Efforts toward a More Consistent and Interoperable Sequence Ontology

Michael Bada[1*] and Karen Eilbeck[2]
mike.bada@ucdenver.edu, keilbeck@genetics.utah.edu
[1] University of Colorado Anschutz Medical Campus, Department of Pharmacology, MS 8303, RC-1 South, 12801 East 17th Avenue, L18-6400A, P.O. Box 6511, Aurora, CO 80045 USA
[2] Department of Biomedical Informatics, Health Sciences Education Bldg, University of Utah, 26 South 2000 East, Suite 5700, Salt Lake City, UT 84112-5750 USA

ABSTRACT

The Sequence Ontology (SO), a member of the Open Biomedical Ontologies (OBOs) library, was developed with the goals of standardizing the vocabulary and semantics of biological-sequence annotation with the goal of increased interoperability for software developers and users of genomic sequences. Here we present our recent developmental approaches to address three issues of import for the SO: (1) representation of molecular sequences versus abstract sequences; (2) integration with the ChEBI ontology, the Protein Ontology, the RNA Ontology, the Gene Ontology, the Chemical Information Ontology, and the Information Artifact Ontology; and (3) consistent representation of DNA, RNA, and peptide sequences and harmonizing their use toward annotation in sequence databases. We anticipate that these efforts will result in a representation of biological sequences that is more consistent not only internally but also with respect to its use in annotations in sequence databases. We further envision increased interoperability of the SO with other OBOs, which would benefit applications beyond sequence annotation.

1 INTRODUCTION

The Sequence Ontology (SO), a member of the Open Biomedical Ontologies (OBO) library (Smith *et al.*, 2007), was developed with the goals of standardizing the vocabulary and semantics of biological-sequence annotation toward interoperability for software developers and users of genomic sequences, which had not been established (Eilbeck *et al.*, 2005). The currency of genomic annotation are the sequence features that provide anchor points to which to attach biological knowledge. Creating the SO has involved the naming and defining of sequence features and establishing topological relationships between these classes with respect to their positions on genomic sequences.

We present our recent developmental efforts seeking to address three issues of import for the SO: (1) representation of molecular versus abstract sequences; (2) integration with other OBOs, particularly the Chemicals of Biological Interest (ChEBI) ontology (de Matos *et al.*, 2010) the Protein Ontology (PRO) (Natale *et al.*, 2011), the RNA Ontology (RNAO) (Hoehndorf *et al.*, 2011), the Gene Ontology (GO) (The Gene Ontology Consortium, 2000), the Chemical Information Ontology (CHEMINF) (Hastings *et al.*, 2011), and the Information Artifact Ontology (IAO) ((http://code.google.com/p/information-artifact-ontology/); and (3) consistent representation of DNA, RNA, and peptide sequences and harmonizing their use in annotations. We anticipate that these efforts will not only result in a more consistent representation of biological sequences but also increased operability with other OBOs, which would be beneficial to the primary use case of sequence annotation and also to other applications, including natural-language processing (Bada and Hunter, 2010) and reasoning with multiple ontologies (*e.g.*, Blondé *et al.*, 2011).

2 RESULTS AND DISCUSSION

The SO is a resource actively maintained by a small group of curators responsive to the requirements and input of the sequence-annotation community; it is currently managed via an SVN repository, where users can download versioned releases and revisions (http://www.sequenceontology.org/resources/index.html). The developers of the SO have begun efforts to harmonize the SO with other resources (Mungall *et al*., 2011); this entails making the SO more consistent both internally and with respect to external resources. We discuss three foci of our recent efforts in this endeavor here.

2.1 Representation of Abstract versus Molecular Sequences

There has been considerable ambiguity with regard to the ontological nature of biological sequences, including the categories of sequences represented in the SO. Hoehndorf *et al.* have posited the existence of three types of sequence: (1) *abstract sequences* are abstract entities that are "independent of space and time: either [they] … are not located in space and time, or they are located everywhere and at all times"; there is, for example, only one instance of the abstract sequence ACA; (2) *syntactic sequences* are sequence representations such as those in biomedical databases and text representations; and (3) *molecular sequences* are physical chains of nucleotides or amino acids (Hoehndorf *et al.*, 2009). In an effort to integrate the SO with the Basic Formal Ontology (BFO) (Grenon *et al.*, 2004)—the upper-level ontology to which OBO developers commit—SO develop-

* To whom correspondence should be addressed

ers have elaborated that sequence types of the SO are *generically dependent continuants*, defined in the BFO as continuants dependent on one or more independent-continuant bearers; thus, a given sequence is an abstract instance (and from here on "abstract" is meant in a wider sense, not the specific sense of Hoehndorf *et al.*) that inheres in each corresponding molecular sequence (Mungall *et al.*, 2011).

