# A BioSequence Ontology from Molecular Structure

Jona THAI [a], Michael GRÜNINGER [a]
[a] *Department of Mechanical and Industrial Engineering, University of Toronto, Ontario, Canada M5S 3G8*

**Abstract.** Gene sequences are a focal point of modern biological research, with applications ranging from diagnostics to gene-driven drug design. An ontology's automated reasoning capability and traceable logic is sure to be an asset to these efforts. However, current biomedical ontologies fail to achieve this potential. This may partly be due to a lack of formal axiomatization of the underlying molecular structure and mereology of gene sequences, despite an otherwise richly defined vocabulary. In this paper, we propose a new BioSequence Ontology with explicit axiomatization of the underlying path graph structure.

**Keywords.** mereology, sequence ontology, genes, molecules

## 1. Introduction

If we passed an arbitrary gene sequence as a query into a biomedical ontology, would it be able to reason about its ancestry and recognize an open reading frame? Removing the possibility of recognizing a well-known or pre-programmed sequence, most biomedical ontologies simply do not have the axiomatization that is necessary and sufficient to reason about concepts like parthood, betweenness or even directionality. Moreover, gene sequences are curious things. They are simultaneously a molecule, abstract information and (to a computer at least) a string of letters. Even if we chose to focus on the molecular aspect of a gene sequence, it can be either linear or circular, which have not-so-subtle nuances in understanding of betweenness. Hence, the BioSequence Ontology proposed in this paper has three goals: (1) Explicit axiomatization of the mereotopology of gene sequences (2) Design a gene sequence ontology based on molecular structure (3) Provide representation for both linear and circular gene sequences.

## 2. Motivating Scenarios

What is a gene? It is commonly agreed to be a sequence of nucleotides in DNA or RNA that encode the synthesis of a gene product, including and not limited to RNA or protein. To break down this definition, we have to look into definitions of key terms. For instance, consider the word "sequence". Formally, a sequence is a collection of elements, or members in which repetitions are allowed and order matters. Hence, a sequence can be built of sub-sequences, which are further built of sub-sub-sequences and so on so forth. This

is important because it sheds light as to what kind of molecules are structurally plausible and what are not.

*Motivating Scenario 1: Sequence of Sequences*

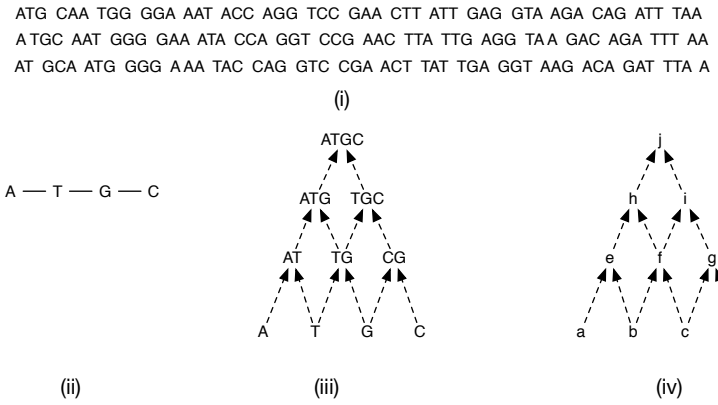


**Figure 1.** Mereology of gene sequences.

A DNA molecule is understood to have a double helix, double strand structure. However, these strands are complementary and essentially hold the same information, so it is of interest to focus on the structure of a single strand. RNA also happens to be single-stranded. Figure (i) shows a an arbitrary sequence split into triplets (codons). Figure (ii) then zooms into the connection between 4 nucleotides. As displayed by Figure (iii), this set of nucleotides can be split into two codons, depending on how the sequence is read, with ”TG” as the overlapping pair of elements. However, ”A” and ”C” are never considered as a pair of elements as ”TG” is in between them. In a classical mereology, ”AC” would be considered as a possible pair. This clearly shows that gene sequence mereology differs subtly from classical mereology due to this emphasis on a connected substructure. Another word of interest is ”encode”. In this context, encoding is achieved through a series of processes, namely transcription, translation, and splicing to build a product, namely RNA or protein. These processes are explored more in depth below.

*Motivating Scenario 2: Splicing*

Splicing is a process that converts messenger RNA (mRNA) to mature RNA, through the removal of introns(non-coding regions) and joining together exons(coding region). It is a fundamental process in making proteins that takes place after transcription and before translation.

