An Ontological Characterization of a Conceptual Model of the Human Genome

Alberto García S.[1][0000−0001−5910−4363], Giancarlo Guizzardi[2,3][0000−0002−3452−553X], Oscar Pastor[1][0000−0002−1320−8471], Veda C. Storey[4][0000−0002−8735−1553], and Anna Bernasconi[1,5][0000−0001−8016−5750]

[1] Universitat Politècnica de València, Valencia, Spain
{algarsi3,opastor}@pros.upv.es, abernas@upvnet.upv.es
[2] Free University of Bozen-Bolzano, Bolzano, Italy Giancarlo.Guizzardi@unibz.it
[3] University of Twente, Twente, The Netherlands g.guizzardi@utwente.nl
[4] Georgia State University, Atlanta, Georgia USA vstorey@gsu.edu
[5] Politecnico di Milano, Milan, Italy anna.bernasconi@polimi.it

Abstract. The ability to sequence the human genome is a scientific, historical breakthrough. Although the human genome mapping is available to all scientists, information about it can be difficult to share. The Conceptual Schema of the Human Genome represents the concepts required to holistically understand the human genome. We report on our continued efforts to ensure that the human genome can be meaningfully shared by conducting an ontological analysis and enrichment of the conceptual model to facilitate domain understanding and data exchange among heterogeneous systems. The analysis and enrichment process is supported by the ontology-driven conceptual modeling language, OntoUML, to gain ontological clarity and demonstrated on a relevant section of the Pathways view of the schema. Consistent with the overall objective of designing a sound genomics information system, the results lead to major modeling implications for the: characterization of biological entities; changes in biological entities over time; and representation of chemical compounds. Our research shows that the inclusion of a strong ontological foundation in a conceptual model contributes to the design of complex systems.

Keywords: Ontological Analysis · Conceptual Schema of the Human Genome · Ontological Foundation · Heterogeneous Data Representation

1 Introduction

The modeling of the human genome is a fascinating and extremely important area of research due to its potential to impact all of mankind through improved treatments and possibly, removal of diseases. In essence, this modeling is contributing to understanding life itself. Progressing research on the human genome, however, is challenged for many reasons, perhaps the greatest of which is the fact that the body of knowledge surrounding the human genome constantly changes

and evolves as scientists and researchers all over the world work with it. Furthermore, the terminology and concepts employed in genomics can be imprecise and change continuously. So does the scope and complexity of the modeling required to represent them. The definitions of terms needed to characterize any phenomena rely on the experience of the domain experts who use and interpret them. Definitions may be purposely abstract to reflect the limited, and constantly changing, knowledge of the domain. They cannot simply be translated into an unambiguous representation of knowledge. However, a fundamental prerequisite for analyzing and understanding any complex domain, is to facilitate a shared understanding among the people who work in that domain.

The most common artifacts used for representing concepts are lightweight ontologies (logical specifications in a form of Description Logics) and thesauruses of controlled vocabulary [15], because they provide standard concepts and definitions. However, they can only correctly represent a minor portion of relevant facts in genomics [4]. Conceptual models are appropriate because they facilitate the exchange of information [13], while providing a sound basis to make a conceptualization process explicit and facilitate a shared understanding of a domain [11]. For the human genome domain, applying conceptual modeling can: improve communication among physicians, geneticists, biologists, and other researchers; assist in knowledge transfer; and, ultimately, enable efficient exploitation of information for progressing the understanding of the human genome [12].

Prior research has created a Conceptual Schema of the Human Genome (CSHG) [2]. This research extends the conceptual model by making the definition of the relevant concepts precise, explicit and understandable. We conduct an ontological analysis and enrichment of the Pathways view of the current model. We use "ontological" in a strong sense. Our analysis aims at revealing and explicitly modeling aspects related to the *nature* and *real-world semantics* of entity types and relations by employing the conceptual modeling language OntoUML [5], which is grounded in the Unified Foundational Ontology (UFO) [7]. The contribution is to reformulate the conceptual model, showing how a foundational ontology brings ontological clarity to complex models by facilitating domain understanding and data exchange among heterogeneous systems [6].

2 Conceptual Schema of the Genome (CSHG)

The Conceptual Schema of the Human Genome (CSHG) [3] focused on representing the most relevant concepts of genomics. Creating this holistic schema required the integration of conceptual components that represent the relevant data that connect the genome structure (genotype) with its expression of real world behavior (phenotype). Evolution of the CSHG resulted in five views: structural, variation, transcription, proteome, bibliography, and pathway. Using conceptual views allows us to focus on specific dimensions of interest, which have many practical uses such as identifying and managing genomic variations related to the treatment of Alzheimer's [9], developing a conceptual model-based frame-

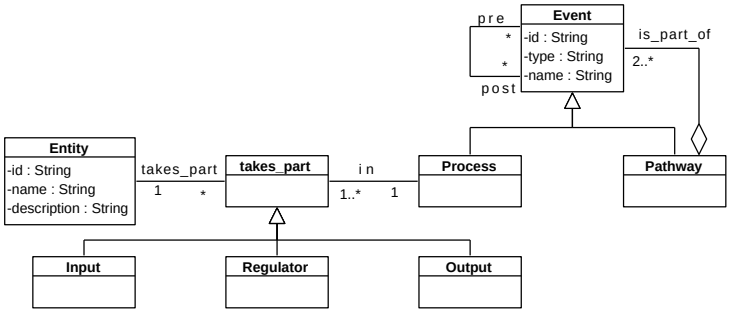

Fig. 1. Subset of selected classes from the Pathways view of the CSHG

work to improve data quality processes for precision medicine [10], of reporting early diagnosis of alcohol sensitivity [14].