However, there are several issues making this conceptualization as the basis for SO representation problematic. In their framing of SO concepts as generically dependent continuants, SO developers acknowledged a discordance in that sequence attributes, which are explicitly represented in the SO and used in the formal definitions of corresponding sequences, actually apply to the molecular sequences. For example, a `wild_type_rescue_gene` is a `rescue_gene` that `has_quality wild_type`; that is, "wild-type" describes molecular sequences, not the abstract sequences that refer to the molecular sequences. More straightforwardly, biologists fundamentally regard sequences such as genes, exons, mutations, transcripts, and peptides as molecular entities, as evidenced in, *e.g.*, their definitions in biology textbooks, and this conceptualization is reflected in the natural-language definitions for most of the SO classes in their current official state.

We argue that since the molecular sequences are the more fundamental concepts (indeed, the generically dependent sequences depend upon them for their existence), they should be explicitly represented. That being said, there are at least a small number of SO classes whose conceptualizations as molecular sequences do not seem sensible. For example, `match` is defined as a "region of sequence, aligned to another sequence with some statistical significance, using an algorithm such as BLAST or SIM4". (Annotations using this concept are typically used to provide supporting evidence to computational gene models.) For a given match, there may be a molecular sequence that directly corresponds to it (though there may not be, *e.g.*, if gaps are permitted in this conceptualization), but since this matching occurs computationally, it seems much more sensible to represent it as a type of abstract sequence. Thus, some abstract sequences will be needed to represent the full set of concepts of the SO. Our solution to this is to represent biological sequences in two parallel ontologies, one containing the large majority of classes that can exist as molecular sequences and the other containing the corresponding abstract sequences for all of these molecular sequences and also the small number of classes that make more sense as abstract sequences.

The former will be an evolution of the Sequence Ontology:Molecules (SOM) effort, a small ontology representing molecules of genomic origin (Mungall *et al.*, 2011), which will accommodate not only its current more circumscribed domain but also all of the molecular-sequence concepts to which the SO refers. It will therefore be renamed the Molecular Sequence Ontology (MSO), as SO concepts refer to molecular sequences (*i.e.*, sequences at the molecular level), but most of them refer to parts of molecules rather than proper molecules themselves. Significantly, the many formal cross-product definitions of SO concepts (*e.g.*, the aforementioned `wild_type_rescue_gene`) will be transferred to their corresponding MSO concepts, as these define the molecular sequences. As we discuss in the next section, the sequence concepts of this ontology will be the bridge to the GO, PRO, RNAO, and ChEBI ontology.

The corresponding abstract sequences of these molecular sequences will remain the province of the SO; therefore, SO concepts will continue to be generically dependent continuants. This has the advantage of minimizing disruption to annotation efforts with the SO, as all current SO terms will continue to exist in the SO (whereas the concepts that are more sensible as abstract sequences will not be correspondingly represented in the MSO). In an effort toward usability, corresponding abstract and molecular sequences will be identically named but use their respective namespaces. As the current SO cross-product definitions will be transferred to their corresponding MSO concepts, SO concepts will instead be formally defined in terms of analogous MSO concepts, as will be shown in the next section. Since SO concepts will be necessarily and sufficiently defined in terms of their corresponding MSO concepts, an OWL reasoner will be able to automatically generate the hierarchy of the former from the latter, so the two parallel sequence hierarchies will not have to both be manually curated. In addition to linking to the MSO, the concepts of the SO will be connected to the CHEMINF ontology, and thus indirectly to the IAO, as described in the next section.

2.2 Integration with ChEBI, PRO, RNAO, GO, CHEMINF, and IAO

Many of the OBOs have been impressively developed, but lack of formal linkage among them is a serious issue, and we seek to (directly or indirectly) link both the MSO and SO to other OBOs. As for the former, among the neighboring ontologies with which we envision integration are the ChEBI ontology and the PRO, RNAO, and GO. The first of these is the primary OBO representing molecules, molecular parts, atoms, subatomic particles, and biochemical roles and applications of these entities, and all MSO concepts will be subclasses of the ChEBI class `molecular entity`. The current official top-level sequence term in the SO is the fuzzily named `region`, defined as a sequence feature with an extent greater than zero, which will be renamed to the more precise `monomeric sequence`, *i.e.*, a sequence of biological monomers; this concept will be fundamentally subdivided into `monomeric sequence molecule`, representing sequences that are whole molecules, and `monomeric subsequence`, representing proper parts of monomeric sequence molecules. (We are aware that a mol-