*Motivating Scenario 3: Transcription*

In a nutshell, transcription is the process of copying DNA to RNA. This is achieved with RNA polymerase (RNA maker, essentially) binding to a section on the DNA strand called the promoter. Since DNA has a double helix structure (2 strands), the RNA is copied by generating a complementary strand of the template DNA strand.

*Motivating Scenario 4: Translation*

Post-splicing, translation is the process of building an amino acid chain from the transcribed RNA. This is achieved via tRNA attaching to codons after the start codon, and assigning the according amino acid.

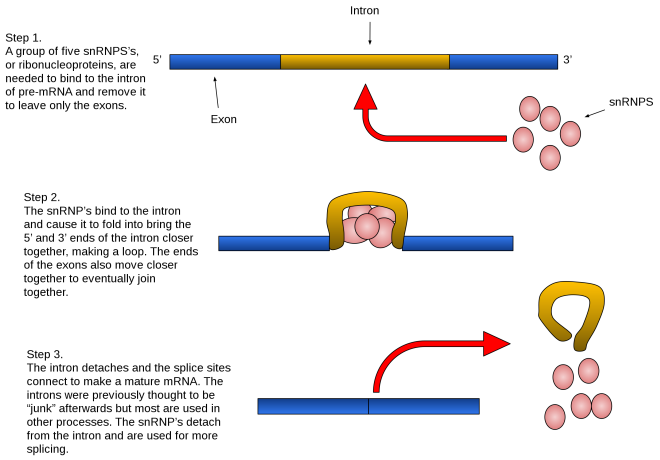


**Figure 2.** Splicing
[12]

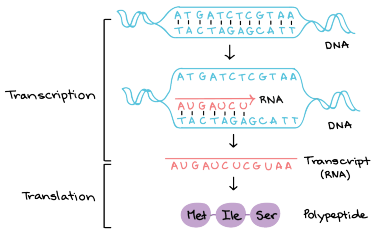


**Figure 3.** Transcription & Translation
[5]

Coincidentally, processing the gene sequence as above results in an amino acid chain or RNA, which is also a sequence. In other words, these processes can essentially be interpreted as sequence manipulation to build other sequences.

### *2.1. Semantic Requirements/Intended Models*

Based on the above motivating scenarios, we can conclude that genes can be interpreted and operated on as sequences. However, in reality, genes are also physical molecules with chemical structures that are not necessarily sequence-like. In fact, spatial aspects of a molecule is akin to a graph i.e. chemical graph theory.
This begs the question - How can we preserve the intuitive semantics of a sequence while remaining true to its molecular properties, to maximize inference capabilities and minimizing complexity? This motivates the following competency questions.

1. *What type of overlap is present with ATP6 and ATP8 genes?*
   An overlapping gene exists when an expressible nucleotide sequence of a gene overlaps with another expressible nucleotide sequence in another gene. This is

natural so that a sequence can contribute to more than one gene product. This competency question tests the ontology's ability to reason about parthood and directionality within a gene sequence.

2. *What do the DNA codons ATT and TTA code for?*
   The RNA sequence ATT codes for Isoleucine whereas TTA codes for Leucine. This competency question is designed to ensure the ontology's ability reason about directionality and sequence convexity (between-ness).
3. *What are the introns/exons in the given sequence?*
   Introns are non-coding sequences of DNA that are removed via splicing. Exons are coding sequences of DNA that provide instructions for making a product (protein). This coding sequence tests the ontology's ability to reason and classify key components of gene structure in regards to function/process.
4. *What are conserved sequences present in the histone h1 protein given the amino acid sequence.*
   Histone H1 is one of the five main histone protein families which are components of chromatin in eukaryotic cells. It is simultaenously one of the most conserved and most variable histone across species. A conserved sequence are identical or similar sequences in nucleic acids (DNA and RNA) or proteins across species.
5. *What are conservative replacements for the conserved sequence in the gene coding for the h1 histone protein?*
   A conservative replacement in an amino acid chain involves swapping an amino acid with another that has similar biochemical properties e.g. glycine and alanine. This competency question tests the ontology's ability to reason about molecular structure.
6. *Is this gene sequence circular?* A circular gene sequence is essentially a sequence where the 5' and 3' ends have been joined together to form a loop. Examples of natural occurring circular sequences are eccDNA and plasmids. This competency question tests the ontology's ability to reason about between-ness and correlate molecular structure with gene sequence.
7. *How is 5'UTR and the start codon related?*
   The 5' UTR (untranslated region) refers to a phosphate group attached to the 5 carbon of the ribose ring in a DNA/RNA molecule, and the start of the open reading frame. The start codon is first codon of a messenger RNA (mRNA) transcript translated by a ribosome during translation. The 5' UTR is directly upstream of the start codon.

