Entities with Genetic Information – An Initial Perspective from the Core Theme of Continuity and Change in Biology

Frank Loebe\textsuperscript{1*} and Vinay K. Chaudhri\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, University of Leipzig, Germany
\textsuperscript{2}SRI International, Menlo Park, California, USA

ABSTRACT

Genetic information is transferred through biological processes studied at different levels of granularity. Focusing on the continuity and change of genetic information and its phenotypic expression thus yields a cross-cutting theme in biology. The inspiring aim of this work is to integrate this theme into the knowledge base of an electronic biology textbook. The paper describes first steps in this regard. We present a definition of the theme to set its scope, sketch competency questions for future evaluation, and systematically identify a set of information (bearing) units as key notions of the theme.

Due to the central role of genetic information in biomedical research, continuing the work towards the taxonomic classification of those key notions connects with their occurrence in biomedical ontologies. Besides the taxonomy in the knowledge base, we briefly survey selected ontologies regarding the notions of gene, DNA, and DNA sequence. Beyond specific analyses, this leads us to conclude with the need for novel solutions in dealing with multiple, but integrated perspectives on biological terms and for the efficient management of detailed ontological distinctions.

1 INTRODUCTION

Genetic information is central in much of biomedical research and meanwhile constitutes a large set of databases with a vast amount of genetic data. Supporting the annotation of genetic information has led to the immense success of the Gene Ontology (GO)\cite{1,16}, consequentially pioneering the overall field of biomedical ontologies. GO and many other bio-ontologies are applied by practicing biologists and biomedical researchers. On the one hand, they already serve certain needs of researchers, on the other hand new applications and user requests promote the continued development of the ontologies. They evolve not only in the light of new terms that need to be integrated, but further, for example, based on the alternative analysis of ontology terms and their relationships.

In the present work we aim at analyzing and representing continuity and change in genetic information, identified as a core biology theme by the United States College Board\cite{17}. We focus on providing a scoping definition of the theme itself and propose a way to organize its notions of major relevance. On that basis, we initially study the relationship with representative biomedical ontologies overlapping this domain.

A large effort that aims at question answering and inference sets the scene for our work. Its immediate application is the creation of an electronic version of Campbell Biology (abbreviated by KB-Bio, as a system). In order to answer questions from that KB, the biological content must be captured at a deeper level of detail, involving a much higher amount of formally captured interrelationships than is commonly found in biomedical ontologies. Our work covers an initial stage in enriching its content to eventually include the core theme of continuity and change.

There are eight core themes in \textit{Campbell Biology} as they were defined by the U.S. College Board for advanced placement courses\cite{17}. Other core themes include structure and function\cite{6}, energy transfer and regulation\cite{7}, etc. Some of them have already been represented in the KB. They all capture a coherent sub-domain of biological knowledge and have a specific thematic focus.

Representing a core theme starts with refining the brief definition of the theme by the College Board for clarity about that focus. A conceptualization of respective biological processes, entities, and their interrelations must then be devised and represented in the KB, and tested by demonstrating that the representation is feasible for the purposes requested of the knowledge base. Our goal in these regards is a set of ontology-based modeling patterns that serve as sample cases and guidelines for converting the overall biological knowledge of the theme into extensions of the KB. The latter work is distributed over a larger team of domain experts trained in encoding knowledge accordingly. To a large extent this can proceed chapterwise, given the basis of the core theme representation.

In this paper, our aims are two-fold. In the first part, we detail our methods in designing a core theme representation for the knowledge base in section 2 and then describe our initial results for the core theme of continuity and change in section 3. This includes a scoping definition of the theme, an outline of competency questions for later evaluation, and the identification of key entities together with summarizing characterizations, including distinct views as available from\cite{13}. The second part of the paper, presented in section 4, focuses on the taxonomic placement of selected key entities in comparison with existing ontologies in the biomedical domain. On that basis we discuss and conclude with general, interrelated research tasks in ontological analysis and ontology representation and argue that solutions to them would be of major impact, not only in the biomedical domain, but with immediate applications therein.

