Modeling of Continuity and Change in Biology

Vinay K. Chaudhri[1] and Frank Loebe[2]

[1] Artificial Intelligence Center, SRI International, Menlo Park, CA, 94025, USA
[2] Department of Computer Science, University of Leipzig, 04009 Leipzig, Germany

Abstract. Continuity and change is a core theme in biology that refers to how genetic information is carried forward. This paper reports on our initial steps toward representing this core theme and describes the methodological background and open challenges. We define continuity and change from a conceptual modeling perspective, identify its facets that require further ontological work, and present competency questions designed to check the adequacy of its representation. Moreover, we explore whether continuity and change must be explicitly represented as primitives in the representation of biological processes or whether it can be inferred from the process structure.

Introduction

The ability to model genetic information has received considerable attention in the recent literature. A specific example is Gene Ontology (GO), which has been tremendously successful and has been widely adopted for a variety of annotation projects [1, 31]. GO serves the needs of practicing biologists and biomedical researchers by providing highly specific vocabulary for gene products and functions. We report here on our work on representing continuity and change in genetic information, a core theme in biology, as described in an introductory college textbook.

Our work is complementary to molecular and cellular level representations supported in GO because the textbook covers knowledge at organismal, species and population levels. Further, we model knowledge at a much deeper level than GO, by using an extensive set of relationships which are valuable for question answering and inference.

In the context of our immediate application, the work presented here contributes to the enrichment of the content contained in an electronic textbook. *Campbell Biology* [27] is a widely used textbook for introductory biology courses in the United States. Substantial parts of its contents have already been transformed into a knowledge base that is the basis of the prototype of an intelligent textbook that supports question answering and serves as a learning tool for students [8].

Campbell Biology is organized into eight core themes that were defined for advanced placement courses by the United States College Board [16]. Continuity and change is one of these eight core themes. Other core themes include structure and function, energy transfer, regulation, etc. A core theme captures a coherent sub-domain of

biological knowledge and has a specific thematic focus. This framework enables developing conceptual models and representations with respect to a theme that are self-contained and interrelated, thereby offering value beyond their specific use in the electronic version of the textbook.

Representing a core theme begins with a conceptualization of biological processes, entities, and their interrelations. The representation must then be shown useful for the purposes requested of the knowledge base (KB). Both tasks involve a variety of challenges, from general methodological aspects to specific modeling decisions.

In this paper, we provide an initial analysis for conceptualizing continuity and change. We focus on presenting our methodology as well as modeling challenges that arise from this core theme. We begin by giving some background on relevant existing parts of the knowledge base and the steps of core theme design with constraining aspects. We then define what is meant by continuity and change in the context of biological processes. This definition sets the scope for our conceptual analysis, where we next consider examples of entities and processes already represented in the KB and discuss whether continuity and change could be derived by automated inference or whether novel explicit encoding needs to be introduced. We define the concepts of continuity and change based on how the textbook and biology teachers define these concepts and then state them from a knowledge-engineering perspective. We also consider a few example questions that can be answered using the representations developed so far. A discussion of related work and the modeling challenges follows. We conclude with open problems and directions for future research.

1 Methods and Knowledge Base

The starting point for our work is an upper ontology and a partial encoding of the textbook knowledge that overlaps with the continuity and change concepts. This KB, called `KB_Bio_101`, is a rich biological ontology that acts as the central resource for the electronic textbook [12]. Therefore, an immediate concern in developing our representations for continuity and change is reusing existing representations when appropriate and practical. We describe the most relevant components of the KB first, as this approach eases the formulation of the methodological steps that follow in the remainder of the paper.

1.1 Component Library

Component Library (CLIB) is a foundational component of `KB_Bio_101` that serves as an important starting point for the representation work in our context. CLIB is an upper ontology which is linguistically motivated and designed to support the representation of knowledge for automated reasoning [5]. CLIB adopts four simple top level distinctions that are comparable to other widely known upper ontologies [7, 30]: (1) *entities* (things that are); (2) *events* (things that happen); (3) *relations* (associations between things); and (4) *roles* (ways in which entities participate in events).

In addition to these distinctions, CLIB provides a vocabulary of actions and semantic relationships that has proven to be easy to use by domain experts [21]. For instance,

the class Action (a direct subclass of Event) has 42 direct subclasses and 147 subclasses altogether. Examples of direct subclasses include Create, Impair, and Move. Other subclasses include Copy (which is a subclass of Create) and Break (a subclass of Damage which is a subclass of Impair).

CLIB provides a vocabulary to define the participants of an action that is inspired by a comprehensive study of case roles in linguistics [4]. These relations include agent, object, instrument, raw-material, result, source, destination, and site. Syntactic and semantic definitions for these relations are available elsewhere [11]. As an example, we consider the definition of raw-material. The semantic definition of raw-material is *any entity that is consumed as an input to a process*. The syntactic definition of raw-material is either to be the grammatical object of verbs such as *to use* or *to consume*, or to be preceded by *using*.

We consider two distinguishing features of CLIB that make it especially suitable for the work considered here. First, CLIB offers a good coverage of domain-independent actions that are needed to describe the biological processes [11]. Second, it is accompanied by a systematic account of guidelines for knowledge engineers to model the semantic relationships supported in it [11].

CLIB provides good vocabulary to encode the basic process structure (i.e., steps in a process and their relationships) and participants (i.e., entities that participate in different steps). However, it does not provide adequate guidelines and vocabulary to represent knowledge about continuity and change. After surveying many of the available ontologies, we found that no other ontology addresses these concepts adequately for our purposes.