Based on these competency questions, we can construct ontological commitments to drive the design of the ontology.

1. The ontology must represent the properties of functional groups, connections between functional groups, connections between functional groups with sequences, components of sequences and connections between sequences and sequences.
2. Functional groups, molecules, and chains of molecules(sequences) are the primitives of the ontology. Functional groups are the lowest level of abstraction in this ontology.
3. Axioms in the ontology are fully interpretable in the Molecular Structure Ontology (MoST) [3] as all sequences are molecules.

4. The ontology must represent the 5' and 3' relation: the 5' end of a sequence is the beginning of the open reading frame and the 3' end is the end of the open reading frame. This introduces the concept of directionality into the ontology.
5. The ontology must represent the "between" relation. To understand what a gene sequence does, it requires understanding the concept of an open reading frame (between start and stop codons). Sequences often have overlapping reading frames so that a nucleotide sequence can code for more than one protein.

## 3. Existing Work

### *3.1. Evaluation wrt semantic requirements*

The idea to represent gene sequences through an ontology is not a novel one. In fact, two of the most notable biomedical ontologies are dedicated to this purpose - the Gene Ontology (GO) [1] and the Sequence Ontology(SO) [4] and Molecular Sequence Ontology(MSO) [2].
The GO was a pioneer in creating an extensive, consistent vocabulary on gene sequences and is widely used, particularly in annotation. Definitions within the ontology are defined in terms of each other, akin to a directed acyclic graph in mathematics. This ensures consistency with labeling.
However, the GO had no explicit mathematical definitions to represent part-hood in a gene sequence. This was addressed in the SO, which recognized the need for a properly defined mereology for increased automated reasoning capacity. This topic was first raised in a paper by Hoehndorf [10]. Here, he explicitly stated the difference between molecular, abstract and syntactic aspects of a gene sequence and provided a set of axioms. However, this axiomatization was disputed by the original creators of the Sequence Ontology in a later-released paper [9]. Mungall then released another paper [8], which recognized the deep similarity between time intervals and gene sequences, as both have an underlying connected substructure. However, there were no explicit axiomatizations of the mereology and ordering.
Developments in the Molecular Sequence Ontology is fairly recent, and meant to be the molecular counterpart to the Sequence Ontology's genome annotation. [2]

### *3.2. Relationship to Upper Ontologies*

One may wonder why the MSO cannot simply be added on as a feature to the SO. This is perhaps largely due to SO and MSO's alignment to the Basic Formal Ontology (BFO). In BFO, concepts are classified as continuant or occurrent. This works great in many cases, but in this case, it would be rather limiting to classify a gene sequence as only either abstract information(sequence) or physical entity(molecule).
As highlighted earlier, we are committed to accurately represent gene sequences as what they are - sequences and molecules. To achieve that, we need to build upon an ontology of molecular chemistry, with rigorous definitions of molecular structure and parthood. A candidate we considered was Chemical Entities of Biological Interest(CheBI) [6], however it did not quite fulfill our requirements for a mathematically rigorous definition of parthood relations. After consideration, only the Molecular Structure Ontology

(MoST) [3] best fit our requirements for expressivity and axioms about structural properties. Its ability to represent functional groups in chemical graph theory grants us the luxury of setting that as the BioSequence Ontology's lowest level of abstraction. This will improve the readability and usability of the BioSequence Ontology. Coincidentally, MoST is independent from an upper ontology. As such, the BioSequence Ontology is also foundationless, the effect of which will be studied.

## 4. The BioSequence Ontology

### *4.1. Overview*

In this section we will discuss the signature and design of the BioSequence Ontology based on the requirements developed in previous sections. This ontology is axiomatized in Common Logic syntax and made available in a repository on COLORE link.

### *4.2. MoST*

What is DNA? It is a sequence of nucleotides, and hence it is also a molecule. To accurately represent a molecule, we need to build upon a knowledge base of atoms and bonds. To achieve this, we build upon the existing Molecular Structure Ontology (MoST). It is a foundation-less ontology that represents molecules as molecular graphs.
Since the focus of the BioSequence Ontology is the semantics of gene sequences, the lowest level of abstraction would be on a *functional_group* level. Atoms and bond types would be beyond our scope, and this information is inherited from MoST. Hence, most or all entities within the BioSequence Ontology can be represented as a skeleton, or connected graph of atoms. This is achieved with the *mol* and $tether$ relations.

