Applying the Principles of an Ontology-Based Approach to a Conceptual Schema of Human Genome

Ana Mª Martínez Ferrandis[1], Oscar Pastor López[1], and Giancarlo Guizzardi[2]

[1] Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València, Spain
{amartinez,opastor}@dsic.upv.es
[2] Ontology and Conceptual Modeling Research Group,
Federal University of Espirito Santo, Brazil
gguizzardi@inf.ufes.br

Abstract. Understanding the Human Genome is currently a significant challenge. Having a Conceptual Schema of Human Genome (CSHG) is in this context a first step to link a sound Information Systems Design approach with Bioinformatics. But this is not enough. The use of an adequate ontological commitment is essential to fix the real-world semantics of the analyzed domain. Starting from a concrete proposal for CSHG, the main goal of this paper is to apply the principles of a foundational ontology, as it is UFO, to make explicit the ontological commitments underlying the concepts represented in the Conceptual Schema. As demonstrated in the paper, this ontological analysis is also able to highlight some conceptual drawbacks present in the initial version of the CSHG.

Keywords: Ontology, UFO, Conceptual Modeling, Information Systems, Bioinformatics.

1 Introduction

Genomics is one of the most interesting research areas of the Bioinformatics field. Understanding the Human Genome is currently a significant research challenge but with far reaching implications such as to provide answers to questions like what the concepts that explain ours essential characteristics as species are, or how to prevent disease within a personalized medicine context. Given the large amount of data involved in such as task, and the need to structure, store and manage this data correctly, the application of sound conceptual modeling principles is made necessary. In fact, the most remarkable properties of the genomic field research are the tremendous quantity of data available, its dispersion and the continuous evolution of the involved concepts. Thus, having a Conceptual Schema of the Human Genome (CSHG) is in this context a first step to link a sound Information Systems Design approach with Bioinformatics. Moreover, given the complexity of the involved notions as well the clear need for autonomous data interoperability, it is essential that these notions are well understood and that their underlying real-world semantics are made explicit.

W. Ng, V.C. Storey, and J. Trujillo (Eds.): ER 2013, LNCS 8217, pp. 471–478, 2013.

In this paper, we start from a concrete proposal for a CSHG [1] and illustrate how a foundational ontology (UFO) [2] can be used to make explicit the ontological commitments underlying the concepts that are represented in the Conceptual Schema. The benefits of such an approach are twofold. On one side, we can improve consistency and understandability of the CSHG through *conceptual clarification*. On the other side, this approach can identify a number of conceptual drawbacks present in the initial version of the quoted CSHG that have been put in evidence and corrected.

In the field of Biology, ontologies are often used as repositories of data, vocabularies, taxonomies, etc. A well-known, relevant example is the Gene Ontology [3]. Nevertheless, in contrast with the Gene Ontology, we here strongly advocate the use of foundational ontologies to characterize the real-world semantics that are used in the specification of conceptual genomic models. In this spirit, the work presented here is very much in line with approaches such as in [4], which promote the use of foundational ontologies to avoid errors in the curation and creation of domain models in the biomedical field. However, we here take one step forward from a conceptual modeling point of view, namely, we show how the benefits of using these foundational theories can be systematically carried out to conceptual modeling by employing an ontologically well-founded conceptual modeling language (OntoUML) [2].

The remainder of this article is organized as follows. In Section 2, we briefly present the foundation ontology UFO and its relation to the OntoUML conceptual modeling language. Section 3 explains the use of Conceptual Models in the specification of the Genomics Domain, starting with the Conceptual Schema of the Human Genome (CSHG). In Section 4, we present the main contribution of this paper, namely, the ontological analysis of the CSHG using the approach introduced in section 2. Section 5 presents some final considerations.

2 OntoUML as Tool for an Ontological Analysis of the CSHG

In recent years, there has been a growing interest in the application of Foundational Ontologies, i.e., formal ontological theories in the philosophical sense, for providing real-world semantics for conceptual modeling languages, and theoretically sound foundations and methodological guidelines for evaluating and improving the individual models produced using these languages. OntoUML [2] is an example of a conceptual modeling language whose metamodel has been designed to comply with the ontological distinctions and axiomatic theories put forth by a theoretically well-grounded Foundational Ontology [5]. This language has been successfully employed in a number of projects in several different domains including Heart Electrophysiology, Petroleum and Gas, Software Engineering, News Information Management, among many others. Asides from the language itself defined with an explicit metamodel embedded with ontological constraints, the OntoUML approach includes a number of ontology-based patterns and anti-patterns (modeling patterns, analysis patterns, transformation patterns and validation anti-patterns) as well as a number of automated tools for model construction, verification, validation, verbalization and code generation.

