Toward a Richer Representation of Sequence Variation in the Sequence Ontology

Michael Bada[1] and Karen Eilbeck[2]

[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) is a member ontology of the Open Biomedical Ontologies (OBO) library that is charged with formally representing types of biomacromolecular sequences and their associated attributes as well as the interrelationships among these. By providing a common vocabulary and set of definitions, it is widely used to facilitate accurate storage, processing, and exchange of sequence data. While some parts of the SO are quite mature, particularly the independently defined sequence features, the current representation of sequence variation is more problematic in that the representation of corresponding variations, sequences containing these variations, and processes resulting in these variations and their interrelationships is incomplete. Additionally, corresponding variations in DNAs, RNAs, and polypeptides and their interrelationships are not represented. We report here on our progress in more completely and precisely representing these concepts, which will allow for more consistent annotation of variant sequence data. Furthermore, formally linking and defining these sequence-variation classes within the OBO framework will enable powerful, logically sound reasoning with various types of variant data as well as with other types of annotated biological data.

Introduction

The Sequence Ontology (SO) [1] is a member ontology of the Open Biomedical Ontologies (OBO) library (http://www.obofoundry.org) [2] that is charged with formally representing types of biomacromolecular sequences and their associated operations and attributes as well as the interrelationships among these. By providing a common vocabulary and set of definitions, the SO is able to be used toward systematic annotation of sequences in biological databases, and it is indeed widely used to facilitate accurate storage, processing, and exchange of sequence data (*e.g.*, http://www.sequenceontology.org/resources/databases.html). Additionally, the explicit specification of the interrelationships among these classes permits sound automated reasoning of data annotated with related SO classes.

Some portions of the SO, particularly the branches representing sequence features, which include subsequences and boundaries, are quite mature. However, the current representation of sequence variation is more problematic in that the representation of corresponding variations, sequences containing these variations, and processes resulting in these variations is incomplete. Additionally, corresponding variations in DNAs, RNAs, and polypeptides and their interrelationships are not represented. Since the SO is the primary OBO of types of biological sequences, which are essential for the representation of biological knowledge, a rigorous model delineating and linking these related concepts is needed, particularly as the OBOs are increasingly connected with each other and incorporated into richer knowledge representations. Furthermore, such a representation is required for consistent annotation of variant sequence data in biological databases. In particular, these terms are needed for the specification of sequences variations in the new Genome Variation Format version (GVF) [3], which seeks to serve as a common flat-file format for the annotation of genomic variants.

We report here on our progress in the representation of sequence variation in the SO. This entails making variations of DNAs, RNAs, and polypeptides and the relationships among them explicit. Additionally, variations, sequences containing these variations, and processes resulting in these variations are clearly arranged in their own hierarchies, and they are explicitly linked through their relationships. Necessary and sufficient definitions will be created for most of these concepts in terms of relations to more atomic concepts; this will allow for automatic classification of these terms by a reasoner, thus relieving the developers of the SO of a great deal of manual curation and reducing associated human error. The representation of these variant sequence types in all of their complexity will be a significant addition to the SO and resource for sequence-database annotators. Recognizing that not all users of the SO will need or want to work with this ontologically sound but admittedly involved representation, we plan to offer both this full version and a "light" version for those who likely will not need to take advantage of the sophisticated inference it enables.

Current Representation of Sequence Variation in the SO

In the SO, one of the most basic classes is sequence_feature, which is defined as an "extent of biological sequence"; this classes subsumes junction, which is a "sequence_feature with an extent of zero", as well as region, a "sequence_feature with an extent greater than zero"; the latter subsumes terms representing familiar concepts such as bases, genes, assemblies, repeats, and binding sites. Also subsumed by sequence_feature is sequence_alteration, which is "a sequence_feature whose extent is the deviation from another sequence"; thus, a sequence alteration is essentially semantically equivalent to a mutation (*i.e.*, the resulting entity, not the process); the vocabulary here reflects the commonly agreed adoption of less contentious terms to describe genetic change [4, http://www.hgvs.org]. There are currently 38 types of sequence alterations currently represented, including insertion, deletion, substitution, inversion, translocation, and copy-number variation.

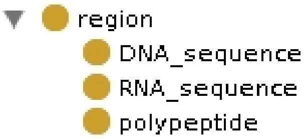
Another basic SO class is sequence_variant, which is a nonexact copy of a sequence feature or genome exhibiting one or more sequence alterations; thus, a sequence variant is defined in terms of its sequence alteration(s). Types of sequence variants currently include functional variants such as those capable of affecting transcription and/or translation and structural variants such as splicing variants and variants capable of affecting polypeptide structure.

Currently, the sequence variants and sequence alterations are orthogonal and unconnected: There are no represented sequence variants for currently represented corresponding types of sequence alterations; for example, a point mutation is an explicitly represented type of sequence alteration, but there is no represented variant containing a point mutation. Conversely, there are no represented sequence alterations for currently represented corresponding sequence variants; for example, a polypeptide posttranslational-processing variant is an explicitly represented type of sequence variant, but there is no correspondingly represented sequence alteration. This is problematic in that this orthogonality forces database curators to annotate mutational sequence data either at the sequence-alteration level or at the sequence-variant level dependent on the type of variation.