**Definition 1**

$$\forall x(mol(x,x)) \tag{1}$$

$$\forall x \forall y(mol(x,y) \supset mol(y,x)) \tag{2}$$

$$\begin{aligned} &\forall g_1 \forall g_2 \forall b((tether(g_1,g_2,b) \equiv (group(g1) \wedge group(g2) \\ &\wedge bond(b) \wedge (g_1 \neq g_2) \wedge \exists a_1 \exists a_2((atom(a_1) \wedge atom(a_2) \wedge mol(a_1,g_1) \\ &\wedge mol(a_2,g_2) \wedge mol(a_1,b) \wedge mol(a_2,b) \wedge \neg mol(b,g_1) \wedge \neg mol(b,g_2)))))). \end{aligned} \tag{3}$$

### *4.3. Signature*

Molecular chemistry background knowledge in hand, the natural next step in building the ontology is to define the primitives, or basic building blocks of the ontology. Every entity in the ontology will be defined in terms of these primitives. Intuitively, these primitives will be $skeleton$, $sequence$, $nucleotide$, $functional\_group$, $triplet$, $coding\_sequence$, $protein$ and $codon$.

**Definition 2**

$$\forall x nucleotide(x) \supset skeleton(x) \tag{4}$$

$$\forall x sequence(x) \equiv skeleton(x) \tag{5}$$

$$\forall x nucleotide(x) \equiv \exists s1 \exists s2 \exists s3 \exists b1 \exists b2 skeleton(x)$$
$$\wedge nucleobase(s1) \wedge sugar(s2) \wedge phosphoric\_acid(s3) \wedge mol(s1,x) \wedge mol(s2,x) \wedge mol(s3,x)$$
$$\wedge tether(s1,s2,b1) \wedge tether(s2,s3,b2) \tag{6}$$

$$\forall x\ sequence(x) \equiv \exists n1 \exists n2 \exists n3 \exists b1 \exists b2 \wedge nucleotide(n1) \wedge nucleotide(n2)$$
$$\wedge nucleotide(n3) \wedge tether(n1,n2,b1) \wedge tether(n2,n3,b2) \tag{7}$$

$$\forall x\ triplet(x) \equiv \exists n1 \exists n2 \exists n3 \exists b1 \exists b2\ nucleotide(n1) \wedge nucleotide(n2) \wedge nucleotide(n3)$$
$$\wedge tether(n1,n2,b1) \wedge tether(n2,n3,b2) \tag{8}$$

$$\forall ccodon(c) \equiv (\exists x,y,z) nucleotide(x) \wedge nucleotide(y) \wedge nucleotide(z) \wedge x \neq y \wedge x \neq z \wedge y \neq z$$
$$\wedge part(x,c) \wedge part(y,c) \wedge part(z,c) \wedge \exists y1 \exists z1\ amino\_acid(y1) \wedge rf(z1) \wedge triplet(c) \wedge mol(c,z1)$$
$$\wedge codes\_for(c,y1) \wedge (\forall w)\ nucleotide(w) \wedge part(w,c) \supset ((w = x) \vee (w = y) \vee (w = z)))) \tag{9}$$

$$\forall x\ cds(x) \equiv \exists y\ protein(y) \wedge sequence(x) \wedge translates\_to(x,y) \tag{10}$$

$$\forall x\ rf(x) \equiv \exists y \exists z\ nucleotide(y) nucleotide(z) \wedge sequence(x) \wedge 5'(y,x) \wedge 3'(z,x) \tag{11}$$

$$\forall x\ orf(x) \equiv \exists y \exists z\ rf(x) \wedge protein(y) \wedge RNA(z) \wedge (transcribes\_to(x,z) \vee translates\_to(z,y)) \tag{12}$$

Next, relationships between the entities need to be defined. These relationship axiom predicates are *between*, *overlap*, *tether*.