In section 4, we introduce some of the OntoUML modeling constructs and briefly elaborate on their ontological semantics as defined in UFO. For a fuller presentation of UFO and OntoUML, containing philosophical justification, empirical support and formal characterization, one should refer to [2,5]. We focus the remainder of this paper to illustrate with a preliminary practical exercise how OntoUML can be used to support an ontological analysis of a particular Conceptual Schema of the Human Genome, making explicit its ontological commitments, fixing a particular real-world semantics for its constructs as well as identifying conceptual problems in terms of uncertainty, inconsistencies, lack of constraints and dubious modeling choices.

3 Conceptual Schema of Human Genome (CSHG)

This section briefly elaborates on the second fundamental component for the analysis presented in this paper, namely, the Conceptual Schema of the Human Genome (CSHG) [1]. This conceptual schema was produced as a result of the Human Genome project developed by Genome Research Group of the "Centro de Investigación en Métodos de Producción de Software (ProS") of the Universitat Politècnica de València. This group is an interdisciplinary group consisting of experts both in the field of genomics and computer science whose main goal is to clearly specify and represent the genomic domain.

The CSHG consists in four different views, namely, the Variation View, the Phenotypic View, the Transcription View, and the Genome View. In the present article, due to space limitations, we focus on an excerpt of the variation view. This view comprises the description of the variations that are found on a gene. Details about this Variation View can be found in [6]. Hereafter, only the needed fragments of the CSHG that were required to understand the analyzed concepts are shown.

4 Discussion and Results

In this section we elaborate on some of the outcomes of our analysis. Due to space limitations, in this section, we restrict our discussion to fragments of the redesigned CSHG. A fuller presentation of the complete ontological analysis and redesign of the original conceptual schema will be presented in a subsequent publication. It is important to also highlight the fact that the generated OntoUML model makes explicit the particular ontological commitments underlying the CSHG as conceived by its creators. Although these particular commitments can be debated, the reason they can be so is exactly because they no longer remain tacit in the creator´s minds.

The human genome is the entire genetic information that a particular individual organism has and that encodes it. It is formed by the set of all the chromosomes on the DNA. In the same manner, chromosomes are formed by a set of genes, which is an ordered sequence of nucleotides in the DNA molecule and contains the information needed for the synthesis of a macromolecule with specific cellular function.

OntoUML makes a fundamental distinction between three different types of substantial entities depending on their unity criteria and the relation they have with their

parts. Here, we focus on two of these distinctions, namely, Functional Complexes and Collectives [2,7]. A collective is an entity characterized by the fact that all its constituent parts instantiate the same type and play the same role w.r.t. the whole (e.g., a forest or a crowd). In contrast, the different parts of a functional complex X are of different types and play different roles w.r.t. to X. Examples of the latter include a human body, a computer, an organization, a TV set. In OntoUML, identity-providing rigid types whose instances are collectives receive the homonymous stereotype; identity-providing rigid types whose instances are functional complexes are stereotyped with the word «kind».

When analyzing the core concepts of the CSHG in light of these distinctions, we can see that allele, gene and chromosome can be seen as collectives and nucleotide as a functional complex (Fig. 1). These distinctions between types used in OntoUML make explicit additional information about the nature of each type. This, in turn, prevents an unwarranted interpretation that nucleotide and gene are of the same ontological nature. By making explicit the ontological nature of the entities, we can also systematically make explicit the different types of parthood relations involving these entities and their respective parts. In Fig. 1b, we have that nucleotides are essential parts of a specific Allele, i.e., besides the relation of parthood, there is an existential dependence relation between an Allele and each of its constituent nucleotides. In other words, a specific Allele can only exist (preserving the same identity) by having each of these nucleotides as parts. In fact, the identity of an allele is defined by the sum and position (sequence) of its parts.