1.2 Knowledge Base Format and Biological Contents

The knowledge base uses a fragment of first order logic which is comparable in expressiveness to datalog with function symbols [14]. The knowledge-engineering effort is structured in a way that the knowledge engineers have access to the full power of the language, but the biologist encoders can only extend the taxonomy, declare classes to be disjoint, assert qualified number constraints [2, ch. 2], and, most importantly, author existential rules [3].

The existential rules are authored through a graphical interface of the sort shown below in Figures 2 and 3. Each graph captures an existential rule: each white node of the graph (e.g., DNA-Replication in Figure 2) is universally quantified, and every other node is existentially quantified. Such a graph has the intuitive meaning that for every instance of the class represented by the universally quantified node, there exist instances of classes represented by the gray nodes that are related to each other by the relations in the graph. A more formal description is available elsewhere (cf. e.g., [10, 21]). Moreover, all concepts in the KB are inserted into its taxonomy, (i.e., a subsumption (poly-)hierarchy of all concepts) the upper levels of which are constituted by CLIB.

Relationship between Structure and Function is a core theme that is already captured in the KB [9] and that offers the greatest potential for reusing the knowledge already represented in the KB for continuity and change. Besides functional interconnections, we have already available representations of mereological modeling of entities

and events. Consequently, for many notions that are central to continuity and change, the KB already contains representations from a structural and functional point of view.

1.3 Methods

Our goal in developing approaches to representing a new core theme within the KB is a set of ontology-based modeling patterns that form the basis for converting the biological knowledge of the theme into extensions of the KB. The latter work can then proceed mainly sequentially over chapters of the textbook and is distributed over a larger team of domain experts trained in encoding knowledge.

We pursue the following steps in designing basic core theme representations for the KB and assessing their adequacy for question answering: (1) Synthesize a textual definition of the core theme; (2) Establish a set of informal competency questions; (3) Select/identify key concepts for the theme; (4) Propose/verify the position of the key concepts in the KB's taxonomy; (5) Draft/determine/refine prototypical concept graphs for core theme coverage; and (6) Conceive of possible reasoning patterns and simulate tests of question answering.

In practice, these steps are not followed in a strict sequence and require multiple iterations. The later steps depend on earlier decisions, but may also have an impact on the former steps because they provide a limited form of validation. Of course, convergence of this process is sought while the steps are repeated several times. Our process corresponds to an evolving approach for core theme representations, in line with the prevailing life cycle model for ontologies as judged in [19, sect. 3.3.8, p.153f.].

The textual definition of the theme provides scope and focus for its coverage in the KB. Competency questions [20] also contribute to this, and they are adopted for two further reasons. First, they are a well-established means for evaluating ontology representations (cf. e.g., [26]). Second, question-answering is the main application. We adapt the idea of informal competency questions by considering a small set of *diagnostic questions* for testing representations and a large set of *educationally useful questions* as determined by biology teachers and students.

Key concepts (and possibly relations) receive particular attention in the remaining steps, as major anchor points of the targeted modeling patterns. Usually, entities and events are identified initially, with those choices leading to respective relations and roles. To include concepts that are relevant to this core theme in the KB, these elements must be placed in the taxonomy. In many cases, key concepts are already present in the KB. Then the task changes to assessing the current concepts: whether their hierarchical position and modeling already supports the perspective of the theme, its addition to the representation, or whether re-engineering is required. In the latter case, retaining control over the effects of changes is especially important. Finally, step 6 is aimed at question-answering and evaluates representation patterns against diagnostic and educationally useful questions.

A side aspect of our work concerns the faithfulness to the original biology textbook [27]. Such fidelity is desirable for the users of the electronic version to recognize consistent matches between replies to their questions from the KB and relevant parts of the text. Further, it serves as a simplifying constraint that limits and stabilizes the

knowledge to be formalized. Future alternative applications of the KB may thus require extensions, for example, to cover more details for specialized domains.

2 Defining Continuity and Change

The first step of our procedure is concerned with finding or developing a definition of the core theme. We start from characterizations provided by the College Board, the authors of the textbook, and the biology teachers we work with.

The College Board syllabus [16] provides the following definition for continuity and change: *all species tend to maintain themselves from generation to generation using the same genetic code. However, there are genetic mechanisms that lead to change over time, or evolution*.

Campbell Biology [27, ch. 1, p. 8 ff.] starts the description of the theme as follows: *The division of cells to form new cells is the foundation for all reproduction and for the growth and repair of multicellular organisms*. After referring to chromosomes as the main carriers of genetic material, the theme outline continues on the structure and function of deoxyribonucleic acid (DNA) together with its ability to store information, further highlighting the processes of replication and gene expression. It also establishes the link between the genome of organisms and genomics, as the study of genes and sets of genes within species, as well as cross-species genome comparison.

The key aspects of continuity and change from the perspective of biology teachers include the following: (1) it involves genetic information; (2) continuity is about the maintenance of the fidelity of the information from generation to generation, cell to cell, organism to organism, or species to species; (3) change is about loss or altering of fidelity of the information from generation to generation, cell to cell, organism to organism, or species to species; (4) it often involves a measurable or observable outcome, and this outcome relates directly to continuity or change of information.