**Definition 3**

$$\forall x \forall y\ tandem\_overlap(x,y) \equiv sequence(x) \wedge sequence(y) \wedge x \neq y$$
$$\wedge \exists n1 \exists n2\ nucleotide(n1) \wedge nucleotide(n2)$$
$$\wedge 3'(n1,x) \wedge 5'(n2,y) \wedge overlap(n1,n2) \tag{13}$$

$$\forall x \forall y\ convergent\_overlap(x,y) \equiv sequence(x) \wedge sequence(y) \wedge x \neq y$$
$$\wedge \exists n1 \exists n2\ nucleotide(n1) \wedge nucleotide(n2)$$
$$\wedge 3'(n1,x) \wedge 3'(n2,y) \wedge overlap(n1,n2) \tag{14}$$

$$\forall x \forall y\ divergent\_overlap(x,y) \equiv sequence(x) \wedge sequence(y) \wedge x \neq y$$
$$\wedge \exists n1 \exists n2\ nucleotide(n1) \wedge nucleotide(n2)$$
$$\wedge 5'(n1,x) \wedge 5'(n2,y) \wedge overlap(n1,n2) \tag{15}$$

$$\forall x \forall y\ in\_phase(x,y) \equiv sequence(x) \wedge sequence(y) \wedge (rf(x) = rf(y)) \tag{16}$$

$$\forall x \forall y\ out\_phase(x,y) \equiv sequence(x) \wedge sequence(y) \wedge (rf(x) \neq rf(y)) \tag{17}$$

### *4.4. Mereology*

A major goal of the BioSequence Ontology is to explicitly axiomatize the parthood relations between entities of the ontology. The most unique point of the mereology of gene sequences is that it is composed of convex intervals; in other words, the mereological

sum of sequence A and sequence B implies that A and B are already connected through a bond. We achieve this by modeling gene sequences after path graphs, defined below [7].

**Definition 4** *Let* $\mathbb{H} = \langle V, \mathbf{E} \rangle$ *be a simple graph.*

$\mathbb{H}$ *is a path iff there exists a sequence* $\mathbf{x_1}, ..., \mathbf{x_n}$ *such that* $(\mathbf{x_i}, \mathbf{x_{i+1}}) \in \mathbf{E}$.

$\mathbb{H}$ *is connected iff for any two vertices* $\mathbf{x}, \mathbf{y} \in V$*, there exists an induced subgraph that is a path containing* $\mathbf{x}, \mathbf{y}$.

In a nutshell, the following definitions illustrate the chain-like connected-substructure of a sequence.

**Definition 5** *A partial ordering* $\mathbb{P} = \langle V, \leq \rangle$ *is properly chain semimodular iff*[1]

1. $\mathbb{P}$ *is atom-height, that is, the cardinality of all maximal chains in* $\mathbb{P}$ *is equal to the cardinality of the set of atoms in* $\mathbb{P}$*;*
2. *for each* $\mathbf{x} \in V$*,* $\langle U^{\mathbb{P}}[\mathbf{x}], \leq \rangle$ *is an upper semimodular lattice:*
   (a) *any two elements* $\mathbf{y}, \mathbf{z}$ *have a least upper bound and a greatest lower bound in* $U^{\mathbb{P}}[\mathbf{x}]$*;*
   (b) *if* $\mathbf{z}$ *covers the greatest lower bound of* $\mathbf{z}$ *and* $\mathbf{y}$*, then the least upper bound of* $\mathbf{z}$ *and* $\mathbf{y}$ *covers* $\mathbf{y}$*;*
3. *for each* $\mathbf{x} \in V$*,* $\langle U^{\mathbb{P}}[\mathbf{x}], \leq \rangle \cong \mathbf{m} \times \mathbf{n}$.

$\mathfrak{M}^{proper\_chain\_semimodular}$ *denotes the class of properly chain semimodular partial orderings.*

The following theorem illustrates the verification of the mereotopology by establishing a bijection between our definition of part-hood and the models of our ontology for gene sequences. We obtain the axioms of our ontology from $T_{cisco}$, an ontology for mereotopology of connected substructures [7].