Additionally, there is currently no separation of the corresponding DNA, RNA, and polypeptide sequence alterations and sequence variants nor their interrelationships. This is important because some sequence alterations and variants are appropriate only at some of these levels. For example, a silent mutation is appropriate at the DNA and RNA levels but not at the polypeptide level, as a silent mutation does not result in an amino-acid change in the resulting polypeptide; thus, there is no resulting polypeptide alteration or variant. As another example, it would not be appropriate to use one of the transcript-variant terms (*e.g.*, transcript_processing_variant) at the polypeptide level. If corresponding DNA, RNA, and polypeptide sequence alterations and variants are separated and correctly linked, annotation tools could then use this accurately represented sequence knowledge to aid database curators by helping them avoid using terms not appropriate for their task, *e.g.*, by allowing them to only use terms at the DNA level if they are annotating DNA sequences.

A More Refined Representation of Sequence Variation in the SO

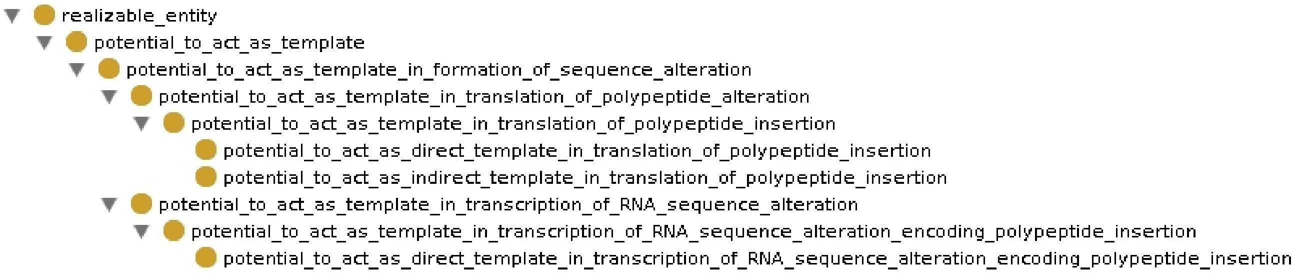
Our proposal for a more refined representation of sequence variation in the SO begins with a straightforward classification of DNA_sequence, RNA_sequence, and polypeptide as subclasses of region (which is synonymous with sequence), as seen in Figure 1.



- region
    - DNA_sequence
    - RNA_sequence
    - polypeptide

Figure 1. Partial hierarchy for region, which is synonymous with sequence[1].

Another basic set of classes required for this representation is a hierarchy of realizable entities. OBOs are committed to using the Basic Formal Ontology (BFO) as a common upper ontology of general concepts to maximize consistency and interoperability among the various OBOs [5]. Within the BFO is a class named realizable_entity, an instance of which is an abstract entity that inherently exists within another entity or group of entities but is not exhibited in full at every time in which it exists within the entity or group of entities; this includes dispositions, functions, and roles. We next introduce a hierarchy of potentials to act as templates in the formation of various types of sequence alterations, as shown in Figure 2. For this part of the representation and for the following parts, we present the generic framework for all sequence alterations and variants; additionally, as a specific example of a type of sequence variation currently represented in the SO, we include the corresponding classes for polypeptide insertion.



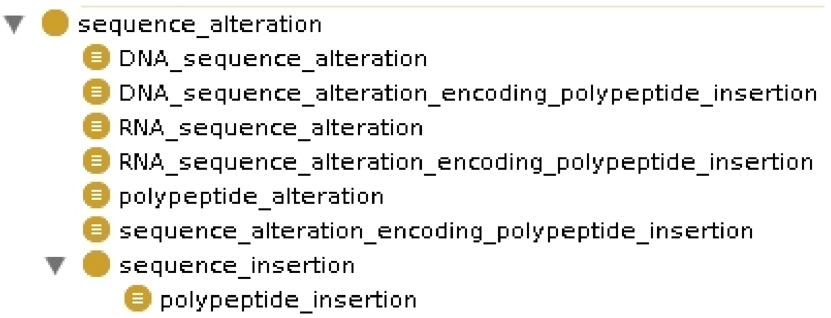
- realizable_entity
    - potential_to_act_as_template
        - potential_to_act_as_template_in_formation_of_sequence_alteration
            - potential_to_act_as_template_in_translation_of_polypeptide_alteration
                - potential_to_act_as_template_in_translation_of_polypeptide_insertion
                    - potential_to_act_as_direct_template_in_translation_of_polypeptide_insertion
                    - potential_to_act_as_indirect_template_in_translation_of_polypeptide_insertion
            - potential_to_act_as_template_in_transcription_of_RNA_sequence_alteration
                - potential_to_act_as_template_in_transcription_of_RNA_sequence_alteration_encoding_polypeptide_insertion
                    - potential_to_act_as_direct_template_in_transcription_of_RNA_sequence_alteration_encoding_polypeptide_insertion

Figure 2. Hierarchy of potentials to act as templates in the formation of sequence alterations.