We synthesized the three different perspectives into a definition, which evolved during later steps in our procedure to the following result:

> Continuity and change concerns genetic information and its phenotypic expression, where the basic form of genetic information is given by nucleotide sequences. Continuity and change are considered with respect to inheritance, more precisely regarding the flow of genetic information: (i) within events occurring in the transition from generation to generation at different levels, namely of cells (or viruses), organisms, or populations; and (ii) in the evolutionary development of species. Information flow requires information units (i.e., entities that carry information), which are complemented by observable effects of that information. We distinguish: (a) sub-cellular information units that are physical parts of cells (or viruses) (e.g., nucleic acid molecules (DNA, RNA), genes/alleles, and chromosomes); (b) aggregated information units that are derived from the sub-cellular units (e.g., genotype, genome, gene pool); and (c) traits/phenotypes of organisms, which are determined by genetic information. On this basis, *continuity* refers to maintaining the sameness of genetic information as well as of information units themselves, the latter supporting

the former, and of phenotypes. *Change* refers to events generating differences in genetic information, in information units (if affecting carried information), or in resulting phenotypic characteristics.

3 Representation from the Perspective of Continuity and Change

This section concerns steps 3–5 of our procedure (see Section 1.3). We start by describing the representation of entities that are involved in continuity and change followed by the representation of the processes. There is no formal separation among those two kinds. Concept graphs can relate entities to processes and vice versa, cf. Figure 3. Diagnostic and educationally useful questions as resulting from step 2 are discussed in combination with initial tests based on these representations (step 6) in the subsequent Section 4 on question answering.

3.1 Representing Entities in Continuity and Change

The key entities involved in continuity and change that are covered in [27] are genetic information units, namely nucleotide, codon/anti-codon, DNA, RNA, DNA strand, RNA strand, allele, gene, chromatid, chromosome, genotype, genome, and gene pool. These entities span across different levels of biological organization. For example, while a nucleotide corresponds to the molecular level, gene pool is defined only at the population level. From the point of view of inheritance at the level of organisms, the notions of trait and phenotype should be included, as well. To represent all those concepts in the KB, we need to find suitable positions for them in the taxonomy as well as provide their detailed definitions, eventually in the form of concept graphs.

Figure 1 shows the current positioning of (most of) these entities in the taxonomy of the KB. Inspecting some of the corresponding concept graphs, both Genotype and Gene-Pool have an Allele as an element. Gene-Pool aggregates the Alleles at the level of a population, whereas Genotype aggregates them at the level of an individual. A Chromosome has-part a DNA which has-part a DNA-Strand which in turn has-part a Gene. To define the relationship between Gene and Allele, a novel domain-specific relationship called has-variant is introduced.

Though the current modeling is backed up by *Campbell Biology* [27], this support does not automatically render the representation adequate from the perspective of continuity and change. One reason for that is the use of concepts in different, partially incompatible "flavors" throughout the book, especially such key notions that occur with high frequency. This is not surprising for a natural language source, but it counteracts finding a consistent model. For example, for the concept of gene, we found statements supporting at least four competing views of whether gene is subsumed by DNA, DNA region, Nucleic Acid Sequence, or, generally, Information. The situation is similar for most key concepts in continuity and change.

In some cases, we can consult existing ontologies or ontological analyses, such as [25] for gene or [22] for biological sequences. This can be instructive in terms of possibly useful distinctions or novel aspects. For example, [22] distinguishes sequences composed of molecules (molecular tokens) from sequence representations (e.g., as string

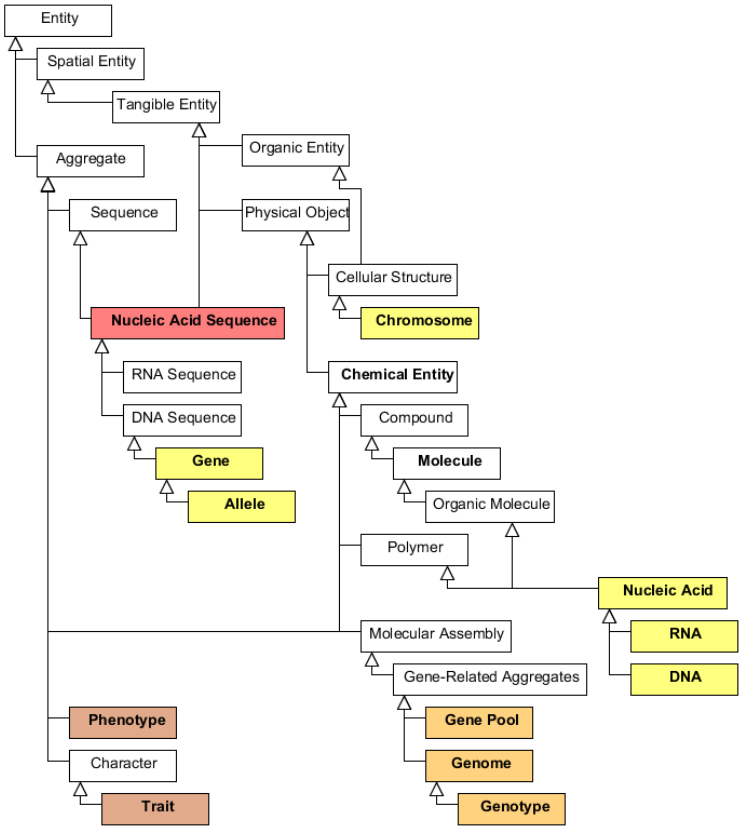

Fig. 1. Key entities associated with continuity and change in the ontology. The coloring reflects the distinction of (a) sub-cellular and (b) aggregated information units, and (c) traits / phenotypes.

tokens) and abstract sequences (as types that can be shared among tokens, bearing information). However, often drawing a distinction that leads to two (or more) concepts instead of one beforehand entails a multiplication of model parts that depend on the "split" concept and remain largely analogous otherwise. Accordingly, the four views on gene or the three on sequences cannot be adopted automatically by four or three distinct concepts. Another brief illustration concerns a major decision for representing continuity and change, namely whether to separate the aspect of information into distinct concepts. This approach may be reasonable for having an explicit entity that could be the object of continuity in certain processes. Contrariwise, all key concepts are information units, which may easily lead to their duplication into additional information entities.