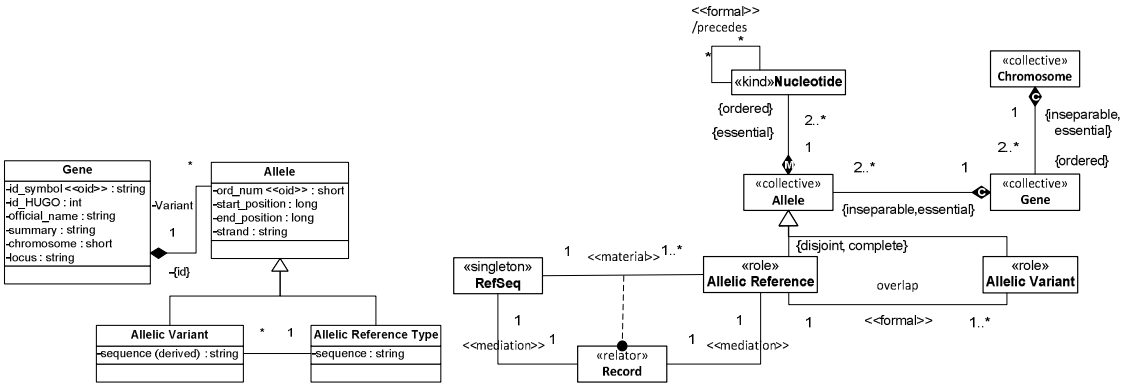

Fig. 1. (a-left) A fragment of the CSHG and **(b-right)** its counterpart in OntoUML (core concepts of the variation view)

In the previous version of the CSHG (Fig. 1a), the notion of nucleotide was not included. Instead, the attribute *sequence* was used to express this idea of collective of nucleotides. The existential dependence from an allele to a set of nucleotides represents explicitly the relation between the identity criteria of an allele and the ordered sequence of its constituent nucleotides. In the original model, this identity criterion was artificially represented by an assigned object identifier.

Making explicit the different types of entities in the OntoUML model clearly sets out the different types of part-whole relations involving them. OntoUML prescribes

four different types of part-whole relations: *subQuantityOf* (defined between quantities), *memberOf* (defined by individuals and the collectives they compose), *subCollectiveOf* (defined between collectives) and *componentOf* (defined between functional complexes and their parts). As demonstrated in [2], each of these different types of part-whole relations is correlated with different types of meta-properties regarding existential dependence, transitivity, shareability, among others.

Still regarding part-whole relations, Fig. 1b models that there is a mutual existential dependence between a Gene and its constituent Alleles, i.e., an Allele must be part of particular Gene and a Gene must be composed of that specific set of Alleles. Analogously, an individual gene must be part of a specific chromosome and a chromosome must be composed of that specific set of genes in every situation that it exists. Finally, in OntoUML, we have that a memberOf relation is never transitive, but also that subCollectiveOf relations are transitive [7, 8]. For this reason, in the model of Fig. 1b, we have that ultimately a chromosome can be seen as an ordered sequence of nucleotides. However, we also have that none of the parts of a nucleotide are parts of an Allele, of a Gene or of a Chromosome. In the previous version of the CSHG, all the aforementioned information remained tacit in the modeler's mind.

Alleles are specialized as *Allelic Variant* and *Allelic Reference Type*. The last is an allele that works as a stable foundation for reporting mutations, in the sense that all the alleles that are different from it (but still related to the same gene) would be classified as *Allelic Variant* and those differences would be reported as genetic variations. But what makes an allele be considered an allele of reference? The RefSeq project [9] defines the alleles to be used as standards for well-characterized genes. So, an allele becomes an *Allelic Reference* if there is a record in RefSeq for this allele. Indirectly, this record also makes the remaining alleles from a gene an *Allelic Variant*. In OntoUML, both *Allelic Reference* and *Allelic Variant* are considered types of *Role*. A role is an anti-rigid type (i.e., a type describing contingent properties of its instances) and a relationally dependent one (i.e., a type defined in terms of a relational condition) [2]. In Fig. 1, *Allelic Reference* is a role (contingently) played by an allele when referred by (*related to*) a record in RefSeq. Moreover, an *Allelic Variant* is a role played by an allele when related to the same gene as an *Allelic Reference*. Finally, an entity like *record* in Fig. 1 is modeled in OntoUML by using the notion of a *relator*. A relator is the objectification of a relational property and represents the so-called *truthmaker* of a material relation [2]. So, for instance, in the same way that an entity such as a particular marriage (a particular bundle of commitments and claims) is the truthmaker of the relation *is-married-to* between the individuals John and Mary, the presence of a RefSeq record represents here a binding between RefSeq and a particular allele, thus, making true the relation between an allele playing the role of *Allelic Reference* and RefSeq. The relations of mediation between the presence of a RefSeq record, RefSeq and the corresponding Allelic Reference (Fig. 2) are relations of existential dependence (the presence of this record depends on RefSeq and on the Allelic Reference) and constitutes the aforementioned binding.