These potentials can be used to formally define sequence alterations, specific types of which are presented in Figure 3. Note that this is a mostly flat classification of these sequence alterations at this stage. Note also that all but one of these subsumed sequence-alteration classes are necessarily and sufficiently (*i.e.*, completely) defined, as indicated by the equivalency characters ("≡") rendered within the orange-circle class icons, to which we shall return later. (Conversely, sequence_insertion is here directly asserted to be a subclass of sequence_alteration.)

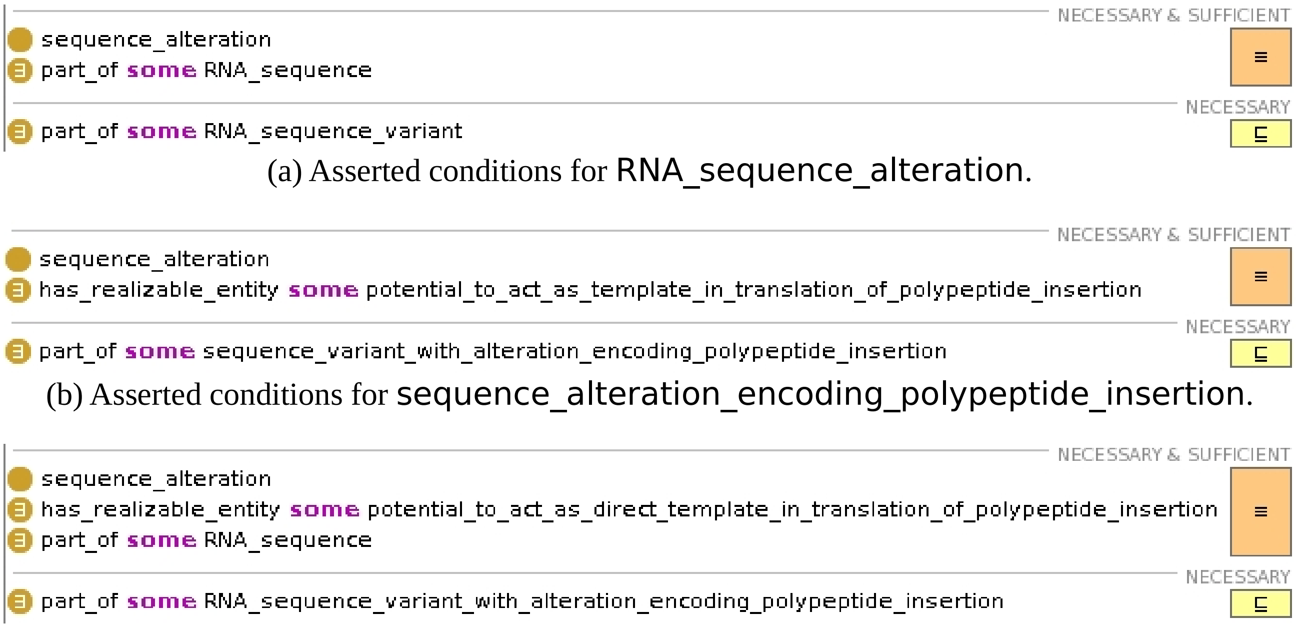
Figure 4 shows the sets of asserted conditions for three of the completely defined types of sequence alterations. For example, in Figure 4(a), a RNA sequence alteration is necessarily and sufficiently defined as a sequence alteration that is a part of a RNA sequence; furthermore, it is necessarily asserted to be part of an RNA sequence variant (which will be discussed later). Note that Figures 4(b) and 4(c) depend upon realizable entities for their definitions. For example, in Figure 4(b), a sequence alteration encoding a polypeptide insertion is a sequence alteration that has the potential to act as a template in the translation of a polypeptide insertion.

[1] All ontology figures are screenshots from Protege-OWL. An orange circle indicates a class, and a subclass is shown as indented relative to a superclass.



- sequence_alteration
  - DNA_sequence_alteration
  - DNA_sequence_alteration_encoding_polypeptide_insertion
  - RNA_sequence_alteration
  - RNA_sequence_alteration_encoding_polypeptide_insertion
  - polypeptide_alteration
  - sequence_alteration_encoding_polypeptide_insertion
  - sequence_insertion
    - polypeptide_insertion

Figure 3. Hierarchy of sequence alterations.



NECESSARY & SUFFICIENT

sequence_alteration
part_of **some** RNA_sequence ≡

NECESSARY

part_of **some** RNA_sequence_variant ⊑

(a) Asserted conditions for RNA_sequence_alteration.

NECESSARY & SUFFICIENT

sequence_alteration
has_realizable_entity **some** potential_to_act_as_template_in_translation_of_polypeptide_insertion ≡

NECESSARY

part_of **some** sequence_variant_with_alteration_encoding_polypeptide_insertion ⊑

(b) Asserted conditions for sequence_alteration_encoding_polypeptide_insertion.

