Efforts toward a More Consistent and Interoperable Sequence Ontology

Mike Bada, Ph.D.
University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Karen Eilbeck, Ph.D.
University of Utah, Salt Lake City, UT, USA

The Sequence Ontology (SO)

- Member of the OBO library and candidate for the OBO Foundry
- Developed with goals of standardizing vocabulary and semantics of biological sequence annotation and increasing interoperability among software developers and users
- Represents types of biomacromolecular sequences, qualities & sequence variation

Use of the SO

- Has a sizable user base in the model-organism database community
- Genome Model Organism Database (GMOD) schemas, formats & tools rely on the SO
- Other uses of the SO have sprouted, including those in natural-language processing & reasoning with multiple ontologies

Motivation

- The SO is mature in ways, but issues remain:
  - Abstract vs. physical nature of sequences remain muddled
  - The SO is not well-integrated with other OBOs
  - Representation and use of corresponding DNA, RNA & peptide sequence is inconsistent

Current Efforts

- We are seeking to address these issues by:
  - Representing both abstract and physical sequences & linking them appropriately
  - Integrating (or setting up for integration) with ChEBI, PRO, RNAO, GO, CHEMINF & IAO
  - Consistently representing corresponding DNA, RNA & peptide sequences & harmonizing their use in annotations

The Nature of Sequences

- Hoehndorf *et al.* (*BMC* Bioinformatics, 2011) have posited 3 types of sequences:
  - *Abstract sequences*: 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"; only one instance of the abstract sequence ACA
  - *Syntactic sequences*: sequence representations as those in biomedical databases & text representations
  - *Molecular sequences*: Physical chains of nucleotides or amino acides

The Nature of Sequences

- In previous attempts to integrate with the BFO, SO developers have elaborated that SO sequences are *generically dependent continuants*, defined as continuants dependent on one or more independent-continuant bearers

(Mungall *et al.*, *J Biomed Inform*, 2011)

The Nature of Sequences

- However, this conceptualization is problematic:
  - SO developers have acknowledged a discordance in sequence attributes, *e.g.*, `wild_type_rescue_gene` is a `rescue_gene` that `has_quality wild_type`
  - More straightforwardly, biologists fundamentally regard sequences as molecular entities
  - This molecular view is reflected in the natural-language definitions in the SO

The Nature of Sequences

- Since molecular sequences are the more fundamental concepts, we argue that they should be explicitly represented
- However, there are at least a small number of SO classes whose conceptualizations as molecular entities do not seem sensible, *e.g.*, `match`
- Thus, some abstract sequences will be needed
- Our proposed solution is to represent sequences in two parallel ontologies

Parallel Sequence Ontologies

- One will be an evolution of the Sequence Ontology:Molecules (SOM) effort, with an enlarged domain of all SO molecular-sequence concepts
- This will be renamed the Molecular Sequence Ontology (MSO), since most SO concepts refer to parts of molecules
- The many formal definitions of SO concepts will be transferred to the MSO
- MSO concepts will be bridges to GO, PRO, RNAO & ChEBI

Parallel Sequence Ontologies

- The corresponding abstract sequences will remain in SO, which should minimize disruption to SO annotation efforts
- Corresponding abstract & molecular sequences will be identically named but use their respective namespaces
- SO concepts will be formally defined in terms of corresponding MSO concepts

Parallel Sequence Ontologies

- Since SO concepts will be formally defined in terms of MSO concepts, an OWL reasoner will be able to automatically generate the hierarchyof the former from the latter
- In addition to linking to the MSO, SO concepts will be connected to CHEMINF & thus indirectly to IAO

Integration with Other OBOs

- For the MSO, we seek integration with ChEBI, PRO, RNAO & GO
- MSO concepts will be subclasses of `CHEBI:molecular entity`
- A current high-level SO class is `region`, defined as a sequence feature with an extent greater than zero, which will be more precisely renamed to `monomeric sequence`

Integration of MSO with ChEBI

- "sequence" can refer to either a whole sequence or a proper subsequence, which will be encapsulated in the top-level `monomeric sequence`
- `monomeric sequence` will be fundamentally divided into `monomeric sequence molecule` and `monomeric subsequence`
- This subdivision will enable us to assert equivalency of specific existing ChEBI macromolecular classes & specific MSO subclasses of `monomeric molecule`

Integration of MSO with ChEBI

- We can further link MSO to ChEBI by defining MSO sequence types in terms of constituent ChEBI monomers, *e.g.*:

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

Integration of MSO with PRO & RNAO

- The PRO could link to the MSO by making its top-level `protein` a subclass of `MSO:peptide sequence molecule`
- Likewise, the RNAO will be able to integrate with the MSO by subclassing RNA-specific sequences and structures from more general MSO concepts

Integration of MSO & GO

- GO classes representing processes operating on sequences will be able to rely on relevant MSO classes, *e.g.*, for `RNA processing`, which is "[a]ny process involved in the conversion of one or more primary RNA transcripts into one or more mature RNA molecules":

```
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'
```

Integration of MSO & GO

- These can be seen as extensions to the OBO cross-product effort (Mungall *et al.*, *J Biomed Inform*, 2011)
- Currently, there are a plethora of vetted cross-product definitions among a number of OBOs as well as among concepts within the SO, but none among SO concepts and those of external OBOs

Integration of SO, CHEMINF & IAO

- SO concepts will be subsumed by `information about a chemical entity` from CHEMINF
- SO will thus be indirectly connected to IAO, as this CHEMINF class is itself a subclass of `IAO:information content entity`, which is “an entity that is generically dependent on some artifact and stands in relation of aboutness to some entity”

Integration of SO, CHEMINF & IAO

- We will use `denotes`, a subrelation of the IAO's fundamental `is_about` relation, to formally define most SO concepts in terms of the MSO, *e.g.*,

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

Integration of SO, CHEMINF & IAO

- Hypothetical, improbable & even impossible abtract sequences could be created, but we consider this an orthogonal issue, and there have been recent efforts to address this
- Since SO concepts will be formally defined in terms of MSO concepts, the classification of the former will be automatically generated from the latter

DNA, RNA & Peptide Sequences

- Sequences are annotated overwhelmingly at the genomic level, even with many SO concepts at RNA & peptide levels
- There are implicit semantics in that the given RNA- or peptide-level concept annotation holds for the RNA or peptide sequence *coded by* the annotated DNA sequence

DNA, RNA & Peptide Sequences

- RNA- and peptide-level SO classes are informally defined as RNA/peptide seqeunces, but they are sometimes subsumed by DNA concepts, *e.g.*, `transcript` is a `gene_member_region`
- We are seeking to make the SO more consistent in terms of both the ontology itself and its use in sequence annotations

DNA, RNA & Peptide Sequences

- Natural-language definitions of concepts should match formal structure, so either the RNA-level definition of `transcript` should change, or it should not be subsumed by a DNA-level concept
- We argue that classes should be defined as they are canonically conceptualized, so `transcript` should be defined at the RNA level
- Its formal classification should reflect this

DNA, RNA & Peptide Sequences

- For this classification, we have created a set of sequence classes defined in terms of type of monomer
- Currently, monomer type is represented by a set of polymer attributes & sequences are attributed these qualities, *e.g.*, `DNA`, `RNA` and `peptidyl` are all qualities, and `RNA chromosome` is formally defined as:

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

DNA, RNA & Peptide Sequences

- For each type of monomer, we are creating a primary sequence class, *e.g.*, `DNA sequence`, `RNA sequence`, `peptide sequence`
- Rather than relying on qualities for specifying monomer type, we will use existing ChEBI monomer classes
- As many monomer types are already represented in ChEBI, this reduces effort on our end & abides by the principle of OBO orthogonality

DNA, RNA & Peptide Sequences

- Monomeric sequences are thus subdivided into two orthogonal axes:
  - Whole molecules vs. proper subsequences
  - Monomer types
- However, all these direct subclasses will be necessarily and sufficiently defined, enabling automatic classification

DNA, RNA & Peptide Sequences

- To address issue of discordance between represented concepts & annotated sequences, we propose:
  - Creating corresponding DNA, RNA & peptide sequence classes
  - Defining/linking them accordingly
  - Guiding annotators to their proper use
- To minimize confusion, name in parallel, *e.g.*, for `polypeptide domain`, also create `DNA coding for polypeptide domain` & `RNA coding for polypeptide domain`