2 METHODS

2.1 Core Theme Design

In the overall process of designing and checking basic core theme representations, i.e., targeting at a set of ontology-based modeling patterns for a particular theme, we pursue these steps, cf. also\cite{8}:

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 taxonomy of the KB;
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, the steps are not followed in a strict sequence and their results interact with each other, based on forms of dependence and validation occurring between the steps. While results on earlier steps may be improved by revisiting them based on later steps, thus going through some iteration, we seek convergence of this overall process.

In the sections below we present the current status of our work for the theme of continuity and change, focusing on steps 1-4. Examples for step 5 and partially step 6 can be found in [8]. Given that continuity and change follows other core themes that have been captured in the KB, we actually start from that as a rich biological ontology already. Hence, one concern in core theme design is the reuse of representations, where appropriate and practical. Next we sketch relevant aspects of the KB.

### 2.2 Component Library, KB Format and Content

The foundational component of the KB is the Component Library (CLIB), an upper ontology which is linguistically motivated and designed to support the representation of knowledge for automated reasoning [3]. CLIB adopts four simple top level distinctions: (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 [10]. It is useful to encode the mereological structure of events and entities as well as ways in which entities participate in different steps of an event.

The overall KB uses a representation based on first order logic. Most importantly, biologist encoders can edit existential rules [2] through a graphical user interface. Each graph captures a single rule, i.e., a sentence quantified universally in one variable and existentially in all others, cf. [7, 10]. All classes (unary predicates) are arranged in a taxonomy, a subsumption poly-hierarchy, the upper levels of which are constituted by CLIB. Turning to biological content, for many notions central to continuity and change the KB already contains representations from a structural and functional point of view, cf. [6].

### 3 RESULTS

#### 3.1 Definition of Continuity and Change

We synthesized three perspectives into a scoping definition of the core theme, see step 1 in section 2. The College Board syllabus [17] states: 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** [13, 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 that we have consulted with 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.

This input evolved in later steps to our proposed theme definition: 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.

#### 3.2 Informal Competency Questions

Competency questions are intended for validating proposed representations for a core theme. We distinguish diagnostic questions that are close to basic tasks and aim at testing representation adequacy, and educationally useful questions that are gathered by convening a focus group of teachers and students and serve as use cases in evaluation studies after a substantial amount of the core theme content in [13] will have been encoded in the KB.

Our suite of educationally useful questions on continuity and change has approximately 100 different questions. We provide four of them for illustration.

E1 What happens if crossing over occurs in the middle of a gene?
E2 What would happen to the chromosomes in eukaryotes if telomerase were lacking?
E3 What is the difference between a translocation and an inversion?
E4 Because of 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.

The diagnostic questions were abstracted from the educationally useful questions, thereby taking sets of diagnostic questions for other core themes into account. At present, these six schemes of diagnostic questions (with variables \( X \) and \( Y \) for events or entities) are available for checking modeling patterns:

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 \( Y \)?

While reported in some more detail elsewhere [8], we note that initial alternatives of representation patterns for continuity and change have been designed and checked against instances of these diagnostic questions. Sample processes under consideration include DNA replication, meiosis, and natural selection.

### 3.3 Structural Levels in Biology and Information Units

For the identification of key notions (step 3 in section 2) we have conducted a corpus analysis of Campbell Biology [13], gathering 169 terms for potential further analysis. While this set of terms includes entity as well as event types, we focus on entities (in the sense of CLIB) in the sequel.

Filtering key entity types from the set of terms is based on two aspects. First, genetic information in an understanding as sequences of nucleotide types is one central subject to continuity, complemented by phenotypic features as inherent expression of genetic information. Secondly, we adopt the structural biological levels of organization from [13], see Table 1, to organize what we abbreviate as information units, namely entities bearing genetic information (or phenotypes). For example, we consider ‘DNA strand’, understood as a polynucleotide molecule, as an information unit because it comprises a chain of nucleotide monomers, each of which instantiates a class of nucleotides that is associated with the genetic code, such as adenine. Due to that connection nucleotide classes can be perceived as types in the sense of the type-token distinction [18] and individual polynucleotide molecules are considered as tokens of sequences of (occurrences of) nucleotide types, cf. [18, sect. 8].