NECESSARY & SUFFICIENT

sequence_alteration
has_realizable_entity **some** potential_to_act_as_direct_template_in_translation_of_polypeptide_insertion
part_of **some** RNA_sequence ≡

NECESSARY

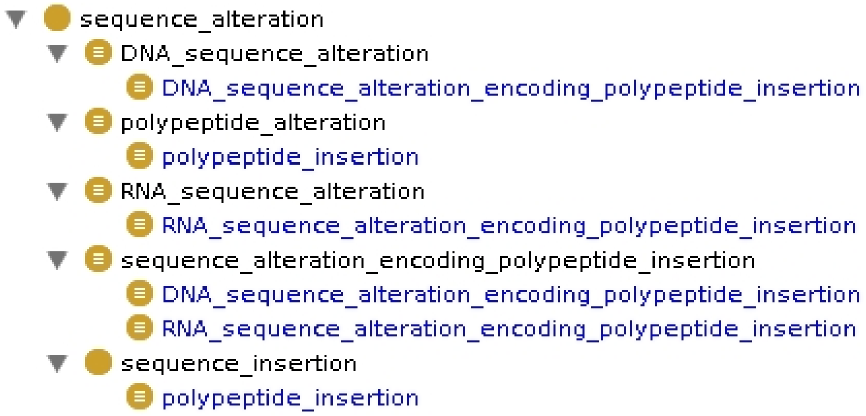
part_of **some** RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion ⊑

(c) Asserted conditions for RNA_sequence_alteration_encoding_polypeptide_insertion.

Figure 4. Asserted conditions for three completely defined sequence alterations.

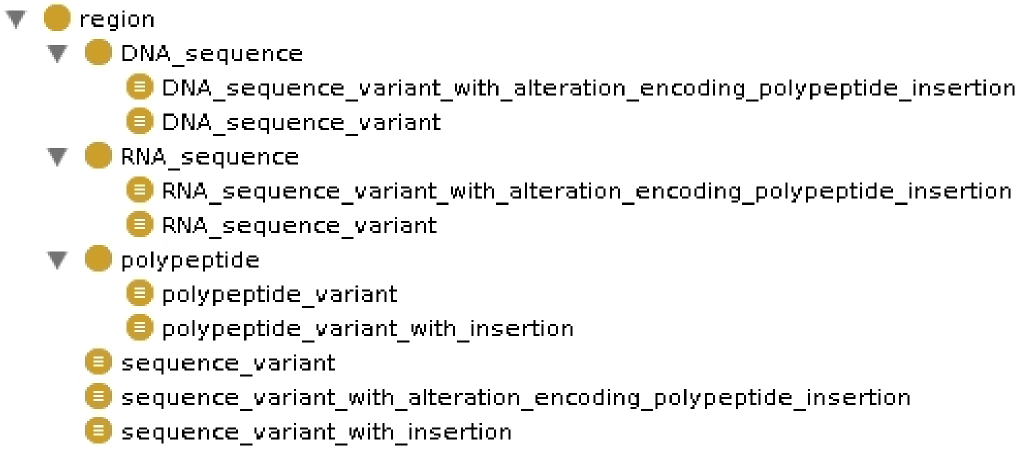
A principal advantage of creating necessary and sufficient definitions is that a reasoner can precisely classify terms with such definitions, relieving the ontology developer of manual effort and reducing human errors, particularly when working with large, complex ontologies. Figure 5 shows the hierarchy of sequence alterations after being classified by a reasoner, in which classes whose positions in the hierarchy have changed as supported by newly made inferences are rendered in blue. For example, RNA_sequence_alteration_encoding_polypeptide_insertion has been inferred to be a subclass of RNA_sequence_alteration, even though this was not directly asserted, as seen in Figure 4(c).

Sequence variants can now be completely defined in terms of the sequence alterations they possess; that is, they are regions (*i.e.*, sequences) with sequence alterations. Figure 6 shows the (unclassified) hierarchy of sequence variants, subsumed by types of region. Note that each type of sequence variant is completely defined. Furthermore, sequence variants may possess realizable entities corresponding to those possessed by their respective alterations, so a hierarchy of realizable entities for sequence variants is introduced in Figure 7.



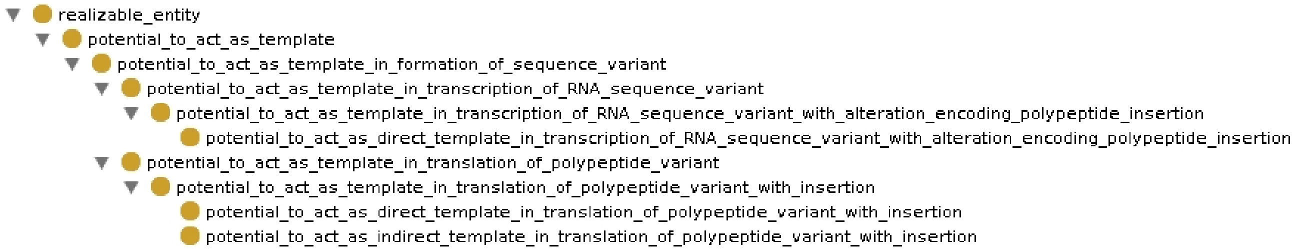
```
sequence_alteration
  DNA_sequence_alteration
    DNA_sequence_alteration_encoding_polypeptide_insertion
  polypeptide_alteration
    polypeptide_insertion
  RNA_sequence_alteration
    RNA_sequence_alteration_encoding_polypeptide_insertion
  sequence_alteration_encoding_polypeptide_insertion
    DNA_sequence_alteration_encoding_polypeptide_insertion
    RNA_sequence_alteration_encoding_polypeptide_insertion
  sequence_insertion
    polypeptide_insertion
```