ecule technically refers to an electrically neutral polyatomic entity and that biological sequences cannot be guaranteed to be electrically neutral (and likely are not); we are referring to a broader sense of molecules that is also reflected in the ChEBI term `macromolecule`, which has the concept of a molecule incorporated into its name but is also not guaranteed to be electrically neutral and is therefore not subsumed by `molecule`.) In biological parlance, "sequence" can refer to either a whole sequence or a proper subsequence, and this ambiguity is encapsulated in the top-level class `monomeric sequence`. (In fact, we can formally define this class as this union of `monomeric sequence molecule` and `monomeric subsequence`.) This fundamental subdivision of `monomeric sequence` allows us to more richly link the SO to ChEBI: Within the latter, sequences such as nucleic acids and peptides are represented as entire molecules but not as proper subsequences, and so this subdivision will enable us to assert the equivalency of specific existing ChEBI macromolecular classes and specific MSO subclasses of `monomeric molecule`. We can further link the MSO to ChEBI by defining sequence types in terms of their constituent monomers. We have created `has_proper_monomeric_part` as a subrelation of `has_proper_part` to use in such definitions, *e.g.*:

```
MSO:'peptide sequence' subclassOf
  MSO:'monomeric sequence' and
  has_proper_monomeric_part
    some CHEBI:'amino-acid residue' and
  has_proper_monomeric_part
    only CHEBI:'amino-acid residue'
```

The MSO will additionally be able to be linked to other OBOs representing more specific types of biological sequences. The PRO, an OBO which focuses on protein classes and complexes, could link to the MSO by making its top-level `protein` a subclass of `MSO:peptide sequence molecule`. (As the PRO also represents protein variants, isoforms, and modified forms, we envision that the MSO will be further linked to the PRO relying on our representation of sequence variation; a discussion of this is beyond the scope of this paper, but we have done preliminary work in a richer representation of sequence variation in the SO (Bada and Eilbeck, 2010).) Likewise, the RNAO will be able to be integrated with the MSO by subclassing its RNA-specific sequences and structures from the more general corresponding concepts of the MSO.

The molecular sequences represented in the MSO will also be able to be utilized by the GO: GO classes representing processes operating on sequences, particularly many subsumed by `macromolecule metabolic process` or `regulation of macromolecule metabolic process`, will be able to rely on relevant SO classes for their formal definitions. For example, the GO class `RNA processing` is currently informally defined as "[a]ny process involved in the conversion of one or more primary RNA transcripts into one or more mature RNA molecules"; it could be formally defined using the SO classes `primary transcript` and `mature transcript`, *e.g.*:

```
GO:'RNA processing' subclassOf
  GO:'biological_process' and
  part_of
    GO:'biological_process' and
    results_in_derivation_from
      some MSO:'primary transcript' and
    results_in_derivation_to
      some MSO:'mature transcript'
```

Here, we have defined an RNA processing as a biological process that is part of a biological process and that results in the derivation from at least one primary transcript to at least one mature transcript. (We have included the parthood expression to model the involvement mentioned in the informal definition, and we take advantage of the fact that something is part of itself for instances of RNA processing that are not proper parts of instances of composite RNA processing. Also, here we have created the occurrent-to-continuant relations `results_in_derivation_from` and `results_in_derivation_to` as extrapolations of the continuant-to-continuant relation `derives_from`, which is defined in the OBO Relation Ontology (Smith *et al.*, 2005)). Such GO definitions relying on the MSO, as well as the previously discussed MSO definitions relying on the ChEBI ontology can be seen as extensions of the OBO cross-product effort (Mungall *et al.*, 2011). There are a plethora of vetted (but still mostly unofficial) cross-product definitions among a number of OBOs (http://www.berkeleybop.org/ontologies/#logical_definitions) as well as those among concepts within the SO, but presently none among SO concepts and those of external ontologies, an issue that this proposal will help to address.