A genetic variation is described as a difference between an *Allelic Variant* and its *Allelic Reference*. So, variations are *fiat entities* but which are existentially dependent on a particular allelic reference and a particular allelic variant. As referable entities

which are existentially dependent on multiple entities, a variation is also represented here as a relator (Fig. 2). In other words, a variation is constituted by a number of nucleotides which are part of the *Allelic Variant* and which vary *in relation to* the *Allelic Reference*. This analysis reveals a conceptual mistake in the previous version of the CSHG: since the sequence of the *Allelic Variant* is modeled as derived by the application of the variations that relate it with its *Allelic Reference Type*, the *Allelic Variant* would be characterized as existentially dependent on the variations and not the other way around. Notice that, since all parts of an allele are essential to it, we have that an *Allelic Variant* is indeed also existentially dependent on the nucleotides that constitute a variant. The notion of variant itself, however, is a relational notion that depends on both the Allelic Reference and Allelic Variant.

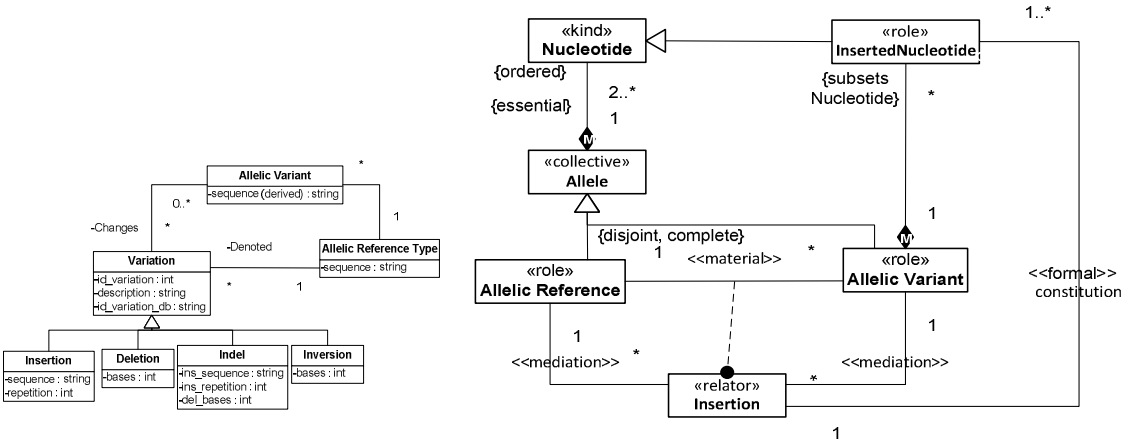

Fig. 2. (a-left) A fragment of the CSHG and **(b-right)** its counterpart in OntoUML (insertion as a type of variation)

Genetic variations can be further characterized depending on their type: insertions, deletions, indels and inversions. In the CSHG, they are simply represented as subtypes of the *Variation* class. This incompleteness in the model, however, leaves implicit the fact that each of these variations is derived from different types of *base relations*. In the redesigned model, we use the OntoUML relator construct to represent explicitly each category of variation with its characterizing relations. The example that we show in Fig. 2 is the case of *insertions*. In the case of insertions, we have that the nucleotides that *constitute* an insertion are parts of the allelic variant mediated by this variation (i.e., of which this variation depends). In the model of Fig. 2, these nucleotides are said to play the role of inserted nucleotides w.r.t. the allelic variant. Moreover, the part of relation between the former and the latter is explicitly represented in that model. Since an Allelic Variant is a collective, this is an example of a *memberOf* relation [7, 8]. This model should also include a constraint that the nucleotides that play the role of *Inserted Nucleotides* w.r.t. an *Allelic Variant* must *constitute* the insertion which mediates that Allelic Variant. Moreover, the nucleotides which play the role of Inserted Nucleotides in an insertion and which are members of an Allelic Variant are necessarily member of that specific allele playing the role of Allelic Variant. This inclusion constraint is represented via association

subseting in Fig. 2b. The need for this constraint can be automatically detected in an OntoUML model since its absence would include in the model an instance of a predefined validation OntoUML anti-pattern [10].

Finally, in the human genome, there is also the notion of *conservative regions*, which are regions that have been in the genome for ages without alteration and which are expected to remain the same in the allelic reference and its variants. We use here a formal relation from the mereological theories underlying OntoUML to model that there exists a relation of *(non-proper) overlapping* between an *Allelic Reference* and its *Allelic Variant* (Fig. 1 and Fig. 2). In other words, Allelic Reference and Allelic Variant must share a common part. If there is no overlapping between sequences, then the two alleles belong to different genes. This constraint is of significance when talking about the nature of the alleles and genes, another feature which remains implicit in the previous version of the CSHG.

