Journal of Biomedical Informatics 44 (2011) 87–93

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evolution of the Sequence Ontology terms and relationships

Christopher J. Mungall [a], Colin Batchelor [b], Karen Eilbeck [c,*]

[a] Lawrence Berkeley National Laboratory, Mail Stop 64R0121, Berkeley, CA 94720, USA
[b] Royal Society of Chemistry, Thomas Graham House, Cambridge CB4 0WF, UK
[c] Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA

ARTICLE INFO

Article history:
Received 23 October 2009
Available online 10 March 2010

Keywords:
Sequence Ontology
Biomedical ontology
Genome annotation

ABSTRACT

The Sequence Ontology is an established ontology, with a large user community, for the purpose of genomic annotation. We are reforming the ontology to provide better terms and relationships to describe the features of biological sequence, for both genomic and derived sequence. The SO is working within the guidelines of the OBO Foundry to provide interoperability between SO and the other related OBO ontologies. Here, we report changes and improvements made to SO including new relationships to better define the mereological, spatial and temporal aspects of biological sequence.

1. Introduction

Genomic data was notorious for the multitude of file formats that expressed the same kind of data in different ways. Each gene prediction algorithm for example, exported the gene models in either a different format from other groups, or when they used the same format, the terms often had slightly different meanings. Data integration between groups was therefore not straightforward. Likewise, validation of annotations relied on the programmers understanding the nuances of each kind of annotation and hard-coding their programs to match. The Sequence Ontology (SO) [1] was initiated in 2003 to provide the terms, and relations that obtain between terms, to describe biological sequences. The main purpose was to unify the vocabulary used in genomic annotations, specifically genomic databases and flat file data exchange formats. The Sequence Ontology Project provides a forum for the genomic annotation community to discuss and agree on terminology to describe the biological sequence they manage, in the form of mailing lists, trackers and workshops.

The purpose of annotating a genome is to find and record the parts of the genome that are biologically significant. In this way researchers can make sense of what would just be a very long string of letters. For example, after annotation, a researcher will be able to know which of the sequence variants fall in coding or non coding sequence and perform subsequent analyses accordingly. A genome annotation anchors knowledge about the genomic sequence and the sequence of molecules derived from the genome onto a linear representation of the replicon (chromosome, plasmid etc.) using base pair coordinates to capture the position. A **sequence_feature** is a region or a boundary of sequence that can be located in coordinates on biological sequence, and SO was initially created as an ontology of these sequence feature types and their attributes.

The SO has a large user community of established model organism databases and newer 'emerging model organism' systems who use on the Generic Model Organism Database (GMOD) [2] suite of tools to annotate and disseminate their genetic information. GMOD is a group that provides an open source collection of tools for dealing with genomic data. GMOD schemas and exchange formats rely on the SO to type their features such as the Chado database [3] with its related XML formats and the tab delimited flat file exchange format Generic Feature Format (GFF3) [4]. Several GMOD tools use GFF3, for example GBrowse [5]. SO is also used by genome integration projects such as Flymine [6], modENCODE [7] and the BRC pathogen data repository [8,9]. There are other uses for SO such as natural language processing initiatives that use the SO terminology [10,11].

Genome annotations specify the coordinates of sequence features that are manifest in one or more of the kinds of molecule defined by the central dogma. For example, although an intron is manifest as an RNA molecule, the coordinates of the intron can be projected onto the genomic sequence. The term labels chosen for SO were those in use by the genome annotation community, thus "transcript", "intron" and so on were chosen as labels for the sequence feature types corresponding to genome regions encoding actual transcript and intron molecules. This polysemy does not cause problems when SO is used purely for genome annotation, but is potentially confusing when it is used in the context of other ontologies.

The current version of SO uses a subsumption hierarchy to describe the kinds of features and a meronomy to describe their

* Corresponding author. Address: Department of Human Genetics, Building 533, 15N 2030 East, Salt Lake City, UT 84108, USA. Fax: +1 801 581 7796.
E-mail address: keilbeck@genetics.utah.edu (K. Eilbeck).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.03.002

part-whole structures. Sequence features were related by their genomic position. For example **polypeptide** (which referred to the sequence that corresponds to a polypeptide molecule) and **transcript** (which referred to the sequence that corresponds to an RNA molecule) were described only by genomic context, that is the region of the genome that encodes their sequence. This excluded the post-genomic topology of these features: how the topology of the features changes, as the sequence is expressed by different molecules.

The SO is one of the original members of the OBO Library, a collection of orthogonal, interoperable ontologies developed according to a shared set of principles. These later evolved into the OBO Foundry principles [12] which include a common syntax, a data-versioning system, collaborative development, and adherence to the same set of defined relationships [13]. The OBO Foundry ontology developers attempt to accurately represent biological reality. Membership in the OBO Foundry represents a commitment to adhere to common ontology design principles and agree to reform where necessary. The OBO Foundry spans the biomedical domain in steps of granularity from the molecule to the organism, and also extends into the realm of experimental measurements, instrumentation and protocol. The OBO Foundry also partitions ontologies according to their relationship to time. Continuants endure through time, whereas occurrents, which include processes, unfold through time in stages. Anatomical entities such as cells and organs are continuants, as are molecules.

