Recent developments in the Sequence Ontology

Colin Batchelor
Royal Society of Chemistry, UK
batchelorc@rsc.org

Karen Eilbeck
University of Utah, US
eilbeck@fruitfly.org

Overview

- What is SO?
- More-biologically-motivated relations
- Alignment with RO and topological relations
- Alignment with BFO: ontology curation experiments
- Open discussion

What is SO? (1)

The Sequence Ontology organizes the kinds of, parts of and properties of biological sequence.

It deals with universal types.

The instances may be DNA molecules, RNA molecules or polypeptides.

What is SO? (2)

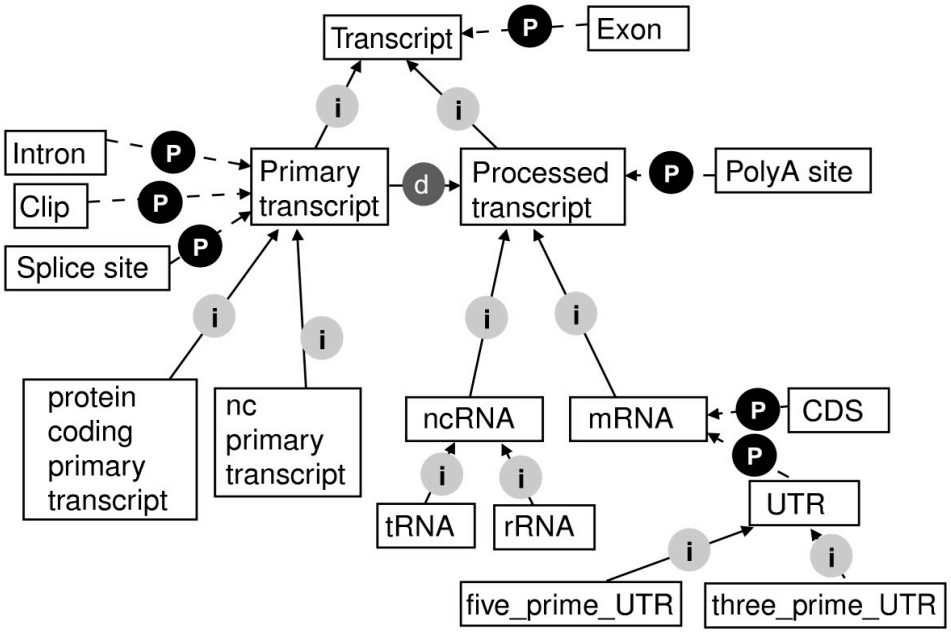
SO has been a pioneer in computable (cross-product) definitions of terms.

```
[Term]
id: SO:0000078
name: polycistronic_transcript
def: "A transcript that is polycistronic." [SO:xp]
intersection_of: SO:0000673 ! transcript
intersection_of: has_quality SO:0000880 ! Polycistronic
```

We are not restricted to dependent continuants for our differentiae:

```
[Term]
id: SO:0000111
name: transposable_element_gene
def: "A gene encoded within a transposable element. For example gag, int,
    env and pol are the transposable element genes of the TY element in
    yeast." [SO:ke]
intersection_of: SO:0000704 ! gene
intersection_of: part_of SO:0000101 ! transposable_element
```