Figure 5. Automatically classified hierarchy of sequence alterations.



```
region
  DNA_sequence
    DNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
    DNA_sequence_variant
  RNA_sequence
    RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
    RNA_sequence_variant
  polypeptide
    polypeptide_variant
    polypeptide_variant_with_insertion
  sequence_variant
  sequence_variant_with_alteration_encoding_polypeptide_insertion
  sequence_variant_with_insertion
```

Figure 6. Hierarchy of sequence variants.

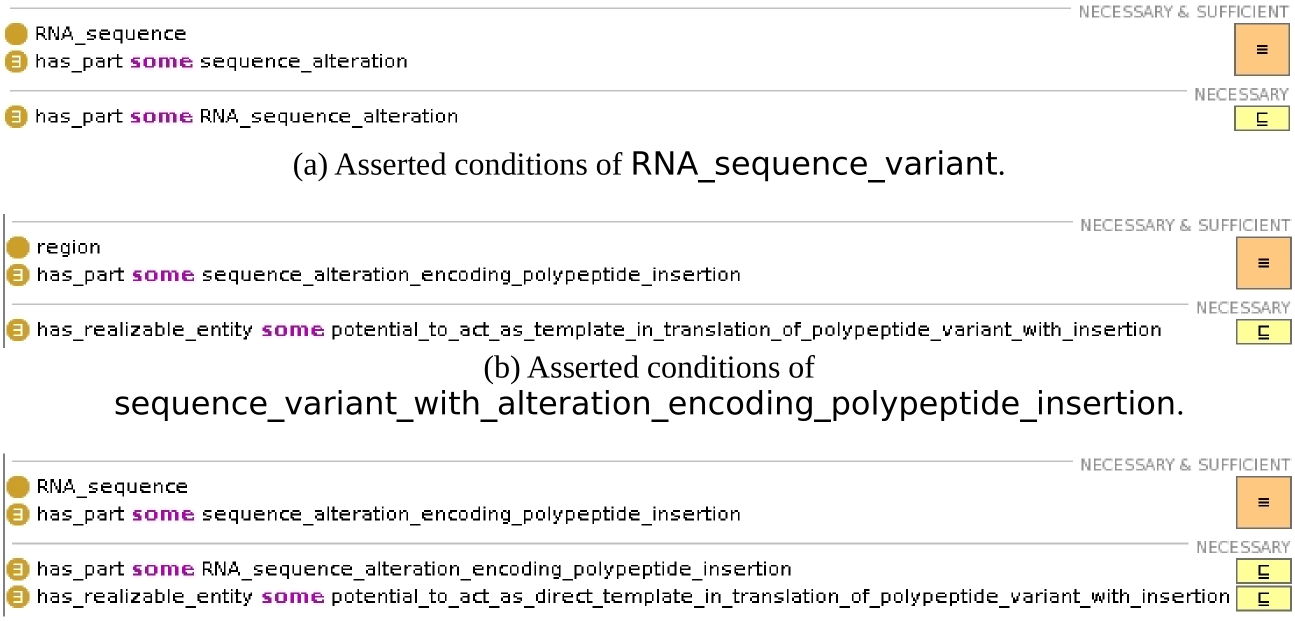


```
realizable_entity
  potential_to_act_as_template
    potential_to_act_as_template_in_formation_of_sequence_variant
      potential_to_act_as_template_in_transcription_of_RNA_sequence_variant
        potential_to_act_as_template_in_transcription_of_RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
          potential_to_act_as_direct_template_in_transcription_of_RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
      potential_to_act_as_template_in_translation_of_polypeptide_variant
        potential_to_act_as_template_in_translation_of_polypeptide_variant_with_insertion
          potential_to_act_as_direct_template_in_translation_of_polypeptide_variant_with_insertion
          potential_to_act_as_indirect_template_in_translation_of_polypeptide_variant_with_insertion
```

Figure 7. Hierarchy of potentials to act as templates in the formation of sequence variants.

Figure 8 shows the definitions for three of these sequence variants. For example, in Figure 8(a), a RNA sequence variant is necessarily and sufficiently defined to be an RNA sequence that has a sequence alteration; additionally, it is necessarily asserted to have an RNA sequence alteration. Furthermore, a sequence variant class may have an additional necessary condition asserting that an instance of the sequence variant has a potential to act as a type of template. For example, a sequence variant with an alteration encoding a polypeptide insertion has the potential to act as a template in the formation of a polypeptide variant with an insertion, as shown in Figure 8(b).

Figure 9 shows these sequence variants classified within the region hierarchy by a reasoner, in which classes whose positions in the hierarchy have changed as supported by new inferences are rendered in blue, as before. For example, RNA_sequence_variant_with_alteration_ encoding_ polypeptide_insertion has been inferred to be a subclass of RNA_sequence_variant, even though this was not directly asserted, as seen in Figure 9(c).