The SO is orthogonal to the neighbor ontologies within the OBO Foundry which represent molecular continuants. Chemical Entities of Biological Interest (ChEBI) is a dictionary of small molecules [14]. The RNA Ontology [15] represents the secondary and tertiary motifs of RNA as well as describing the interactions between bases for base pairing and stacking. The Protein Ontology (PRO) defines the forms of proteins and the evolutionary relationships between protein families [16]. These ontologies are themselves orthogonal to ontologies of processes, such as the Biological Process (BP) and Molecular Function (MF) subsets of the Gene Ontology (GO) [17]. The GO BP ontology represents processes of relevance to SO, such as transcription, gene expression and splicing.

In order to best divide work between curators of neighboring ontologies, and to ensure that SO can reuse material from these ontologies and *vice versa*, the ontologies must all adhere to the same principles. In this paper, we will describe how we have been developing the Sequence Ontology in two respects, first to promote interoperability and second to provide a solid framework to describe how sequences change over the course of genomic and post-genomic processes. The rest of the paper is structured as follows: in Section 2 we describe the OBO Foundry standards we have been adopting. In Section 3 we describe new relations for post-genome topology and in Section 4 we describe the relation of SO to neighboring ontologies.

2. Coordinated reform of SO to OBO standards

The SO, like other pre-existing ontologies has begun to undergo reform to meet the OBO Foundry standards.

2.1. Conformation to an upper ontology

Upper ontologies such as Basic Formal Ontology (BFO) [18] provide a formal structure upon which to base domain ontologies. BFO provides a hierarchy of upper-level abstract classes. Classes in domain-specific ontologies can be defined as sub-classes of appropriate abstract classes and inherit their properties. This allows the multiple independently developed ontologies of the OBO Foundry to be linked together. The development of SO preceded the adoption of BFO by the OBO Foundry, so it was necessary to align SO to BFO *post hoc*. In order to do this, a fundamental question must be answered: what kind of entity is a sequence feature? This is not a trivial question and suggested answers have ranged from: molecules or molecule regions, the physical pattern of electrons in a computer or purely abstract mathematical forms. None of these solutions was biologically satisfying. 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. This can be seen as analogous to the distinction between the physical content of a book, and the words written in that book.

BFO divides continuants into **independent continuants** and **dependent continuants**. The former include physical objects such as molecules, and the latter include entities such physical qualities, shapes and functions. The relation that links these is called *inheres_in*, and we say that for example my temperature *inheres_in* me, or that I am the bearer of my temperature. Dependent continuants are broken down into **specifically dependent continuants** (SDCs) and **generically dependent continuants** (GDCs). What differentiates these is the number of bearers – a SDC has a single bearer, and ceases to exist when that bearer ceases to exist (thus the shape of a particular apple disappears after the apple is eaten). A GDC can have multiple bearers, and can continue to exist when bearers cease to exist, so long as there is at least one bearer. A given genomic sequence may be borne by a DNA molecule, an RNA molecule, a polypeptide chain, or indeed by other molecules or systems that are not products of the replication machinery of the cell, for example the set of instructions that drive a solid-phase nucleic acid synthesis device. For this reason we take biological sequences to be GDCs (Fig. 1). One of the consequences of this decision is that genes such as the gene denoted by the NCBI Gene ID 6469 (human Shh) are *individuals* rather than *types*.

The other SO root classes have also been aligned to BFO, as shown in Fig. 1. We take **sequence_collection**, which is a non-contiguous set of sequences, and **sequence_variant**, such as a mutation, to be the same sort of thing as a sequence_feature, and hence a GDC. For the moment we are treating **sequence_attribute** as an intrinsic property of the molecule that bears the sequence, hence in BFO terms a quality, but this is under review. Lastly, the **sequence_variant_effect**, for example a structural change or a change in transcription, need not necessarily happen so we treat it as a disposition.

2.2. Definitions

We now define new terms according to the OBO Foundry guidelines for definitions. Initially the terms in SO were either defined by a member of the developer community, or taken directly from a reputable website or textbook, giving the ISBN or the URL as the cross-reference. This has led to inconsistency between the definitions, and sometimes inconsistency between the definition and placement of the term within the ontology. This especially led to confusion over the kind of entity described by a feature, whether it was a molecule or a sequence, as there was not conformity in the definitions. For example, **mRNA** was defined as: *Messenger RNA is the intermediate molecule between DNA and protein. It includes UTR and coding sequences. It does not contain introns.* This has been updated to '*Messenger RNA sequence is a mature transcript sequence, a portion of which is coding. It may include UTR but not intron sequence*'. The OBO Foundry recommends that terms be defined with respect to the *is_a* parent, and the attributes that differentiate the term from its parent and sibling terms, called the differentiae. This practice forces a self check on the whether the position of the term in the ontology agrees with the defined