Table 2 displays the entities that we have identified, based on [13], as the main units bearing genetic information (or their phenotypic expression) at different levels of organization. The third column contains condensed characterizations of, partially multiple, readings of the respective terms.

| molecules | organisms |
| sub cellular entities | populations |
| cells | communities |
| tissues | ecosystems |
| organs and organ systems | the biosphere |

Table 1. Structural levels of biological organization, derived from [13]. Each term \( X \) stands for the respective “level of \( X \)”.

Table 2 as well as further analysis forms the basis for determining or checking the position of the key notions in the taxonomy of the KB. This fourth step in our procedure in section 2 links with established biomedical ontologies as a major kind of related work. Moreover, while most of the notions can already be found in KB-Bio, cf. its column in Table 3, we see room for further analysis and discuss related general representation problems in the next section.

### 4 DISCUSSION

#### 4.1 Information Units in Biomedical Ontologies

Regarding related work, we are not aware of any immediately similar approaches, neither of the overall project of establishing an electronic textbook for biology, nor of the topic of focus on continuity and change as a whole. However, there is a wide range of biomedical ontologies that cover various notions identified during corpus analysis, among them the information units identified in Table 2. In order to establish relations to KB-Bio, we have started an initial analysis of the coverage and representation of information units in biomedical ontologies.

Besides KB-Bio, Table 3 shows the taxonomic classification of gene, DNA, and DNA sequence in the Sequence Ontology (SO)\(^1\) [9, 12], BioTop\(^2\) [4], General Formal Ontology for Biology (GFO-Bio)\(^3\), Ontology of Genes and Genomes (OGG)\(^4\), Gene Regulation Ontology (GRO)\(^5\), National Cancer Institute Thesaurus (NCIt)\(^6\), and SNOMED-CT\(^7\), all retrieved via the NCBO BioPortal\(^8\). Those three terms are selected for their central role, polysemous use in [13], and good coverage in the selected ontologies.

Taking further terms into account, SO and NCIt apparently have very elaborate sets of classes with labels that are relevant for continuity and change. But clearly, one must pay attention to the background of each system to interpret its constituents correctly. For example, SO is rooted in genome annotation, such that special readings of the terms apply [12, p. 87], also visible in Table 3 in the upper-level classification in SO, cf. also [12, p. 88–89].

All shown ontologies have a term ‘DNA’, and all include ‘gene’, except for GFO-Bio (due to pending analysis). Across the ontologies, the relationship between ‘gene’ and ‘DNA’ or ‘nucleic acid’ varies strongly, looking at whether subsumption is declared at or which level ‘gene’ and ‘DNA’ have a common ancestor in the taxonomy. Frequently, ‘material object’ is among the common ancestors.

We can neither provide an in-depth comparison of these existing systems here, nor could we argue in detail for one or another solution without taking the systems much more into account. However,

---

\(^1\) http://www.sequenceontology.org/ (Jan 29, 2014)
\(^2\) http://www.imbi.uni-freiburg.de/ontology/biotop/ (May 20, 2014)
\(^3\) http://www.onto-med.de/ontologies/gfo-bio/ (Oct 11, 2010)
\(^4\) http://ogg.googlecode.com/ (May 22, 2014)
\(^6\) http://sitools.nema.org/ (Feb 21, 2014)
\(^7\) http://IDDLE2014.org/ (May 30, 2014)
\(^8\) http://bioportal.bioontology.org
the noted variety can be complemented with further observations, indicating more general problems.