DNA, RNA & Peptide Sequences

- To link these, one option 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
     some 'RNA coding for polypeptide domain'
```

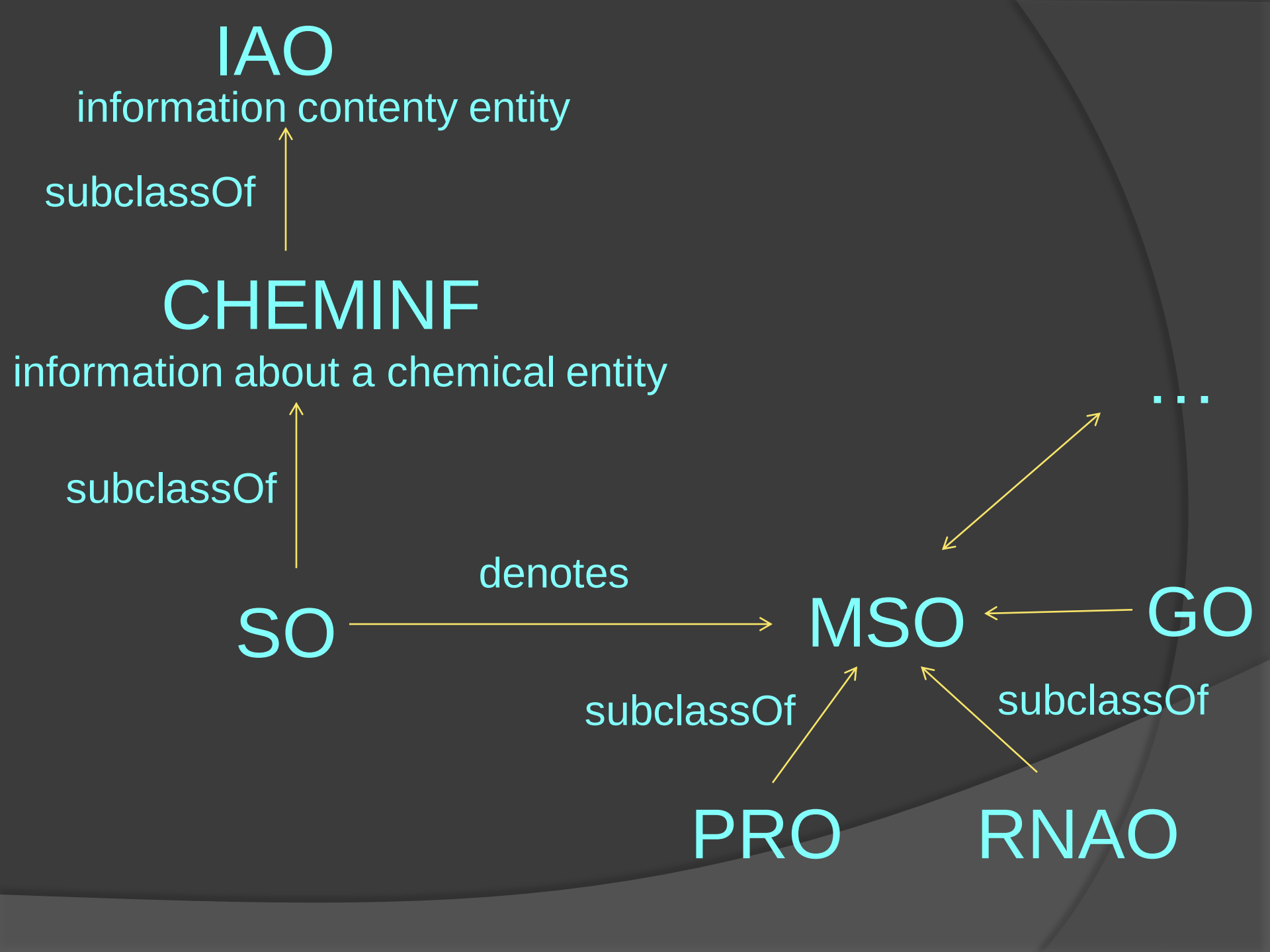
- However, this seems odd and circular

DNA, RNA & Peptide Sequences

- The other option is to state each association in the reverse direction, *e.g.*,

```
'RNA coding for polypeptide domain'
   subclassOf
      'RNA sequence' and
       template_for only 'polypeptide_domain'
```

- This seems more sensible in that it reflects the class
- A universal (only) rather than an existential (some) restriction would be needed


IAO
information contenty entity
subclassOf
CHEMINF
information about a chemical entity
subclassOf
SO
denotes
MSO
GO
...
subclassOf
subclassOf
PRO
RNAO

Conclusions

- Our recent efforts in the continuing development of the SO:
  - Representation of molecular vs. abstract sequences
  - Integration of the SO with ChEBI, PRO, RNAO, GO, CHEMINF & IAO
  - Consistent representation & use of corresponding DNA, RNA & peptide sequences
- In addition to increasing interoperability of SO with other OBOs, we anticipate that this will improve consistency both internally and with respect to external resources

Thanks!