5 Conclusion

In this paper, we start from a concrete proposal for a Conceptual Schema for the Human Genome and illustrate how a principled ontological analysis can be used to make explicit the ontological commitments underlying the concepts that are represented in that schema. Moreover, the paper illustrates an approach in which this ontological analysis is performed systematically and is integrated in a classical conceptual modeling engineering activity throughout the use of an ontologically well-founded conceptual modeling language termed OntoUML as well as its associated methodological tools.

In the redesign of the CSHG as an OntoUML model, a number of implicit assumptions in the original model were made explicit as well as a number of conceptual drawbacks were identified. For instance, the introduction of the RefSeq and the Record was instrumental for expressing that Allelic Reference and Allelic Variant are contingent roles played by an allele in relational contexts: in order for an allele to be an allelic reference it must be referred by a RefSeq record; in order for an allele to be an allelic variant it must be related to the same gene and non-properly overlap with an allelic reference. Moreover, the use of the formal relation of non-proper overlapping between an Allelic Reference and its Allelic Variant represents the conservative regions on the DNA, making explicit the constraint of the non-existence of "extreme variations" in the domain. Furthermore, modeling the Variation as a relator also expresses its existential dependency on a specific Allelic Variant and on a specific an Allelic Reference. This highlights the doubtful choice of considering the sequence of the Allelic Variant as derived by the application of the variations that relate it with its Allelic Reference Type. Finally, the use of the mereological relations of memberOf and subcollectiveOf to represent parthood between concepts such as Gene, Nucleotide, Allele and Chromosome makes explicit the notion of a chromosome as an ordered sequence of nucleotides.

The analysis presented here concentrates on a small fragment of the Variation view of CSHG. In a future work, we shall present a full analysis of CSHG contemplating

all its constituent views. Once we have a complete OntoUML version of the CSHG, we pretend to conduct a full validation with domain experts by using the OntoUML approach of model validation via visual simulation [10]. Finally, after validation, we intend to use the OntoUML tool set to automatically generate OWL specifications for the CSHG. These specifications, in turn, will be employed to support semantic annotation and automated reasoning in a Human Genome Wiki environment.

Acknowledgements: This work has been developed with the support of MICINN under the project PROS-Req TIN2010-19130-C02-02 and the Programa de Apoyo a la Investigación y Desarrollo (PAID-00-12) de la Universitat Politècnica de València, and co-financed with ERDF. The third author has been supported by FAPES (PRONEX Grant #52272362/2011).

References

1. Pastor, O., Levin, A.M., Casamayor, J.C., Celma, M., Eraso, L.E., Villanueva, M.J., Perez-Alonso, M.: Enforcing conceptual modeling to improve the understanding of human genome. In: 2010 Fourth International Conference on Research Challenges in Information Science (RCIS), pp. 85–92. IEEE (2010)
2. Guizzardi, G.: Ontological foundations for structural conceptual models. CTIT, Centre for Telematics and Information Technology (2005)
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.: Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
4. Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H., Kelso, J.: A top-level ontology of functions and its application in the Open Biomedical Ontologies. Bioinformatics 22, e66–e73 (2006)
5. Guizzardi, G., Wagner, G.: Using the Unified Foundational Ontology (UFO) as a foundation for general conceptual modeling languages. In: Theory and Applications of Ontology: Computer Applications, pp. 175–196. Springer (2010)
6. Pastor, O., Levin, A.M., Casamayor, J.C., Celma, M., Kroon, M.: A Conceptual Modeling Approach to Improve Human Genome Understanding. In: Handbook of Conceptual Modeling, pp. 517–541. Springer (2011)
7. Guizzardi, G.: Ontological foundations for conceptual part-whole relations: the case of collectives and their parts. In: Mouratidis, H., Rolland, C. (eds.) CAiSE 2011. LNCS, vol. 6741, pp. 138–153. Springer, Heidelberg (2011)
8. Guizzardi, G.: Representing collectives and their members in UML conceptual models: an ontological analysis. In: Trujillo, J., et al. (eds.) ER 2010. LNCS, vol. 6413, pp. 265–274. Springer, Heidelberg (2010)
9. QPruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35, D61–D65 (2007)
10. Sales, T.P., Barcelos, P.P.F., Guizzardi, G.: Identification of Semantic Anti-Patterns in Ontology-Driven Conceptual Modeling via Visual Simulation. In: 4th International Workshop on Ontology-Driven Information Systems (ODISE 2012), Graz, Austria (2012)