In general, we tackle this problem of a good trade-off for adequately fine-grained, but minimal conceptual distinctions by cycling through steps 4–6 of our procedure, aiming at a converging model. For example, the current modeling of Gene (still as a single concept) in the KB supports views as a tangible entity as well as a nucleotide

sequence, based on multiple inheritance in the case of Nucleic Acid Sequence. It dispenses with gene information as a distinct concept. Nevertheless, we see the general trade-off problem as a challenge with further potential for improved methodological approaches.

3.2 Representing Continuity and Change in Processes

Key processes that involve continuity and change of genetic information are DNA replication, mutation, meiosis, sexual reproduction, and natural selection. Each of these is a complex process that can be modeled using a combination of primitive actions from CLIB. A straightforward approach to capture continuity and change in these processes is to identify various CLIB actions as pertaining to continuity and change. For example, the CLIB actions of Copy and Duplicate involve continuity and Add, Delete, etc. involve making a change. An automated reasoning procedure can then identify which of those actions involve genetic information and use that information for reasoning about continuity and change. An alternative approach would be to use continuity and change themselves as primitive actions and include them as part of the process representation. This approach has a greater likelihood of producing correct inferences but also requires additional representation work. We will illustrate these design choices in our approach by taking DNA replication as an example.

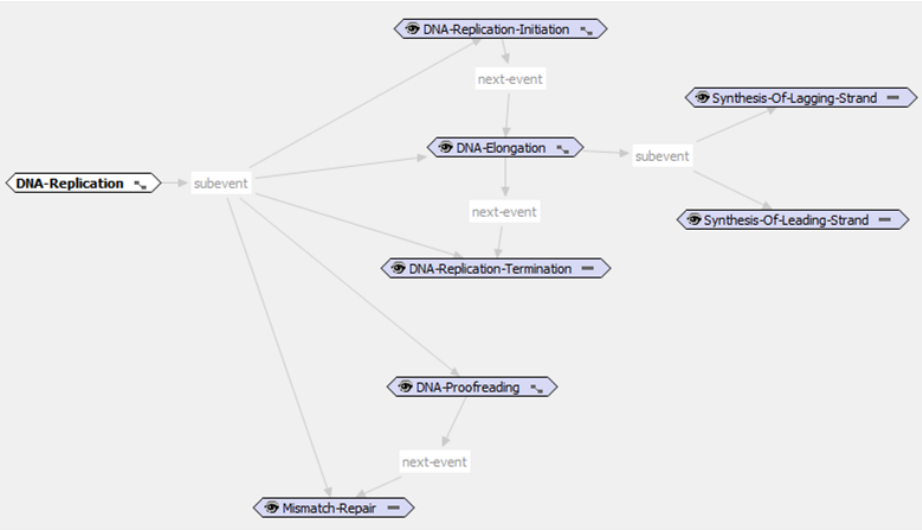

Fig. 2. Structure of major steps in DNA replication.

In Figure 2, we show the major steps of DNA-Replication. The copying process of DNA, accounting for the continuity of genetic information, is spread across its three major steps: (1) initiation, (2) elongation, and (3) termination. The bulk of the copying happens during Synthesis-Of-Leading-Strand and Synthesis-Of-Lagging-Strand.

However, DNA replication does not lead to perfect copies of DNA. A "regular" change is due to the incapability of reaching the very end of the chromosomes, such that the telomeres (regions of repetitive DNA at the ends of chromosomes) are shortened by each replication. Sometimes, errors occur in the replication process. A process called DNA-Proofreading verifies the newly constructed DNA and, in case of errors, corrects them by invoking a process called Mismatch-Repair. These steps, however, are not able to correct all the errors. The effectiveness of the repair process can be explicitly modeled by assigning it a specific property. Any errors that are left uncorrected represent mutations of the DNA. Not all mutations are due to replication errors, though.

If we were to infer the continuity during DNA-Replication, we will have to gather specific copying operations in each of its steps. To infer the changes, we will have to observe that the process can have some errors that are not corrected by the built-in mechanisms. Creating axioms that allow for such automated inferences is complex.