While the MSO will enable integration with ChEBI, PRO, RNAO, and GO, the SO concepts will be made subclasses of the class `information about a chemical entity` from the CHEMINF ontology, an OBO focusing on the representation of informational chemical entities manipulated in computational algorithms and procedures, as well as the algorithms and procedures themselves; the SO will thus be indirectly connected to the IAO, as this CHEMINF class is itself a subclass of the IAO's `information content entity` class, defined as "an entity that is generically dependent on some artifact and stands in relation of aboutness to some entity". We will use `denotes`, a subrelation of the IAO's fundamental `is_about` relation, to formally define the large majority of the concepts of the SO in terms of those of the MSO, *e.g.*:

```
SO:transcript subclassOf
  CHEMINF:'information about a
    chemical entity' and
  denotes some MSO:transcript
```

Thus, an abstract transcript sequence is a chemical information content entity that denotes a molecular transcript sequence. Hypothetical, improbable, and even impossible abstract sequences could be created, which may seem problematic given that we have modeled them as information content entities, which are defined to be "about" something. However, we consider this an orthogonal issue not limited to sequences, and there have been recent efforts to address this issue (Dumontier and Hoehndorf, 2010; Ceusters and Smith, 2010; Hastings *et al.*, 2011). As stated in the previous section, since these SO concepts will be formally defined in terms of their corresponding MSO concepts, the classification of the former will be able to be automatically generated from the latter.

2.3 Consistent Representation of DNA, RNA, and Peptide Sequences and Their Use in Annotation

Sequences are annotated overwhelmingly at the genomic level. Perhaps unintuitively, these sequences are annotated at the DNA level even with the many SO concepts at the RNA and peptide levels (*e.g.*, `splice site`, `polypeptide domain`), with the implied semantics that the RNA- or peptide-level concept holds for the RNA or peptide sequence that corresponds to the DNA sequence denoted by the annotated sequence. These RNA- and peptide-level classes are informally defined as RNA and peptide sequences, respectively, as one would expect, yet they are sometimes subsumed by DNA-level concepts; for example, `transcript` is a subclass of `gene member region`. As part of our efforts toward making the SO more consistent in terms of both the ontology itself as well as its use in sequence annotations, we are addressing this conceptual tangle by consistently representing these sequence types and preparing for their use by annotators.

It is clear that the natural-language definitions of these concepts should match their formal structure, and thus, either the RNA-level definition of `transcript` should change, or it should not be subsumed by a DNA-level concept. We argue that these classes should be defined as they are canonically conceptualized, so `transcript` should be defined at the RNA level. Its classification should also reflect this, so it should be subsumed by some more generic RNA concept rather than by `gene member region`. Therefore, we are properly classifying these concepts.

For this classification, we have created a set of sequence classes consistently defined in terms of type of monomer. Currently, monomer type is represented by a set of polymer attributes, and sequences are attributed these qualities; for example, `DNA`, `RNA`, and `peptidyl` are all subclasses of `polymer attribute`, and, *e.g.*, `RNA chromosome` is formally defined as a chromosome that has an RNA quality:

```
'RNA chromosome' subclassOf
  chromosome and
  has_quality some RNA
```

For each type of monomer, we are creating a primary sequence class (*e.g.*, `DNA sequence`, `RNA sequence`, `peptide sequence`), which was not previously explicitly represented. Rather than relying on qualities for specifying the monomer types of the sequences, we are using ChEBI classes that represent the monomers, as exemplified by the definition of `peptide sequence` in the previous section. A wealth of monomer types are already represented in ChEBI (including many noncanonical ones), so this strategy obviates the need for us to explicitly represent them. In addition to reducing effort on our end, it abides by the principle of orthogonality among ontologies of the OBO library. Monomeric sequences are thus subdivided along two orthogonal axes, namely, whether they are whole molecules or proper subsequences (as discussed in the previous section), and by monomer type. However, all these direct subclasses will be necessarily and sufficiently defined.

As previously stated, sequences are overwhelmingly annotated at the DNA level, but this includes the use of RNA- and peptide-level classes such as `splice site` and `polypeptide domain` to mark up DNA sequences. There are several strategies we can take to address this, one of which is to explicitly represent corresponding DNA, RNA, and peptide sequences, link them accordingly, and guide annotators to proper use of these classes. We anticipate that this would be a significant change for annotators, and so as to minimize confusion, we could name these new classes as the sequences on which they are based, appended with "DNA" and "RNA", as appropriate; for example, for `polypeptide domain`, we could create `polypeptide domain DNA` and `polypeptide domain RNA`, representing DNA and RNA sequences, respectively, corresponding to polypeptide domains.