NECESSARY & SUFFICIENT

RNA_sequence
has_part some sequence_alteration ≡

NECESSARY

has_part some RNA_sequence_alteration ⊑

(a) Asserted conditions of RNA_sequence_variant.

NECESSARY & SUFFICIENT

region
has_part some sequence_alteration_encoding_polypeptide_insertion ≡

NECESSARY

has_realizable_entity some potential_to_act_as_template_in_translation_of_polypeptide_variant_with_insertion ⊑

(b) Asserted conditions of sequence_variant_with_alteration_encoding_polypeptide_insertion.

NECESSARY & SUFFICIENT

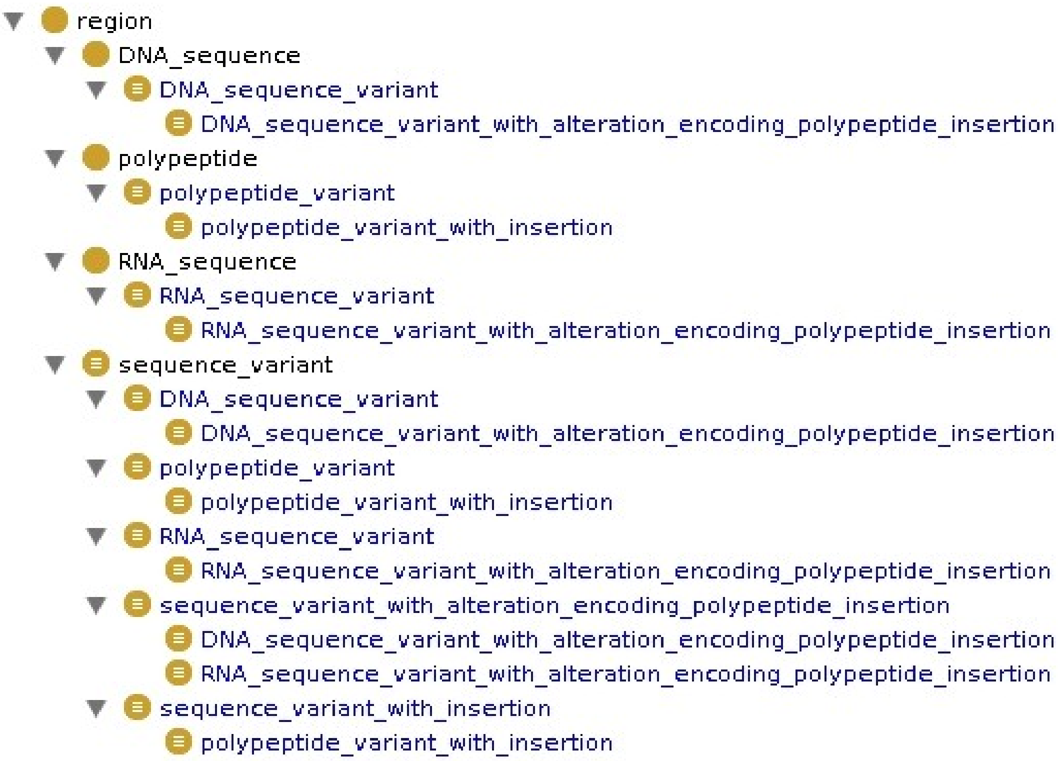
RNA_sequence
has_part some sequence_alteration_encoding_polypeptide_insertion ≡

NECESSARY

has_part some RNA_sequence_alteration_encoding_polypeptide_insertion ⊑
has_realizable_entity some potential_to_act_as_direct_template_in_translation_of_polypeptide_variant_with_insertion ⊑

(c) Asserted conditions of RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion.

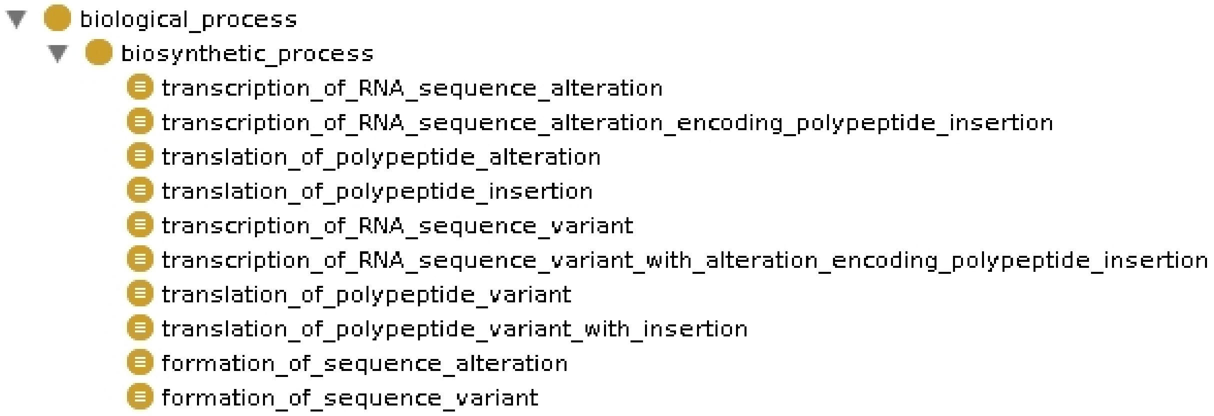
Figure 8. Asserted conditions for three completely defined sequence variants.



```
▼ region
    ▼ DNA_sequence
        ▼ DNA_sequence_variant
              DNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
    ▼ polypeptide
        ▼ polypeptide_variant
              polypeptide_variant_with_insertion
    ▼ RNA_sequence
        ▼ RNA_sequence_variant
              RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
    ▼ sequence_variant
        ▼ DNA_sequence_variant
              DNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
        ▼ polypeptide_variant
              polypeptide_variant_with_insertion
        ▼ RNA_sequence_variant
              RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
        ▼ sequence_variant_with_alteration_encoding_polypeptide_insertion
              DNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
              RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
        ▼ sequence_variant_with_insertion
              polypeptide_variant_with_insertion
```