**Theorem 1** *Let* $T_{cico}$ *be the extension of* $T_{em\_mereology} \cup T_{ub\_mereology}$ *with the sentences:*

[1] We use the following notation:
$U^{\mathbb{P}}[\mathbf{x}] = \{\mathbf{y} : \mathbf{x} \leq \mathbf{y}\} \qquad U^{\mathbb{P}}[X] = \bigcup_{\mathbf{x} \in X} U[\mathbf{x}]$

$$(\forall u,x)\, ppart(u,x) \supset (\exists y)\, atom(y) \wedge part(y,x) \tag{18}$$

$$(\forall x,y)\, covers(x,y) \supset (\exists z)\, atom(z) \wedge ppart(z,x) \wedge \neg part(z,y) \tag{19}$$

$$(\forall x,y,z,u)\, covers(x,y) \wedge atom(z) \wedge ppart(z,x) \wedge \neg part(z,y)$$
$$\wedge atom(u) \wedge ppart(u,x) \wedge \neg part(u,y) \supset (z = u) \tag{20}$$

$$(\forall x,a,b)\, part(x,a) \wedge part(x,b) \supset$$
$$(\exists z)\, part(x,z) \wedge (\forall u)\, (part(z,u) \equiv (part(a,u) \wedge part(b,u))) \tag{21}$$

$$(\forall x,a,b)\, part(x,a) \wedge part(x,b) \supset$$
$$(\exists z)\, part(x,z) \wedge (\forall u)\, (part(u,z) \equiv (part(u,a) \wedge part(u,b))) \tag{22}$$

$$(\forall p,x,y)\, atom(p) \wedge part(x,y) \wedge \neg part(p,y) \supset$$
$$(\exists z)\, part(x,z) \wedge part(p,z) \wedge part(y,z) \wedge covers(z,y) \tag{23}$$

$$(\forall u,x)\, ppart(u,x) \supset (\exists y,z)\, covers(x,y) \wedge covers(x,z) \wedge y \neq z \tag{24}$$

$$(\forall u,x,y,z)\, ppart(u,x) \wedge covers(x,y) \wedge covers(x,z) \wedge covers(x,w)$$
$$\supset (y = z \vee y = w \vee z = w) \tag{25}$$

$$(\forall x,y,z,w)\, covers(y,x) \wedge covers(z,x) \wedge covers(w,x) \supset (y = z \vee y = w \vee z = w) \tag{26}$$

$$(\forall x,y,z,w)\, covers(y,x) \wedge covers(z,x) \wedge y \neq z \wedge overlaps(w,x)$$
$$\supset (\exists u)\, atom(u) \wedge \neg part(u,w) \wedge \neg part(u,x) \tag{27}$$

*There exists a bijection* $\varphi : Mod(T_{cico}) \to \mathfrak{M}^{proper\_semimodular}$ *such that*
$(\mathbf{x},\mathbf{y}) \in \mathbf{part}^{\mathscr{M}}$ *iff* $\mathbf{x} \in L^{\mathbb{P}}[\mathbf{y}]$

It is important to note that this is a nonclassical mereology, since sums do not exist for every pair of underlapping elements. This is distinct from all earlier approaches to parthood that have been taken in biomedical ontologies [11].

### *4.5. Directionality*

A DNA molecule is understood to be composed on two strands that are complementary to each other yet contain the same information. They are essentially inverses of each other. Moreover, as mentioned in Competency Question 2, ATT codes for Isoleucine whereas TTA codes for Leucine. This illustrates how essential explicit ground rules to represent directionality are. Due to topic nuance, we must first define the related vocabulary. This set of definitions define the start (5') and stop (3') ends of a sequence, and notions of upstream and downstream:

$$\forall x \forall y\, nucleotide(x) \wedge sequence(y) \wedge 5'(x,y) \equiv \forall z\, phosphoric\_acid(z) \wedge mol(z,x) \wedge end(z,x) \tag{28}$$

$$forallx \forall y nucleotide(x) \wedge sequence(y) \wedge 3'(x,y) \equiv \forall z\, hydroxyl(z) \wedge mol(z,x) \wedge end(z,x) \tag{29}$$

$$\forall x \forall y\, sequence(x) \wedge sequence(y) \wedge downstream(x,y) \equiv$$
$$\exists n1 \exists n2 \exists b\, nucleotide(n1) \wedge nucleotide(n2) \wedge bond(b) \wedge 5'(n1,x) \wedge (n2,y) \wedge tether(x,y,b) \tag{30}$$

$$\forall x \forall y\, sequence(x) \wedge sequence(y) \wedge upstream(y,x) \equiv$$
$$\exists n1 \exists n2 \exists b\, nucleotide(n1) \wedge nucleotide(n2) \wedge bond(b) \wedge 5'(n1,x) \wedge (n2,y) \wedge tether(x,y,b) \tag{31}$$

However, interpretation of these definitions require a concept of between-ness. Circular between-ness varies slightly from non-circular between-ness as the beginning and end of the sequence is not as explicit