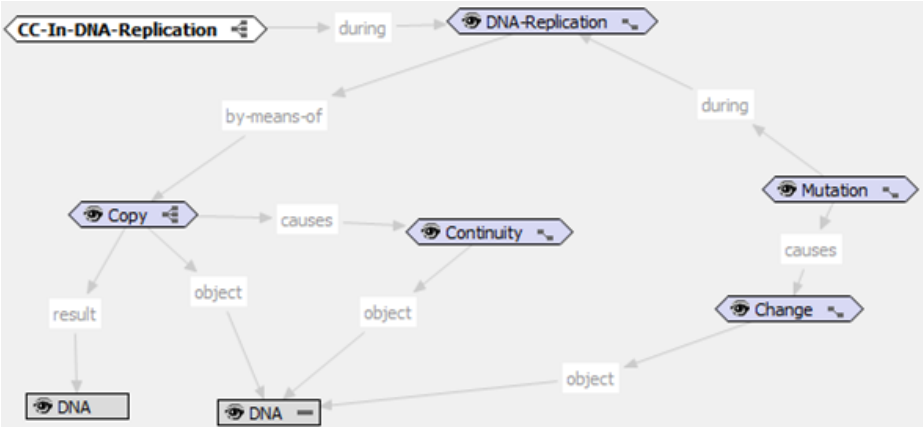

Fig. 3. Continuity and change in DNA replication.

In Figure 3, we introduce a continuity and change perspective on the process of DNA-Replication. In this representation, we first state that CC-In-DNA-Replication is a composite process that happens during DNA-Replication. We next state that a major mechanism in DNA-Replication is Copy (this is indicated by using the by-means-of relationship). The Copy action in this perspective can be viewed as an abstraction of multiple steps in the perspective of Figure 2. The Copy causes the Continuity of DNA. We further indicate that DNA-Replication has a sub-event of Mutation that causes a Change in the DNA. Thus, by introducing this additional perspective, we can more directly state the continuity and change associated with DNA-Replication.

In principle, we could have combined the representation of Figures 2 and 3 into a single model, but that approach would have led to a considerable conceptual clutter. Separating this information into a distinct concept graph provides different modeling views on the KB for human editors. The reasoning is performed over the overall first-order representation of the KB which covers all the concept graphs including entities

and events. For example, the node DNA-Replication in Figures 2 and 3 refers to the very same unary predicate in the KB.

Contrary to differentiating biological concepts as in Section 3.1, introducing graphs for perspectives does not lead to "independent" concepts. DNA-Replication is a biological concept (interrelated with others), whereas CC-In-DNA-Replication has a different status in that it depends on DNA-Replication and is not a naturally occurring concept in *Campbell Biology*. We have found similar perspectives to be useful for modeling energy transfer and regulation [10]. We have performed analogous analysis for the processes of Meiosis, Recombination, Sexual-Reproduction, and Natural-Selection, where a similar strategy seems to be effective. This need to separate the representation of a concept into multiple perspectives is different from the work on connecting independently developed ontologies [17] in the following way: multiple perspectives are defined by the same user and need to be present in the same KB as opposed to being defined by different users and existing in different KBs.

4 Answering Questions Using the Representation

Let us now consider the questions that we wish to answer using the representations of continuity and change. The questions directly address the steps 2 and 6 in our procedure (see Section 1.3). We consider two families of questions: diagnostic and educational. Diagnostic questions are aimed at testing whether the system adequately represents continuity and change, and they are purely driven by the representation. The educationally useful questions are gathered by convening a focus group of teachers and students who provide a set of questions that they wish to pose to a computational tool.

Our current set of diagnostic question patterns for the core theme is as follows. This set results from the first iteration over all steps, with primary feedback from steps 4–6.

D1 What remains the same/changes during X?
D2 What causes the continuity/changes of X during Y?
D3 Describe continuity/changes during a process X.
D4 What is an example of a process that maintains the continuity of/changes X?
D5 What does X contribute to continuity/change of Y?
D6 Which processes contribute to continuity/changes of X during process Y?

We consider example instantiations of these question patterns that involve the concept of DNA-Replication, and discuss resolutions and open aspects of answer generation from the knowledge base (KB). More detailed descriptions of our query answering methods are available elsewhere [13, 15]. All the questions that we consider can be answered using the specification of the behavior of processes. None of the questions that we consider here require executing a process.

D1 will be instantiated as: *What changes during* DNA-Replication? If we use the continuity and change perspective of this process (cf. Figure 3), we can derive a straightforward answer to this question by detecting DNA as changing due to Mutation. By examining the detailed representation of DNA-Replication, one can further infer that the replication and the proofreading processes themselves are faulty and can cause changes. Another type of change, limited to eukaryotes, is that telomeres are shortened every

time DNA replication occurs, because DNA-Polymerases cannot replicate DNA efficiently at the end of linear chromosomes. We will need to capture this aspect in the detailed structural model of DNA-Replication as well as in the continuity and change perspective of the process.

Let us next consider an instantiation of D2: *What causes the continuity of* Gene*s during* DNA-Replication? Relying on the continuity and change perspective of this process again, together with the linkage between the concept graphs of Gene and DNA, we can answer this question at an abstract level by saying that the Copying of DNA causes the continuity of genes. However, for a detailed answer, we will need to examine the more detailed representation of the process to argue that a newly synthesized DNA molecule is identical to the original, thereby maintaining the continuity of genes.

Corresponding arguments that are desirable from the perspective of biology teachers can be found in a sample response to an instance of pattern D3, namely *Describe continuity during* DNA-Replication: "Complementary base pairing and semi-conservative replication ensure that a newly synthesized DNA molecule is identical to the original. The DNA double helix is opened, and each strand serves as template for making a new strand. DNA polymerase enzymes add nucleotides to the new strand according to the base-pairing rules (A-T and C-G), so the continuity of the original DNA molecule is maintained. Additionally, during DNA replication, DNA polymerases proofread each nucleotide against its template as soon as it is added to the growing strand. Upon finding an incorrectly paired nucleotide, the polymerase removes the nucleotide and then resumes synthesis."