### 4.2 Analysis and Representation Problems

Dwelling a little further on ‘gene’, one can easily observe its highly polysemous use in the literature, even just in a summary section in Campbell Biology [13, ch. 17.6]. The latter section suggests at least six distinct views on ‘gene’, among them being DNA, a DNA sequence, a region of DNA, or a locus at a chromosome. Note that we agree with [9, p. 1]: there is less or even no question of what is right or wrong for “the” term ‘gene’. But there is the analytical matter of drawing distinctions, utilizing classes in ontologies correspondingly, and accepting implications of such distinctions. For example, we find it difficult to combine ‘gene’ as subsumed by ‘material object’ with phrases like “the same gene is found in two cells” or “two alleles of the same gene”. The latter case remains problematic if instead ‘gene’ is viewed to be subsumed by ‘DNA sequence’.

<table>
<thead>
<tr>
<th>Levels referred to</th>
<th>Information Units</th>
<th>Description with Information or Inheritance Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecules</td>
<td>nucleotide</td>
<td>A nucleotide is a single molecule/monomer that corresponds to single item/“letter” in an xNA sequence due to its kind (e.g., the well-known types of adenine (A), cytosine (C), guanine (G), and thymine (T)).</td>
</tr>
<tr>
<td>molecules</td>
<td>codon, anti-codon</td>
<td>A(n anti)-codon is a nucleotide triplet that encodes amino acids, i.e., if involved in a particular transcription, a codon is transcribed into a single amino acid. Codons of the same kind may encode/be transcribed into multiple amino acids, however. Start and stop codons do not encode amino acids, but they initiate or terminate the transcription process, respectively.</td>
</tr>
<tr>
<td>molecules</td>
<td>xNA, xNA strand</td>
<td>A ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) molecule is a polymeric molecule, whose primary structure is in most cases linear/unbranched. In these cases it encodes/can be represented by a sequence of the types of its individual nucleotides. At mesoscopic granularity, there is a mass-term view on DNA, i.e., DNA is a substance that can be present in a certain amount.</td>
</tr>
<tr>
<td>molecules</td>
<td>gene, allele</td>
<td>The term gene is highly polysemous, see section 4. An allele (as a sequence S of nucleotide types) is a specific variant of a gene (as a set of possible alternative sequences of nucleotide types), determined as a manifestation of that gene by the location of the allele (as the part of a DNA molecule that encodes/embodies the sequence S) at a chromosome. Proteins are one kind of entity frequently associated with/said to be encoded by individual genes/alleles, in that an individual process of gene expression leads from “reading” a gene (as a part of a DNA strand/molecule) to the synthesis of an individual protein. Thus, the gene (as a sequence of nucleotide types) encodes the specific kind of protein. Another kind of entity that genes are linked with are heritable traits/phenotypic features, depending on the strength of the correlation (and possibly experimentally provable causation) between the presence of one or more genes (as sequences of nucleotide types) in the genome of an organism and the respective traits of that organism.</td>
</tr>
<tr>
<td>sub cellular entities</td>
<td>chromosome, chromatid</td>
<td>A chromosome is a sub cellular entity that contains usually a single DNA molecule, which itself comprises various genes (as parts of the overall molecule in the chromosome). From the information perspective, a chromosome merely packages other information units (genes, xNA), without encoding an (independently defined) information entity corresponding to the overall chromosome (or its DNA molecule) itself.</td>
</tr>
<tr>
<td>cell, organism</td>
<td>genome, genotype</td>
<td>The genome (of a cell) is usually understood as the overall genetic material within a cell, where genetic material is typically equated with “amount of DNA”, in a substance-like understanding of DNA rather than referring to individual DNA molecules. Analogously, one may refer to the genome of an organism. This usually involves the idealization that all cells of an organism comprise exact copies of a single (cellular) genome. The notion of genotype also refers to the overall genetic information within a cell (or an organism, under the same assumption as just indicated), yet under a Mendelian perspective in terms of the specific alleles of all genes available in the cell (or organism). For both, genome and genotype, the “semantics” of their genetic information corresponds to having a “phenotypic blueprint” for the overall organism.</td>
</tr>
<tr>
<td>organism</td>
<td>organism</td>
<td>Organisms themselves can be considered as bearers of genetic information that they inherit from generation to generation. In particular, considering several organisms (of the same or different species), they exhibit differences in their genomes. Depending on the focus of inquiry, this relates to resemblances or differences in terms of specific phenotypic traits or more complex phenotypic aspects, up to the overall phenotype of an organism as an “expression” / the “realized meaning” of its genetic information.</td>
</tr>
<tr>
<td>population</td>
<td>gene pool</td>
<td>The notion of gene pool accounts for an abstraction or aggregation of all alleles that occur in at least one individual within a population. Due to the link between alleles and phenotypic traits in organisms, a gene pool determines a set of possible phenotypes that may be expressed within a population.</td>
</tr>
</tbody>
</table>

