (OBI) [22]. OBI is formalizing the representation of experimental design, protocol, instrumentation, materials, data generated and analyses performed. SO has taken steps to redefine the kinds of biological region to be in alignment with the OBI distinctions. **Region** has thus been subtyped to include: **biological_region**, those mind-independent sequences inhering in the nucleic acid and peptide molecules of you, me and the dinosaurs, **biomaterial_region**, describing those sequences with a specified experimental purpose, acting as a bacterial vector, for example, and **experimental_feature**, describing how sequences were assembled, whether they were a match, contig, supercontig or so forth, and what is known about them. Again, the **biomaterial_region** sequences are defined classes rather than universals and inhere in molecules which have a particular function or role. Those functions and roles are the domain of OBI.

4.4 SO and the formal Ontology of Sequences

Hoehndorf *et al*.[23] have written an interesting “bottom-up” account of axiomatizing sequences and their relations from a logician’s perspective. Two of the assumptions made in the paper allow us to clarify some important points, one about sequence mereogeometry, the other about existential dependence. The first is that they draw a faulty distinction between

[2]Though in ChEBI the only difference, for example, between **glycolipid** (CHEBI:33563) and **neoglycolipid** (CHEBI:51019) is that the latter has been synthetically produced.

the mereology of sequences and the mereology of molecules, in arguing that the sequence *ACAC* has a single sequence *AC* that appears twice (and hence only seven sequences that are parts), as opposed to the molecule *ACAC* which has distinct molecular parts *AC*- and -*AC*, and hence ten molecular parts. Hoehndorf *et al.* read the sequence *ACAC* such that *AC* proper_part_of *ACAC*, which contravenes the weak company principle [24] that

$$AC \text{ proper\_ part\_ of } ACAC => \text{exists}(S') \& S' \text{ proper\_ part\_ of } ACAC \& S' != AC.$$

But this is a misreading, because *ACAC* should really be read as $A_nC_{n+1}A_{n+2}C_{n+3}$. Sequences, as their name suggests, consist of parts *in a particular order*, and the part *AC* that starts at position *n* is clearly distinct from the part *AC* that starts at position *n*+2. The second assumption is that they take junctions to be specifically dependent in the BFO sense on their sequences (by which they mean sequence regions). It is true that junctions existentially depend on the regions they start or end, but the sense of "dependence" intended by BFO's "specifically dependent continuant" is one of inherence, and junctions inhere in molecules just as regions do.

5. Conclusions and future steps

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 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 neighbor 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.

The creation of the SOM subset of terms fills a gap in the OBO Foundry ontologies between SO, ChEBI and RNAO, in describing the physical molecules that are encoded by genomes. This will greatly facilitate inter-ontology relations, and also be useful in defining SO terms. The placement of both SO and SOM into the BFO hierarchy also strengthens the interoperability of the ontologies, and promotes reuse and cross product formation.

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. The daily revisions to the SO are managed using a CVS repository [25], and there is a bi-monthly release schedule for more stable versions [26]. Developers are either committed to using the revisions or releases. SOM is checked into the CVS repository and will undergo releases as required. 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 are given notice of new relationships and structures via the developer mailing list, as this may have adverse effects of pipelines and programs. The relations are added to the ontology before they are used structurally. A webpage addresses the upcoming changes to the SO [27].

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 Biol 2005;6:R44. [PubMed: 15892872]
2. Available at www.gmod.org
3. Mungall CJ, Emmert DB. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 2007;23:i337–46. [PubMed: 17646315]
4. Available at www.sequenceontology.org/gff3.shtml
5. Donlin MJ. Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics 2007;Chapter 9(Unit 9):9. [PubMed: 18428797]
6. Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, McLaren P, North P, Rana D, Riley T, Sullivan J, Watkins X, Woodbridge M, Lilley K, Russell S, Ashburner M, Mizuguchi K, Micklem G. FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biol 2007;8:R129. [PubMed: 17615057]
7. Available at www.modencode.org
8. Available at http://www.brc-central.org/cgi-bin/brc-central/brc_central.cgi
9. IOWG Gff3 Usage Conventions. Available at http://www.pathogenportal.org/gff3-usage-conventions.html
10. Andreas Vlachos, CG.; Lewin, Ian. Ted Briscoe Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles. Pacific Symposium on Biocomputing; 2006. p. 100-111.
11. Kidd R. Changing the face of scientific publishing. Integrative Biology 2009;1:293–295.
12. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5. [PubMed: 17989687]
13. 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 Biol 2005;6:R46. [PubMed: 15892874]
14. 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 Res 2008;36:D344–50. [PubMed: 17932057]
15. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, Major F, Mathews DH, Richardson JS, Williamson JR, Westhof E. The RNA Ontology Consortium: an open invitation to the RNA community. Rna 2006;12:533–41. [PubMed: 16484377]
16. 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. [PubMed: 18047702]
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9. [PubMed: 10802651]

18. Grenon, P.; Smith, B.; Goldberg, L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli, DM., editor. Ontologies in Medicine. Amsterdam: IOS Press; 2004. p. 20-38.