Producing such an answer requires iterating through all the steps that are involved in DNA-Replication, identifying how they are contributing to the continuity of the DNA molecule and explaining them. From the modeling and reasoning perspective, this involves the challenge of determining which levels of (mereological) granularity should be considered when constructing answers. For example, in the above case, including only concepts that have a direct subevent link from DNA-Replication to themselves is insufficient.

An example instance of D4 is: *What is an example of a process that changes* DNA? Based on the continuity and change perspective, this question is straightforward to answer as Mutation. A more elaborate answer will also point out other processes such as insertion or removal of transposons or viral DNA, which will appear in both perspectives.

We can instantiate D5 as follows: *What processes contribute to changes of* DNA *during* DNA-Replication? The answer to this question is that during DNA replication, DNA-Polymerase inserts an incorrect nucleotide approximately once every 1,000 nucleotides. If this error is not corrected by the proofreading process, a Mutation has resulted, causing change in the DNA sequence. This answer can follow if we represent in our model a causal link between failure of proofreading to correct an error in DNA-Replication and Mutation.

Our suite of educationally useful questions on continuity and change has approximately 100 different questions. Here, we consider only a few examples of such questions that are related to DNA-Replication.

E1 What happens if crossing over occurs in the middle of a gene?

E2 What would happen to the chromosomes in eukaryotes if teleromerase were lacking?
E3 What is the difference between a translocation and an inversion?
E4 Due to their structure, DNA polymerases can add nucleotides only to the 5 prime end of a primer of a growing DNA strand, never to the 3 prime end. True or False? Explain in terms of the antiparallel arrangement of the double helix the effect on replication.
E5 A father with blue eyes and blonde hair, a mother with green eyes and brown hair have four children, what are the features the kids would have if DNA replication followed the dispersive model?
E6 The disorder xeroderma pigmentosum is caused by what? Can it be passed from generation to generation in species and organisms? If so, how can this be prevented during DNA replication error correction?

We have done only a preliminary analysis of such questions to identify a few candidate question formats. For example, E1 and E2 follow the format of: What is the effect if X were to happen? E3 has the format of: What is the difference between X and Y? We have an extensive prior work on questions in this form [13]. E4 could be reduced to two sub-questions: Is it true that DNA polymerase can add nucleotides to the 5 prime end of a DNA strand? Is it true that DNA polymerase can add nucleotides to the 3 prime end of a DNA strand? The sub-questions help answer the overall question. E5 requires an explicit representation of the dispersive model and using it to predict how certain features get passed on during replication. E6 requires knowing the connection between xeroderma pigmentosum and the mutations, and reasoning that preventing mutations would require the error correction during replication to be successful.

5 Related Work and Discussion

We look at related work from two angles: (1) biomedical ontologies and models of biological knowledge, and (2) conceptual modeling issues that need to be addressed.

A survey of biomedical ontologies revealed that there is no ontology that deals with the notions of continuity and change we considered. Most entity concepts such as Gene and DNA can be found in multiple biomedical ontologies (e.g., in SNOMED-CT[1], the NCI-Thesaurus[2], the Sequence Ontology[3], the Gene Regulation Ontology[4] (GRO), and in the top-domain ontology BioTop[5] [6], to name a few[6]). Direct reuse (e.g., of fragments of these ontologies) remains limited and usually involves re-engineering for integration into our ontology/knowledge base. Taken together, the existing sources do not provide a consistent, integrated picture. Their coverage is typically limited to classification hierarchies with short, natural language definitions of concepts, in some cases

[1] http://www.ihtsdo.org/snomed-ct
[2] http://ncit.nci.nih.gov/
[3] http://sequenceontology.org/
[4] http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html
[5] http://www.imbi.uni-freiburg.de/ontology/biotop/
[6] All of these are also available from the National Center for Biomedical Ontology (NCBO) through its BioPortal, http://bioportal.bioontology.org/ .

extended with a few semantic relations among them, such as part-of. Our work relies on a much larger set of relations, which enables explicitly stating and reasoning over detailed interconnections among concepts. The same applies to ontological analysis that focuses on very few specific concepts, where we mention [22, 25] in Section 3.1 on representing entities. The depth of such analysis is adequate for our needs, but re-engineering and harmonization with the present KB are still necessary. We identify three major problems that our work addresses:

1. Methodological guidance on explicitly representing ontological distinctions (potentially with multiplicative effects) and/or using "overloaded" concepts, exemplified by Gene, NucleicAcidSequence, and (genetic) information.
2. Supporting different perspectives on concepts (cf. the case of the perspective-based concept graph in Figure 3).
3. Representation of and reasoning across levels of granularity.

In the context of biomedical ontologies, we see no method to address the first issue. As noted, a fundamental question for us concerns how to represent genetic information. However, available conceptualizations vary significantly: BioTop explicitly distinguishes between gene (subsumed by material object) and genetic information (subsumed by information object), whereas gene in GRO is a subconcept of information biopolymer (with an implicit information aspect), itself subsumed by molecular entity. The NCI-Thesaurus comprises the concept gene (as a top-level concept), and no explicit concept of genetic information.

The second issue is related to the first, but is oriented at viewing one concept from different angles. It does not arise before we start the detailed modeling of concepts (i.e., transcending taxonomies). Clearly, multiple perspectives are widely accepted, for example, in conceptual or systems modeling. Using different models for distinct perspectives or views, even in different (sub-)languages, is also the case in the Unified Modeling Language [28] (cf., e.g., the discussion of *viewpoint* [28, p. 678]). But much freedom appears to be left to the modeler as to when and how a model is dissected into several models capturing distinct perspectives.