**Definition 6** *Non-Circular Between-ness*

$$\forall a \forall b \forall c \, between(a,b,c) \supset between(c,b,a) \tag{32}$$

$$\forall a \forall b \forall c \forall d \, between(a,b,d) \wedge between(b,c,d) \supset between(a,b,c) \tag{33}$$

$$\forall a \forall b \forall c \forall d \, between(a,b,c) \wedge between(b,c,d) \wedge (b \neq c) \supset between(a,b,d) \tag{34}$$

$$\forall a \forall b \forall c \forall d \, between(a,b,d) \wedge between(a,c,d) \supset between(a,b,c) \vee between(a,c,b) \tag{35}$$

$$\forall a \forall b \forall c \forall d \, between(a,b,c) \wedge between(a,b,d) \wedge (a \neq b) \supset between(a,c,d) \vee between(a,d,c) \tag{36}$$

**Definition 7** *Circular Betweenness*

$$\forall x \forall y \forall z \, C(x,y,z) \supset \neg C(z,y,x) \tag{37}$$

$$\forall x \forall y \forall z \, C(x,y,z) \supset C(y,z,x) \tag{38}$$

$$\forall x \forall y \forall z \forall w \, C(x,y,z) \wedge C(x,z,w) \supset C(x,y,w) \tag{39}$$

$$\forall x \forall y \forall z \forall u \forall v \, C(x,y,z) \wedge C(x,u,v) \supset C(x,u,y) \vee C(x,y,u) \vee (x = y) \tag{40}$$

## 5. Evaluation

With the proposal of these axioms, questions inadvertently arise. Namely, are these the right models, and why are these the right models? The former, also known as verification, is shown in Theorem 1. Theorem 1 formally demonstrates that the mereology of our models of gene sequences are equivalent to that of a partial linear ordering, or as connected subgraphs of a path graph as mentioned in Definition 4. Validation, or why these are the right models, can be achieved by successful answering of the the competency questions posed in Section 2.1. To achieve this, we will rephrase and encode the competency questions as first-order logic statements. This will be set as the goal in an automated theorem prover e.g. Prover9.

1. *What type of overlap is present with ATP6 and ATP8 genes?* Naturally, ATP6 and ATP8 must first be defined ontologically. ATP6 and ATP8 nucleotide sequence data will be processed via a parsing script, then further defined into terms within our ontology. For example,

$$\forall x ATP6(x) \equiv \exists a AAT(a) \wedge mol(a,x) \wedge \exists c \exists b1 \exists t \exists b2 CTG(c) TTC(t) bond(b1)$$
$$bond(b2) \wedge mol(c,x) \wedge mol(t,x) \wedge tether(a,c,b1) \wedge tether(c,t,b2) \wedge ... \tag{41}$$

   ATP6 is 681 base pairs in length, so for spatial constraint reasons, we will not include the full definition here. A similar definition is provided for ATP8. Now, classes for ATP6 and ATP8 have been defined and can be used for logical inference.
   In other words, this question asks whether there exists a some sequence of nu-

cleotides that is part of ATP6, and a separate sequence of nucleotides that is part of ATP8, that overlap as per definition in previous sections.
In first order logic, this is expressed as:

$$\forall x \forall y ATP6(x) \wedge ATP8(y) \supset \exists s1 \exists s2 sequence(s1) \wedge sequence(s2) \\ \wedge part(s1, ATP6) \wedge part(s2, ATP8) \wedge overlap(s1, s2) \quad (42)$$

2. *What do the DNA codons ATT and TTA code for?*
Again, data regarding ATT and TTA will be ontologically interpreted into relations within the ontology. For instance,

$$\forall x ATT(x) \equiv \exists a \exists t1 \exists t2 \exists b1 \exists b2 adenine(a) \wedge thymine(t1) \wedge thymine(t2) \wedge bond(b1) \\ \wedge bond(b2) codon(x) \wedge mol(a,x) \wedge mol(t1,x) \wedge mol(t2,x) \wedge tether(a,t1,b1) \wedge tether(t1,t2,b2) \quad (43)$$

Then, the question can be rewritten as:

$$\forall x \forall y ATT(x) \wedge TTA(y) \supset \exists z protein(z) code\_for(x,z) \wedge code\_for(y,z) \quad (44)$$

3. *What are the introns/exons in the given sequence?* Introns and exons are nucleotide sequences that are removed and not removed, respectively, during RNA splicing. Processes such as splicing, transcription and translation will be further defined in a follow-up ontology (The BioSequence Process Ontology).