Relations: the original set

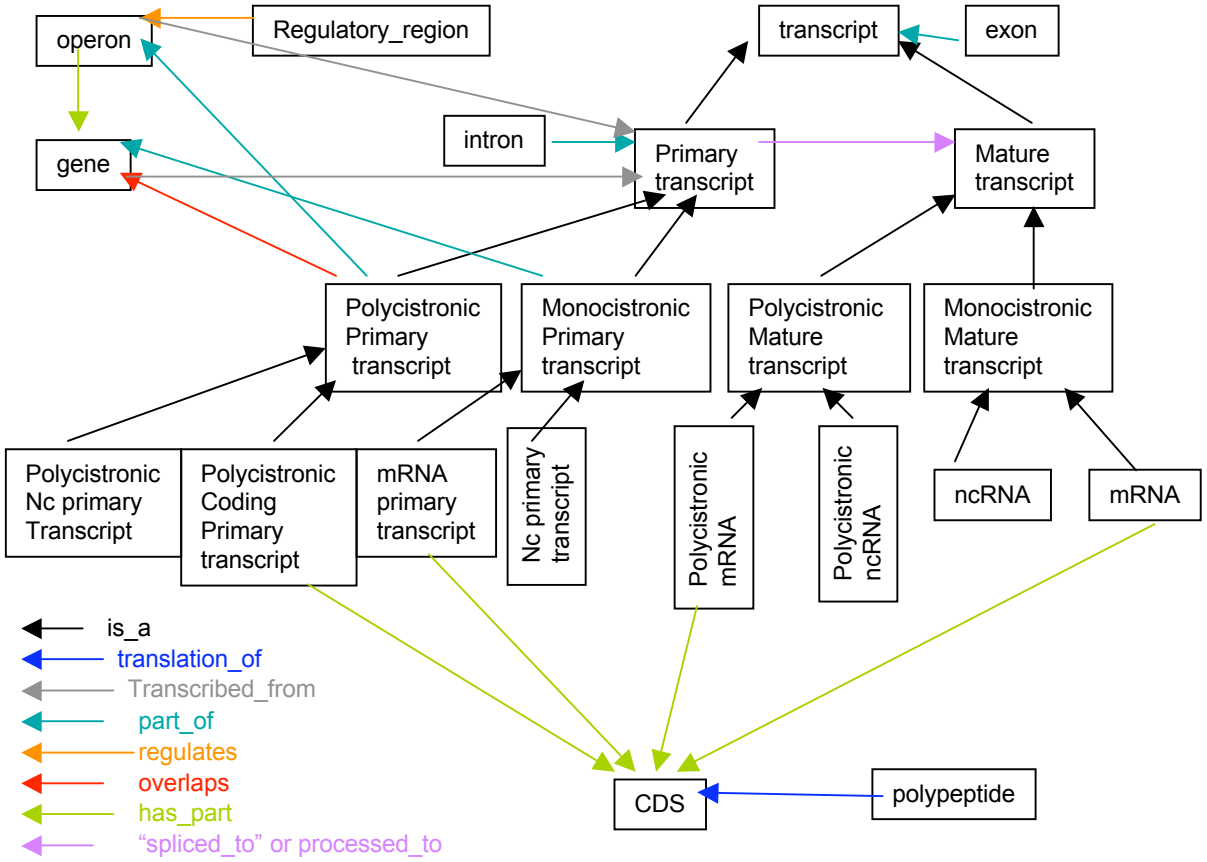


`is_a`: OK

`part_of`: probably not RO compatible in some cases (some–all instead of all–some)

`derives_from`: not well-defined

Relations: a second attempt



Relations: work in progress

- Relations
  - adjacent_to
  - disjoint from
  - encodes
  - evidence_for_feature
    - complete_evidence_for_feature
    - partial_evidence_for_feature
  - has_part
  - has_quality
  - homologous_to
    - non_functional_homolog_of
    - orthologous_to
    - paralogous_to
  - inverse of
  - is_a
  - overlaps
  - part_of
  - processed
  - regulates
  - similar_to
  - transcribed_from
  - translation_of
  - union of
  - variant_of

	transitive	symmetric	reflexive	antisymmetric
disjoint	-	+	-	-
adjacent_to	-	-	+	-
equal	+	-	+	+
inside	+	-	-	-
contains	+	-	-	-
covers	+	-	-	-
covered_by	+	-	-	-
overlaps	-	-	-	-

```
5. contains
|--A------|
  |-B--|
Feature A contains feature B if A and B share interior sequence b
of A's boundary coincides with B's interior.

6. covers
|--A--|
|--B-----|
Feature A covers feature B if both share a common boundary and in
```

See working_draft.obo on CVS:
http://song.cvs.sourceforge.net/song/ontology

Relations: homology

name: directly_descends_from

def: “F `directly_descends_from` F" iff there are O, O" such that O and O" are organisms, F and F" are features, O" is a parent of O and F has been copied from F".”

name: homologous_to

def: "A feature F is homologous to another feature F' if F `descends_from` F" and F' `descends_from` F"."

Alignment with RO

We are making sure all relations operate in the all–some direction.

This involves, for example, replacing `part_of` with `has_part` where appropriate.

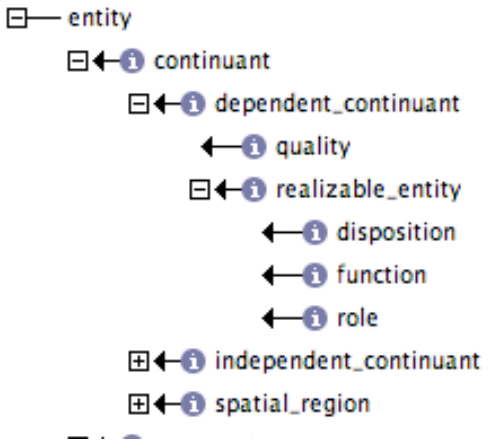
```
[Term]
id: SO:0001250
name: fingerprint_map
def: "A fingerprint_map is a physical map composed of
   restriction fragments." [SO:ke]
[…]
is_a: SO:0001249 ! fragment_assembly
relationship: has_part SO:0000412 ! restriction_fragment
```

Alignment with BFO (1)

Currently: `sequence_feature` terms in SO are related to `sequence_attribute` terms in SO by the `has_quality` relation.

This is not compatible with BFO!

Aim: to classify the `sequence_attribute` terms in SO according to the classes in BFO.



Alignment with BFO (2)

Iterative procedure:

- Two annotators
- Select 30 attributes randomly
- Assign each term into a category (quality, disposition, role, function) *without* consulting the other annotator
- Add justification
- Calculate agreement *adjusted for chance* (the kappa statistic)
- Discuss and modify guidelines
- Repeat

Alignment with BFO (3)

Outcomes (after three rounds):

- Tentative, sequence-specific definitions of BFO classes in working_draft.obo.
- *However*, we still have poor agreement:
  53% (κ = 0.42) on Q *vs.* D *vs.* R *vs.* F.
  77% (κ = 0.46) on Q *vs.* realizable entity
- More work on definitions needed!

Q = quality, D = disposition, R = role, F = function

Any questions?

The kappa statistic

$$\kappa = \frac{\text{actual agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

Widely-used in computational linguistics, psychology, clinical medicine and social sciences to adjust for chance agreement between human annotators. Always lower than or equal to percentage agreement.

κ = 1 indicates perfect agreement.

κ = 0 indicates agreement no better than if people were selecting categories at random.

κ = 0.67 is a lower bound for “acceptable” agreement.