19. Cross-Product Extensions of the Gene Ontology. Available at http://precedings.nature.com/documents/3496/version/1

20. ALLEN JF. Maintaining Knowledge about Temporal Intervals. Communications of the ACM 1983;26:832–843.

21. Egholm M, Buchardt O, Christensen L, Behrens C, Freier SM, Driver DA, Berg RH, Kim SK, Norden B, Nielsen PE. PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. Nature 1993;365:566–8. [PubMed: 7692304]

22. The OBI Consortium. Available at http://purl.obolibrary.org/obo/obi

23. Hoehndorf R, Kelso J, Herre H. The ontology of biological sequences. BMC Bioinformatics 2009;10:377. [PubMed: 19919720]

24. Mereology. Available at http://plato.stanford.edu/archives/sum2009/entries/mereology

25. Sequence Ontology CVS repository. Available at http://song.cvs.sourceforge.net/viewvc/song/ontology/

26. Sequence Ontology Release Repository. Available at http://sourceforge.net/projects/song/files/

27. Sequence Ontology Relations. Available at http://www.sequenceontology.org/resources/proposed_relationships.html

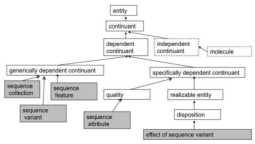

Figure 1.
The Sequence Ontology root terms aligned with the Basic Formal Ontology. SO terms are grey. Within the framework of BFO, GDCs stand in an inheres_in relation to independent continuants (e.g. molecules). Independent continuants are shown in dashed boxes.

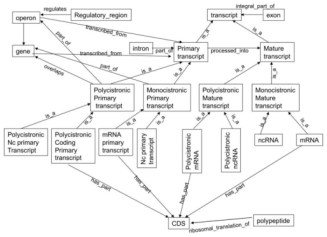

Figure 2.

Post Genome Topology: moving beyond mereology to describe the relation of **sequence_features**. The key **sequence_features** and relations, that describe transcription and translation.

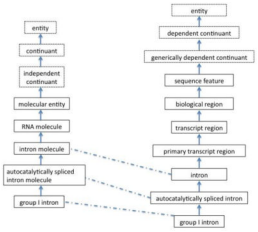

Figure 3.

Comparing the molecular and sequence is_a parents of ID mapped terms in SOM and SO. The SOM subsumption hierarchy for **group I intron** is shown in the left hand column, and SO in the right, In SOM, an **intron molecule** is a kind of **RNA molecule**, where as in SO, an **intron** is a region of sequence. Equivalent terms in SO and SOM have different definitions, cross references and synonyms etc. They also have different relationships. Mappings to BFO are shown as dashed boxes and equivalent ID mappings as dashed lines.

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.

	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**
	ribosomal_translation_of	X ribosomal_translation_of Y - a ribosome reads X and through a series of GO:0030533 ! triplet codon-amino acid adaptor activity processes executed in sequence outputs a Y.	**Polypeptide** *translation_of* **CDS**
	ribosomal_translates_to	Inverse of ribosomal_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 or N terminal boundary	**start_codon** *starts* **CDS**
	finishes	X finishes Y if X is part_of Y and X and Y share a 3′ boundary or C terminal 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	**region** *is_consecutive_sequence_of* **base** **processed_transcript** *is_consective_sequence_of* **exon**

	Name	Definition	example
		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).	

Table 2

The domain of SO, SOM and neighboring ontologies.

	Category				
	Material entity	**Sequence**	**Quality**	**Disposition, function or role**	**Process**
Non-genomically encoded molecule	ChEBI	*none*	*none*	ChEBI	GO
Nucleotide residues, amino acid residues	ChEBI	SO	*none*		
Sub-residue divisions of RNA molecules, base pairs, base stacks	RNAO	*none*	RNAO		
Genomically-encoded nucleic acid	SOM	SO	*none*		
Multiple-residue divisions of genomically-encoded molecules			*none*		
Genomically-encoded polypeptide	PRO		*none*		