There are several options as to how to link such concepts. One is to state each association as the product sequence being created from the template sequence, *e.g.*:

```
'polypeptide domain' subclassOf
  peptide sequence and
  created_from_template_sequence
    some 'polypeptide domain RNA'
```

This states that a polypeptide domain is a subclass of a peptide sequence that is created from an RNA sequence corresponding to a polypeptide domain as a template sequence, which seems odd and circular. The other option is to state each association in the reverse direction, *e.g.*:

```
'polypeptide domain RNA' subclassOf
  'RNA sequence' and
  template_for only 'polypeptide domain'
```

This formal definition seems sensible in that it is reflected in the name of the class. A disadvantage is that the `template_for` restriction is not existential (∃) in that not every RNA sequence corresponding to a polypeptide domain will get translated into a polypeptide domain. Rather, this would have to be made universal (∀), which we believe canonically holds. Relying on this option, we can then link the corresponding DNA and RNA sequences:

```
'polypeptide domain DNA' subclassOf
  'DNA sequence' and
  template_for
    only 'polypeptide domain RNA'
```

Thus, `polypeptide domain DNA` is the class that would be used to annotate a DNA sequence that currently is annotated with `polypeptide domain`. However, as creation and use of explicit corresponding sequences would be a significant change to the ontology and to the annotation process, we will seek community input with regard to this.

CONCLUSIONS

We have presented and discussed our recent efforts in the continuing development of the SO: (1) representation of molecular versus abstract sequences; (2) integration of the SO with ChEBI, PRO, RNA GO, CHEMINF, and IAO; and (3) consistent representation and use of corresponding DNA, RNA, and peptide sequences. In addition to increasing interoperability of the SO with other OBOs, we anticipate that this work will improve the consistency of the SO both internally and and with respect to external resources; these would strengthen the SO as a tool for reasoning with regard to its use toward its primary use case of sequence annotation as well as other applications. As these discussed changes significantly alter the structure and terminology of the ontology, a measured approach must be taken to allow time to update the existing software and protocols that rely on the SO.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of this work by NIH/NHGRI R01HG004341. We also thank Colin Batchelor for his participation in helpful discussions.

REFERENCES

Bada, M. and Eilbeck, K. (2010) Toward a Richer Representation of Sequence Variation in the Sequence Ontology. *Proc 2010 Eur Conf Comp Biol Annotation, Interpretation and Management of Mutations Wkshp*.

Bada, M. and Hunter, L. (2010) Desiderata for Ontologies to Be Used in Semantic Annotation of Biomedical Documents. *J Biomed Inform* **44**(1):94-101.

Blondé, W., Mironov, V., Venkatesan, A., Antenanza, E., De Baets, B., and Kuiper, M. (2011) Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinform* **27**(11):1562-1568.

Ceusters, W. and Smith, B. (2010) Foundations for a realist ontology of mental disease. *J Biomed Semantics* **1**(1), 10.

de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. *Nucleic Acids Res*, 38, D249–D254.

Dumontier, M. and Hoehndorf, R. (2010) Realism for scientific ontologies. *Proc 6th Internat Conf Formal Ontology in Info Systems*, 387-399.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* **6**:R44.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* **25**(1):25-29.

Grenon, P., Smith, B., and Goldberg, L. (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: *Ontologies in Medicine*, 20-38. IOS Press, Amsterdam.

Hastings, J., Batchelor, C., Neuhaus, F., and Steinbeck, C. (2011) What's in an 'is about' Link? Chemical Diagrams and the Information Artifact Ontology. *Proc Internat Conf Biomed Ontology*.

Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., and Dumontier, M. (2011) The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. *PLoS ONE* **6**(10):e25513.

Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, K., Mungall, C.J., Richardson, J.S., Stombaugh, J., Westhof, E., Zirbel, C.L., and Leontis, N.B. (2011) The RNAO Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology* **6**:53-89.

Hoehndorf, R., Kelso, J., and Herre, H. (2009) The ontology of biological sequences. *BMC Bioinform* **10**:377.

Mungall, C.J., Bada, M., Berardini, T.Z., Deegan, J., Ireland, A., Harris, M.A., Hill, D.P., and Lomax, J. (2011) Cross-product extensions of the Gene Ontology. *J Biomed Inform* **44**(1):80-86.

Mungall, C.J., Batchelor, C., and Eilbeck, K. (2011) Evolution of the Sequence Ontology terms and relationships. *J Biomed Inform* **44**:87-93.

Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J.A., Bult, C.J., Caudy, M., Drabkin, H.J., D'Eustachio, P., Evsikov, A.V., Huang, H., Nchoutmboube, J., Roberts, N.V., Smith, B., Zhang, J. and Wu, C.H. (2011) The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res* **39**(Database Issue):D539-545.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11), 1251–1255.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C.J., Neuhaus, F., Rector, A.L., and Rosse, C. (2005) Relations in biomedical ontologies. *Genome Biol* **6**(5):R4.