Figure 9. Automatically classified hierarchy of sequences, including sequence variants.

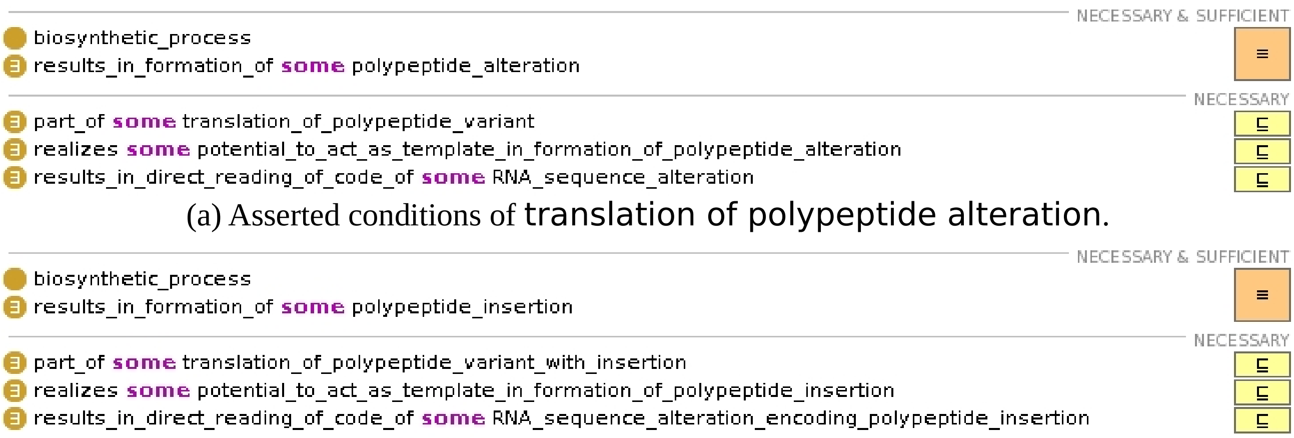
We have now accurately modeled the relationships among corresponding sequence variants and alterations. To accurately model the relationships among alterations at different stages of expression and among variants at different stages of expression, we introduce classes representing the formation of these alterations and variants, which we assert here to be subclasses of biosynthetic_process, a term in the Gene Ontology [6], as shown in Figure 10. For example, a term representing the translation of a polypeptide substitution would link the RNA alteration encoding the substitution to the polypeptide substitution itself. Although this knowledge is not likely to be directly used by sequence-database annotators, it would enable powerful, sound inferences using types of sequence alterations and variants annotated at different stages of expression. Note that all of these classes representing the formation of alterations and variants are completely defined and the hierarchy as directly asserted is completely flat (*i.e.*, under biosynthetic_process).



- biological_process
  - biosynthetic_process
    - transcription_of_RNA_sequence_alteration
    - transcription_of_RNA_sequence_alteration_encoding_polypeptide_insertion
    - translation_of_polypeptide_alteration
    - translation_of_polypeptide_insertion
    - transcription_of_RNA_sequence_variant
    - transcription_of_RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
    - translation_of_polypeptide_variant
    - translation_of_polypeptide_variant_with_insertion
    - formation_of_sequence_alteration
    - formation_of_sequence_variant

Figure 10. Hierarchy of classes representing the formation of sequence alterations and variants.

Figure 11 shows the definitions for two of these formation terms; each has a necessary and sufficient definition in terms of the entity that is created. For example, in Figure 11(a), a translation of a polypeptide alteration is defined to be a biosynthetic process that results in the formation of a polypeptide alteration. Each term representing a formation of a sequence alteration is also necessarily asserted to be part of a formation of the corresponding sequence variant, thus linking these two processes; for example, a translation of a polypeptide alteration is part of a translation of a polypeptide variant, as in Figure 11(a). Additionally, each of these formation terms has an asserted necessary condition that it realizes some realizable entity; for example, in Figure 11(a), it is asserted that a formation of a polypeptide alteration realizes a potential to act as a template in the formation of a polypeptide alteration; that is, this specific potential is realized in this process. Finally, a formation term may also have an asserted necessary condition indicating the template used in the formation process; for example, in Figure 11(a), it is asserted that a translation of a polypeptide alteration results in the direct reading of the code of an RNA sequence alteration. With these complete definitions of these biosynthetic processes, a reasoner can make inferences and automatically classify them precisely, as can be seen in Figure 12.