Addressing the third problem, in [23] an elaborate theory of granularity is presented, based on ontological and logical analysis. However, to apply this approach a domain-granularity framework needs to be derived from the domain- and implementation-independent theory of granularity. We need to evaluate this approach further to judge the benefits in our particular setting, as it seems to be complex. For BioTop, the authors of [29] aim at a neutral position with respect to granularity issues, in the context of integrating biomedical ontologies.

For all three issues, we see the need and potential for novel solutions by methodological means, although these problems have been met in earlier and ongoing work. No generic methods were available in the literature that we could readily apply in our context. Addressing these problems satisfactorily requires further research in conceptual modeling, ontology design patterns, applied ontology, and modeling and using context.

Solutions to the problems of representing information objects, granularity and multiple perspectives are not limited to the domain of continuity and change in biology. Our present focus on biology has an advantage that it gives us concrete versions of

these problems to be solved, with a clear set of computational and application requirements. Our modeling proposals for continuity and change are domain-specific, but the underlying techniques lend themselves to adoption in other areas. As an example, in the context of business process modeling, there is a need for process representations that allow for querying for the continuity (or maintenance) of certain business objects for which similar modeling patterns can be adopted.

6 Summary and Conclusions

Continuity and change is an extremely rich topic area in biology, and in this paper, we have merely scratched its surface, with an eye on modeling methods and open problems. Even though our work is preliminary, it does make several important contributions.

First, we presented a definition of this core theme that unifies the perspectives from U.S. College Board, biology teachers, and *Campbell Biology* [27]. The need to augment and streamline the theme description from the textbook highlights the complexity of biological knowledge and the value added by viewing it from conceptual modeling perspectives. Therefore, we see the comprehensive definition as a contribution in itself.

Second, we outlined a preliminary ontology of the entities involved in continuity and change. An obvious connection exists between these entities and information objects, which is a topic of great current interest in upper ontologies. We hope that our initial analysis will further facilitate the convergence between the theory on information objects and the entities involved in continuity and change.

Third, we argued that supporting multiple perspectives on processes is required, so that certain aspects can be factored out into separate models. Clearly, more work is needed to develop guidelines on how one should decide which information should go into each perspective and how different perspectives should be related to each other.

Fourth, we presented several examples of questions that need to be answered using representations of continuity and change. Many of the diagnostic questions follow in a straightforward manner from the representation. Answering these questions, however, requires reasoning across multiple perspectives, for which we do not yet have an adequate theory. These questions also require reasoning with processes, in some cases how processes behave, and in others, effects if certain processes did or did not happen.

Finally, we outlined our steps in core theme design and related constraints, framing the analysis of continuity and change. Besides supporting multiple perspectives, two other major challenges for systematically improving the current representations were identified, namely more guidance on explicitly representing ontological distinctions (e.g., for the notion of gene) and improved support of granularity.

In summary, we hope that this paper yields a good overview of problems and challenges in representing and reasoning over continuity and change, and that it provides promising starting points for further ontological and methodological research.

Acknowledgments

This work work has been funded by Vulcan Inc. and SRI International. We thank Nikhil Dinesh, Sue Hinojoza and William Webb for numerous discussions that helped develop the ideas presented in this paper.

References

1. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
2. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK, 2nd edition, 2010.
3. Jean-François Baget, Michel Leclère, Marie-Laure Mugnier, and Eric Salvat. On rules with existential variables: Walking the decidability line. *Artificial Intelligence*, 175(9):1620–1654, 2011.
4. Ken Barker, Terry Copeck, Stan Szpakowicz, and Sylvain Delisle. Systematic construction of a versatile case system. *Journal of Natural Language Engineering*, 3(4):279–315, 1997.
5. Ken Barker, Bruce Porter, and Peter Clark. A library of generic concepts for composing knowledge bases. In Yolanda Gil, Mark Musen, and Jude Shavlik, editors, *Proceedings of the First International Conference on Knowledge Capture, K-CAP 2001, Victoria, British Columbia, Canada, Oct 22–23*, pages 14–21, New York, 2001. ACM Press.
6. Elena Beisswanger, Stefan Schulz, Holger Stenzhorn, and Udo Hahn. BioTop: An upper domain ontology for the life sciences: A description of its current structure, contents and interfaces to OBO ontologies. *Applied Ontology*, 3(4):205–212, 2008.
7. Stefano Borgo and Claudio Masolo. Ontological foundations of DOLCE. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, chapter 13, pages 279–295. Springer, Heidelberg, 2010.
8. Vinay K. Chaudhri, Britte Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. *Inquire Biology*: A textbook that answers questions. *AI Magazine*, 34(3):55–72, 2013.
9. Vinay K. Chaudhri, Nikhil Dinesh, and H. Craig Heller. Conceptual models of structure and function. In Klenk and Laird [24], pages 255–271.
10. Vinay K. Chaudhri, Nikhil Dinesh, and Stijn Heymans. Conceptual models of energy transfer and regulation. In Garbacz and Kutz [18]. *In Press*.
11. Vinay K. Chaudhri, Nikhil Dinesh, and Daniela Inclezan. Three lessons for creating a knowledge base to enable explanation, reasoning and dialog. In Klenk and Laird [24], pages 187–203.
12. Vinay K. Chaudhri, Daniel Elenius, Sue Hinojoza, and Michael Wessel. KB Bio 101: Content and challenges. In Garbacz and Kutz [18]. *In Press*.
13. Vinay K. Chaudhri, Stijn Heymans, Adam Overholtzer, Aaron Spaulding, and Michael Wessel. Large-scale analogical reasoning. In Carla E. Brodley and Peter Stone, editors, *Proceedings of 28th AAAI Conference on Artificial Intelligence, AAAI 2014, Québec City, Québec, Canada, Jul 27–31*, pages 359–365, Menlo Park, California, USA, 2014. AAAI Press.
14. Vinay K. Chaudhri, Stijn Heymans, Tran Cao Son, and Michael A. Wessel. Object-oriented knowledge bases in logic programming. *Theory and Practice of Logic Programming*, 13(4-5):Online–Supplement, 2013.
15. Vinay K. Chaudhri, Stijn Heymans, Michael A. Wessel, and Tran Cao Son. Query answering in object oriented knowledge bases in logic programming: Description and challenge for ASP. In Michael Fink and Yuliya Lierler, editors, *Proceedings of the 6th International Workshop on Answer Set Programming and Other Computing Paradigms, ASPOCP 2013, Istanbul, Turkey, Aug 25*, number abs/1312.6138 in arXiv.org, Computing Research Repository (CoRR), Ithaca, New York, 2013. Cornell University Library.
16. College Board. Biology: Course description. `http://apcentral.collegeboard.com/apc/public/repository/ap-biology-course-description.pdf`, 2010.

