SO SO : Sorting Out the Sequence Ontology

In order to sort out some of the problems and holes in the Sequence Ontology (SO) [1], it helps to define some of the kinds of people who would want to use SO as well as the ways in which they might want to use it.

We can start with two examples of potential SO users who might like to *author* new content using the SO ontology.

1. An experimental biologist wants to express a fact about an *actual, existing* piece of DNA or biological sequence characterized or used in her lab.
   1. “I want to assert that a piece of DNA with the following sequence was inserted into that cell.”
   2. “I want to assert that this organism has a particular gene variant.”
   3. “I want to assert that this sequence was read by a sequencer from a sample derived from this cell population.”
2. A synthetic biologist who wants to describe the design of a *theoretical* DNA construct, a new sequence designed for some purpose or function as part of a larger engineering project.
   1. “I want to assert that this DNA construct has the following sequence, and is composed of the following parts.”
   2. “I want to assert that this part of a DNA construct is adapted from this particular DNA sequence in an existing organism.”

There are also potential users who would be interested in *consuming* previously-authored data formulated in the SO ontology.

1. A bioinformatician downloads SO-compliant datasets from other investigators, and tries to combine them into a consistent set of facts about one or more organisms.
   1. “I want to know if these two different strains of yeast contain the same gene?”
   2. “I want to know where are the SNPs in the coding region of this gene?”
   3. “I want to know which portions of the DNA correspond to exons of the edited transcript for this protein?”
   4. “I want to know what the ‘canonical’ sequence for a gene is, and whether the gene’s form in this organism differs from that sequence?”
2. An ontologist examines the SO ontology directly, in order to understand the nature of biological sequence, its features, and their relationships.
   1. “I want to know what kinds of DNA variation exist?”
   2. “I want to know what query, on an SO-formulated knowledgebase, will tell me the complete list of genes of an organism described in that knowledgebase?”
   3. “I want to know what are the known variants of a particular gene?”

Given these different kinds of users, and their different potential uses for the SO ontology, we can ask the question: is SO, as currently formulated, suitably complete and well-defined as to make these uses possible?

We believe that the answer is “no,” and that there are a number of problems preventing SO from being immediately useful. We outline each of these problems below and show how they prevent the use-cases we have outlined above from being accomplished. Finally, we will finish with suggestions for how SO can be improved or extended, in order to make it more easily useful to a wider community.

**Problem #1: SO “features” are *generically dependent continuants* which depend on biological molecules for their existence.**

A quote from some of the maintainers of SO is informative:

“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.” [2] pg. 6

It does not help that this work [2] appears to define terms such as “sequence feature” and “biological sequence” using the term “sequence,” but nowhere defines what a “sequence” actually is (or equivalently, how we recognize whether some individual is a “sequence” *sensu SO*).

This requirement of molecular correlates to SO individuals appears to immediately rule out its use in *synthetic* biology. Synthetic biology involves, among other things, the design and synthesis of DNA molecules engineered to have particular functions or roles within a cell. Although these artificial sequences may have *parts* which were adapted from existing sequences characterized in real organisms, the whole sequence may have been designed *in silico* without existing real molecular counterpart. Synthetic biologists would like to create and exchange the designs for these artificial sequences, and the annotations of their parts, even before actual organisms have been created with genomes that carry these designed parts.

Additionally, this requirement creates difficulties for expressing facts about “canonical sequences.”

One way around this is to say that synthetic biologists are talking about “text sequences” (so-called *information content entities* in the BFO-world), not “biological sequences.” But this answer leads to our second problem with SO.

**Problem #2: SO contains no explicit terms which could be used to annotate *information entities* (such as text sequences), and no description of how such annotation terms could be tied to SO terms.**

In SO terminology, we can easily write that some particular piece of DNA “is” a particular gene. To take the example of “NCBI Gene ID 6469 (human Shh)” from [2], and using an RDF rendering of the appropriate assertions, we can say that

dna:x rdf:type SOM:dna\_molecule .

dna:x ro:bearer\_of [ ro:is\_concretization\_of [ rdf:type gene:6469 ] ] .

However, consider an example line from a GFF file [4]

ctg123 . exon 5000 5500 . + . ID=exon00004

The third field, the “type” of the GFF record, contains a reference (by label) to the SO term *exon*, a subclass of the *feature* term. There is no account in the SO documentation or literature of what this annotation “means” in terms of text sequences, molecules, or anything else.

This record can’t name a particular feature of the exon type, since it appears to have coordinates; coordinates can only be attributes of text sequences, not molecules or biological sequences (i.e. coordinates change with versions of the reference sequence, but this presumably does not entail changes in any correlated molecular structure).

[ elaborate a description of how this could be tied, through text sequences that match biological sequences, to a particular genome somewhere. but no accounting for variation from the “reference,” so that’s weird too. ]

Any account of how text-sequence annotation can be translated into annotation with SO terms leads [as I will have shown] to a description of how text sequences “match up” with biological sequences. Unfortunately, these leads to our next problem.

**Problem #3: The SO account of sequence matching and variation is incoherent and incomplete.**

SO has classes of feature that denote variants. In particular, it has classes that indicate “deletions” or “rearrangements.”

However, SO features must inhere in actual molecules – so what does a “deletion” feature inhere in?

**Problem #4: SO explicitly omits accounts of sequence *patterns*, and confuses *sequences* with *functional qualities*.**

What is a “response element?” If it’s functional, how can it be a *sequence* feature? If it’s a sequence pattern, how do we describe “motifs” that capture classes of sequence patterns?

Solutions

Distinguish text sequences from biological sequences.

Distinguish function from sequence.

Add an account of “sequence alignments” as a way of defining variation.

Talk about “plans” and “processes.”

Use classes, not individuals, to build descriptions.

Say it with “only.”

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 l2005;6: R44.

[2] Mungall CJ, Batchelor C, Eilbeck K. Evolution of Sequence Ontology terms and relationships. Journal of Biomedical Informatics *in press*.

[3] Hoehndorf R, Kelso J, Herre H. The ontology of biological sequences. BMC Bioinformatics l2009;10: 377.

[4] GMOD Wiki. “GFF” http://gmod.org/wiki/GFF3