Table 2. Information units, relevant structural levels of Table 1, and a characterization of their relation to genetic information or its phenotypic expression, primarily based on [13].
From an analytical point of view, we agree in general with [11] that a role-based account of ‘gene’ is adequate (which applies to various other notions, as well). Strong support for that is given by (especially same-strand) *overlapping genes*, which are observed in genomes of various species including human [14], i.e., (with the molecular view on ‘gene’) parts of DNA that are read and translated in different transcription processes. However, role-based analysis usually comes at the price of further discrimination. Distinctions become additionally multiplied by different readings of other terms, cf. e.g. the four views on DNA discussed in [15, p. 112]. Accordingly, ‘gene’ may be related with double-stranded or single-stranded DNA. On the other hand, one may wish to be agnostic about that distinction unless explicitly required.

Problems of this kind have surely been noted earlier. Usually, a trade-off needs to be found between fine-grained distinctions and maintaining comprehensibility and tractability. But large systems such as NCI and SO that are applied in multiple contexts start to incorporate additional distinctions, cf. the SO:Molecules subset [12, p. 87]. Similarly, we expect that answering a variety of questions from KB-Bio will require further distinctions of the key entities in continuity and change.

At the present stage, we see two major interrelated challenges regarding ontological analysis and representation in general (also beyond the biomedical domain). (1) There is the need to analyze “informal notions” from multiple perspectives, but in an integrated way. We expect this to lead to various, ontologically distinguishable core notions of continuity and change.

Problems of this kind have surely been noted earlier. Usually, a trade-off needs to be found between fine-grained distinctions and maintaining comprehensibility and tractability. But large systems such as NCI and SO that are applied in multiple contexts start to incorporate additional distinctions, cf. the SO:Molecules subset [12, p. 87]. Similarly, we expect that answering a variety of questions from KB-Bio will require further distinctions of the key entities in continuity and change.

At the present stage, we see two major interrelated challenges regarding ontological analysis and representation in general (also beyond the biomedical domain). (1) There is the need to analyze “informal notions” from multiple perspectives, but in an integrated way. We expect this to lead to various, ontologically distinguishable core notions of continuity and change.
5 CONCLUSION

Continuity and change in genetic information is a rich topic area in biology. This paper presents some of our preliminary results in designing representations for this core theme for KB-Bio, a knowledge base for answering questions within an electronic biology textbook. We contribute a scoping definition for the theme and outline two types of competency questions for evaluating representations of the theme. Moreover, key entity types in this area are systematically identified and briefly characterized as information units, based on the set of levels of biological organization in Campbell Biology [13].

Clearly, we have only scratched the surface of the topic, leaving much work for the future. Towards a more detailed analysis and a revised classification of the key entity types, a comparison chart of gene, DNA, and DNA sequence in biomedical ontologies is briefly discussed. This confirms our impressions from developing first representations of events [8] that capturing continuity and change would benefit from advancing ontology representation in general, e.g. to support multiple, integrated views on biological terms, and to work efficiently with ontological distinctions drawn among readings of polysemous terms.

ACKNOWLEDGEMENT

This 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 ideas in this paper.

REFERENCES