17. Bernardo Cuenca Grau, Bijan Parsia, and Evren Sirin. Combining OWL ontologies using $\mathcal{E}$-Connections. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):40–59, 2006.
18. Pawel Garbacz and Oliver Kutz, editors. *Proceedings of the 8th International Conference on Formal Ontology in Information Systems, FOIS 2014, Rio de Janeiro, Brazil, Sep 22-25*. IOS Press, Amsterdam, 2014. *In Press*.
19. Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Advanced Information and Knowledge Processing. Springer, London, 2004.
20. Michael Grüninger and Mark S. Fox. Methodology for the design and evaluation of ontologies. In Douglas R. Skuce, editor, *Proceedings of the IJCAI 1995 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada, Aug 19–20*, pages 6.1–10, 1995.
21. David Gunning, Vinay K. Chaudhri, Peter E. Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David D. McDonald, Sunil Mishra, John Pacheco, Bruce Porter, Aaron Spaulding, Dan Tecuci, and Jing Tien. Project Halo update: Progress toward Digital Aristotle. *AI Magazine*, 31(3):33–58, 2010.
22. Robert Hoehndorf, Janet Kelso, and Heinrich Herre. The ontology of biological sequences. *BMC Bioinformatics*, 10:377.1–11, 2009.
23. Catharina Maria Keet. *A Formal Theory of Granularity: Toward enhancing biological and applied life sciences information systems with granularity*. PhD thesis, KRDB Research Centre for Knowledge and Data, Free University of Bolzano, Bolzano, Italy, 2008.
24. Matthew Klenk and John Laird, editors. *Proceedings of the Second Annual Conference on Advances in Cognitive Systems, Baltimore, Maryland, USA, Dec 12–14*. Cognitive Systems Foundation, 2013.
25. Hiroshi Masuya and Riichiro Mizoguchi. An ontology of gene. In Ronald Cornet and Robert Stevens, editors, *Proceedings of the 3rd International Conference on Biomedical Ontology, ICBO 2012, KR-MED Series, Graz, Austria, Jul 21–25*, volume 897 of *CEUR Workshop Proceedings*, Aachen, Germany, 2012. CEUR-WS.org.
26. Fabian Neuhaus, Amanda Vizedom, Ken Baclawski, Mike Bennett, Mike Dean, Michael Denny, Michael Grüninger, Ali Hashemi, Terry Longstreth, Leo Obrst, Steve Ray, Ram Sriram, Todd Schneider, Marcela Vegetti, Matthew West, and Peter Yim. Towards ontology evaluation across the life cycle: The Ontology Summit 2013. *Applied Ontology*, 8(3):179–194, 2013.
27. Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Campbell Biology*. Benjamin Cummings (impr.), Pearson, Boston, 2011.
28. James Rumbaugh, Ivar Jacobson, and Grady Booch. *The Unified Modeling Language Reference Manual*. Addison Wesley, Reading, Massachusetts, 2nd edition, 2005.
29. Stefan Schulz, Martin Boeker, and Holger Stenzhorn. How granularity issues concern biomedical ontology integration. In Stig Kjær Andersen, Gunnar O. Klein, Stefan Schulz, Jos Aarts, and M. Cristina Mazzoleni, editors, *eHealth Beyond the Horizon - Get IT There: Proceedings of MIE2008: The XXIst International Congress of the European Federation for Medical Informatics, Göteborg, Sweden, Aug 25–28*, volume 136 of *Studies in Health Technology and Informatics*, pages 863–868, Amsterdam, 2008. IOS Press.
30. Andrew D. Spear. Ontology for the twenty first century: An introduction with recommendations. Manual, Institute for Formal Ontology and Medical Information Science (IFOMIS), Saarbrücken, Germany, 2006. `http://www.ifomis.org/bfo/documents/manual.pdf`.
31. The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Research*, 36:D440–D444, 2008.