NECESSARY & SUFFICIENT

biosynthetic_process
results_in_formation_of **some** polypeptide_alteration ≡

NECESSARY

part_of **some** translation_of_polypeptide_variant ⊑
realizes **some** potential_to_act_as_template_in_formation_of_polypeptide_alteration ⊑
results_in_direct_reading_of_code_of **some** RNA_sequence_alteration ⊑

(a) Asserted conditions of translation of polypeptide alteration.

NECESSARY & SUFFICIENT

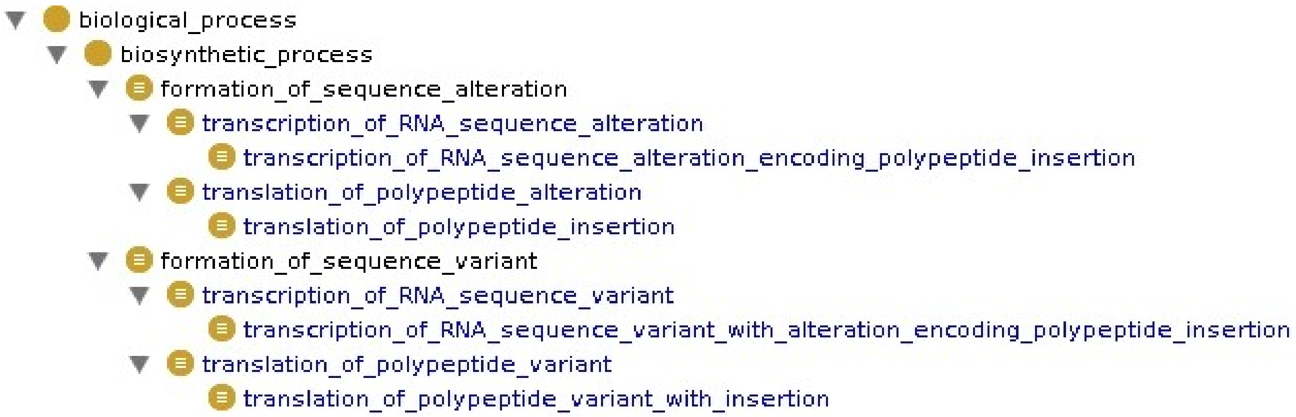
biosynthetic_process
results_in_formation_of **some** polypeptide_insertion ≡

NECESSARY

part_of **some** translation_of_polypeptide_variant_with_insertion ⊑
realizes **some** potential_to_act_as_template_in_formation_of_polypeptide_insertion ⊑
results_in_direct_reading_of_code_of **some** RNA_sequence_alteration_encoding_polypeptide_insertion ⊑

(b) Asserted conditions of translation of polypeptide insertion.

Figure 11. Asserted conditions for two completely defined types of translation.



```
biological_process
  biosynthetic_process
    formation_of_sequence_alteration
      transcription_of_RNA_sequence_alteration
        transcription_of_RNA_sequence_alteration_encoding_polypeptide_insertion
      translation_of_polypeptide_alteration
        translation_of_polypeptide_insertion
    formation_of_sequence_variant
      transcription_of_RNA_sequence_variant
        transcription_of_RNA_sequence_variant_with_alteration_encoding_polypeptide_insertion
      translation_of_polypeptide_variant
        translation_of_polypeptide_variant_with_insertion
```

Figure 12. Automatically classified hierarchy of types of formation.

Conclusions

We have presented examples of our efforts in creating a richer representation of sequence variation in the Sequence Ontology. In this representation, every type of sequence variant has a corresponding sequence alteration to which it is linked, thus giving sequence database curators the option of annotating at either level of granularity and at the same time allowing accurate processing of these different types of data. Additionally, sequence alterations and variants can be defined in terms of their potentials to effect processes, which are realized in these processes. Furthermore, sequence alterations and variants are created at their corresponding DNA, RNA, and polypeptide levels and linked together by their corresponding biosynthetic processes; this will allow annotators to mark up sequence data at any of these stages of expression while using a representation that will allow for sound reasoning using any of these types of data.

Acknowledgments

The authors are grateful for their support from NIH 5R01 GM083649-03 and P41 HG002273.

References

[1] 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 Biology* **6**:R44.

[2] 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. *Nature Biotechnology*, **25**(11):1251-1255.

[3] Reese, M. G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G., Stein, L., Flicek, P., Yandell, M., and Eilbeck, K. A standard variation file format for human genome sequences. *Genome Biology,* in press.

[4] den Dunnen, J. T. and Antonarakis, S. E. (2001) Nomenclature for the description of human sequence variations. *Human Genetics,* **109**(1):121-124.

[5] Grenon, P., Smith, B., Goldberg, L. (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli, D.M., ed. Ontologies in Medicine. Amsterdam: IOS Press, 20-38.

[6] Ashburner, M., Ball, C. A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.