$$\forall x \, sequence(x) \supset \exists i \exists e \, intron(i) \wedge exon(e) \wedge part(i,x) \vee part(e,x) \quad (45)$$

4. *What are conserved sequences present in the histone h1 protein given the amino acid sequence.* As shown in the above examples, histoneh1(x) will be defined as a specific class within the ontology. The question can then be rewritten as:

$$\forall x histoneh1(x) \supset \exists s \wedge conserved\_sequence(s) \wedge part(s,x) \quad (46)$$

5. *What are conservative replacements for the conserved sequence in the gene coding for the h1 histone protein?* A conserved sequence is defined as sequences in nucleic acids that are similar or identical across species. In other words, does there exist a sequence that serves the same role as the conserved sequence in the h1 histone protein? This can be rewritten as:

$$\forall s conserved\_sequence(s) \forall h histoneh1(h) \wedge part(s,x) \\ \supset \exists y conserved\_sequence(y) \wedge s \neq y \wedge (part(y,x) \wedge \neq part(s,x) \quad (47)$$

6. *Is this gene sequence circular?* In other words, this asks if the starting and ending sequence somehow overlap or connect. This can be phrased as:

$$\forall x sequence(x) \supset \exists n_1 \exists n_2 \wedge 5'(n_1,x) \wedge 3'(n_2,x) \wedge mol(n_1,n_2) \quad (48)$$

7. *How is 5'UTR and the start codon related?* In other words, do the 5'UTR of a sequence and the start codon overlap somehow?

$$\forall x start\_codon(x) \supset \exists y 5'UTR(y) \wedge mol(y,x) \tag{49}$$

## 6. Summary

We began this paper claiming that classical mereology was different from the mereology of gene sequences due to its convex nature, which has not been explicitly visited by other biomedical ontologies. We then provided mathematically rigorous definitions to represent this, and validated it with a theorem. This enables our ontology to represent and reason about structural properties such as directionality, between-ness and parthood while maintaining the semantics of a sequence. Explicit axiomatization of circular and linear between-ness also provides representation for circular gene sequences, which is also lacking in earlier approaches. This will be beneficial in applications such as gene-driven drug design. Reasoning on molecular chemistry is achieved by building upon the Molecular Structure Ontology (MoST), which is coincidentally unaligned with an upper ontology. This was a conscious design decision to maximize expressivity and simultaneously represent gene sequences as a physical molecule and a abstract information entity within the same ontology. Aligning with specific ontologies based on design necessity instead of an arbritrary upper ontology proved beneficial in this case, and could serve as a case study for other niche ontologies that don't fit perfectly into a larger framework. For future work, we will explore developing a BioSequence Process Ontology to formalize definitions for processes such as transcription, translation and splicing.

## References


[1] Ashburner M, Ball CA, Blake JA, et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. (2000)

[2] Bada, M., Eilbeck, K.: Efforts toward a more consistent and interoperable sequence ontology. In: ICBO. (2012)

[3] Chui, C., Gruninger, M.: A Molecular Structure Ontology for Medicinal Chemistry. In: Proc. of the 10th Int. Conference on Formal Ontologies in Information Systems (FOIS2016), IOS Press (2016) 317–330

[4] Eilbeck, K., Lewis, S.E., Mungall, C.J. et al.: The Sequence Ontology: a tool for the unification of genome annotations. . (2005)

[5] Khan Academy: Transcription & translation (2020) [Online; accessed April 27, 2020].

[6] Kirill Degtyarenko, Paula de Matos, M.E.J.H.M.Z.A.M.R.A.M.D.M.G.a.M.A.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research (2007)

[7] Michael Gruninger, Carmen Chui, Y.R.J.T.: A mereology of connected structures. In: FOIS. (2020)

[8] Mungall, C.J.: Formalization of Genome Interval Relations. bioRxiv (2014)

[9] Mungall CJ, Batchelor C, E.K.: Evolution of the Sequence Ontology terms and relationships. J Biomed Inform **1** (2011) 87–93

[10] Robert Hoehndorf, J.K..H.H.: The ontology of biological sequences. BMC Bioinformatics **10** (2009)

[11] Stefan Schulz, Anand Kumar, Thomas Bittner: Biomedical ontologies: What part-of is and isn't . **39** (2006)

[12] Wikipedia, the free encyclopedia: Splicing (2020) [Online; accessed April 27, 2020].