Recent Developments in the Sequence Ontology

Colin Batchelor

2012-10-30

What is the Sequence Ontology?

It describes **genomic features**, such as

- Parts of gene models
- Transposons
- Assembly components
- Experimental results relating to genome sequence
- Kinds of variants
- Effects of variants
- Locations of variants

Who uses it?

GMOD schemas, formats (GFF3 and GVF) and tools rely on SO to type features.

As a rich vocabulary of genomic entities such as transcripts, genes, RNAs and so forth, it is very useful in **natural language processing**.

(That is how I got involved.)

Recent points of interest

Variation work – collaboration with Ensembl (see next talk)

Getting our ontological commitments right – most of this talk.

Getting our ontological commitments right

SO is a bit of a jumble of

- **Chemical things** (residues, bases, caps)
- **Dispositional things** (genes, splice sites)
- **Immaterial things** (junctions)
- **Epistemological things** (contig, golden path, virtual sequence)
- **Experimental artefacts** (PCR product, clone, engineered plasmid)

Getting our ontological commitments right

The **leaves** in the ontology are more or less OK, but the state of the tree makes the ontology difficult to use for integration and, perhaps more importantly, makes it difficult to **maintain**.

A first attempt (2009)

I (Colin) started work on SOM (Sequence Ontology: Molecules) which turned out to be much more work than I'd anticipated!

The underlying idea is to reuse ChEBI for the chemistry and GO for the processes.

A first attempt (2009): examples

```
[Term]
id: SOM:0002000
name: ribonucleic acid
def: "A nucleic acid that consists of ribonucleotide subunits." [SO:cb]
comment: this is to supersede CHEBI:33697.
intersection_of: SOM:0002003 ! nucleic acid
intersection_of: has_repeating_subunit CHEBI:55366 ! ribonucleotide residue

[Term]
id: SOM:0000031
name: aptamer
def: "An oligonucleotide that has been selected from a random pool based on
its ability to bind other molecules." [SO:cb]
intersection_of: SOM:0002003 ! nucleic acid
intersection_of: has_disposition SOM:0001998 ! selective binding disposition
```

A first attempt (2009)

If you're really interested in SOM you can get hold of it here:

http://song.cvs.sourceforge.net/viewvc/song/ontology/som.obo?revision=1.5

Current attempt (2011 to present)

Mike Bada (UC Denver) is **funded** to do this.

- SOM to become MSO (Molecular Sequence Ontology) and grow **lots of entities**
- MSO will be a bridge to GO, PR, RNAO and ChEBI.
- The abstract sequences in SO will be formally defined in terms of MSO
- GO will then be able to use MSO entities in definitions!

Sequences as information

- SO entities will be children of “information about a chemical entity” in CHEMINF.
- This is a child of “information content entity” in IAO.

Not within scope
(at least not yet)

Proper formal definitions of annotations, such as in GFF3.

Examples from outwith SO:

FALDO:

https://github.com/JervenBolleman/FALDO

SIO:

http://code.google.com/p/semanticscience/wiki/SIO

Resources

http://www.sequenceontology.org/

The tracker!
http://sourceforge.net/tracker/?atid=810408&group_id=72703&func=browse

File formats:
http://www.sequenceontology.org/resources/gff3.html
http://www.sequenceontology.org/resources/gvf.html
http://www.sequenceontology.org/resources/gvfclin.html