Here, we focus on the Pathways view, which describes the chemical reactions that explain the different molecular processes. This view reflects very critical aspects of the genomics domain, including a *biological event* that addresses how genome elements interact to produce a biological behavior. Given its importance and richness, this view provides an appropriate way to motivate and demonstrate the need for the type of analysis and redesign proposed here. A portion of the Pathways view is depicted in Fig. 1 as a UML class diagram. The model is centered on the notions of entity and event, represented by homonymous classes.

Consistent with the terminology of geneticists, an entity class identifies any possible physical component present in a body and plays a role in an event. In turn, an event class represents the biological events that occur in a body. Events are recursively composed of additional events. An event can be a pathway (complex event, made up of other events) or a process (elementary event). A process is then an atomic, simple event of a given type. A pathway is a more complex type of event that is decomposed into a specific set of events (processes or pathways). A process is a specific interaction between entities. An entity can participate as an input, an output, or a regulator. These associated sets of inputs, outputs, and (optionally) regulators characterize the process functionality. When an entity takes part in a specific process, it assumes one of these three roles.

3 OntoUML

OntoUML is an ontology-driven conceptual modeling language based on the upper ontology Unified Foundational Ontology (UFO, [7,5]). OntoUML uses stereotypes to represent the mapping between its modeling constructs and UFO ontological categories. OntoUML is built upon the fundamental distinction between Types and Individuals. Types are patterns of features that are repeatable across multiple instances. OntoUML includes a theory of higher-order types so first-order types are types instantiated by individuals, whereas higher-order

types (represented by the stereotype TYPE) are instantiated by other types. UFO countenances two fundamental types of individuals: endurants (objects and their existentially dependent reified aspects) and perdurants (events and processes).

Endurants types are classified on two dimensions, sortality (identity) and rigidity. Sortals are types whose instances obey a single identity principle (all of the same KIND); non-sortals are types that classify instances of multiple kinds. A type is rigid if it defines essential characteristics of its instances; anti-rigid if it defines contingent characteristics for all instances. The type person is rigid, but student is anti-rigid. Kinds represent the genuine fundamental types of objects that exist according to a particular conceptualization of a domain. All objects belong to exactly one kind. There can be other static specializations of a kind, namely SUBKINDS; e.g., the kind "gene product" can be specialized into the subkinds "coding RNA" and "non-coding RNA".

Objects can be classified depending on their principle of unity, i.e., the principle binding the parts that form a whole. For example, they can be COLLECTIVES if they are composed of parts (termed *members*) that play the same role with respect to the whole, or FUNCTIONAL COMPLEXES if they are composed of parts (termed *components*) that play different roles with respect to the whole. Finally, objects can be QUANTITIES to represent homeomerous entities (i.e., entities repeatably decomposable into entities of the same kind), such as water, sand, or blood. Since most of the kinds in a domain are those whose instances are functional complexes, we use the stereotype KIND simply to represent them.

Anti-Rigid types are specialized into PHASES and ROLES. Both phases and roles are dynamic types. Phases have intrinsic dynamic classification conditions, i.e., they capture a cluster of change conditions in intrinsic properties. Roles, in contrast, have relational dynamic classification conditions, i.e., they capture a cluster of change conditions bound to changes in a relational context. For instance, a blood cell has multiple phases such as blood stem cell, red blood cell, etc. depending on its maturity (i.e., an intrinsic property). In the case of roles, a person (i.e., an instance of the kind person) can be a patient (role) while participating in a medical treatment.

Phases and Roles are sortals (classify things of the same kind). We can, however, have analogous anti-rigid non-sortal classes, namely, PHASEMIXINS and ROLEMIXINS. As non-sortals, phaseMixins and roleMixins classify instances of multiple kinds. For instance, suppose a protein (kind) and an organic chemical compound (kind) play the role of a regulator in a specific biological process. There are two different roles: the "regulator protein" and the "regulator chemical compound". Both regulate a process so we can abstract them into a new roleMixin, called regulator, from which the other two roles specialize. PhaseMixins and RoleMixins can be thought as refactoring classes (abstracting properties common to entities of multiple kinds) and, hence, they are always *abstract* types (i.e., types that cannot be directly instantiated). We can have refactoring (non-sortal) types that are rigid, i.e., that abstract *essential* properties common to entities of several kinds. These are marked as the CATEGORY stereotype.

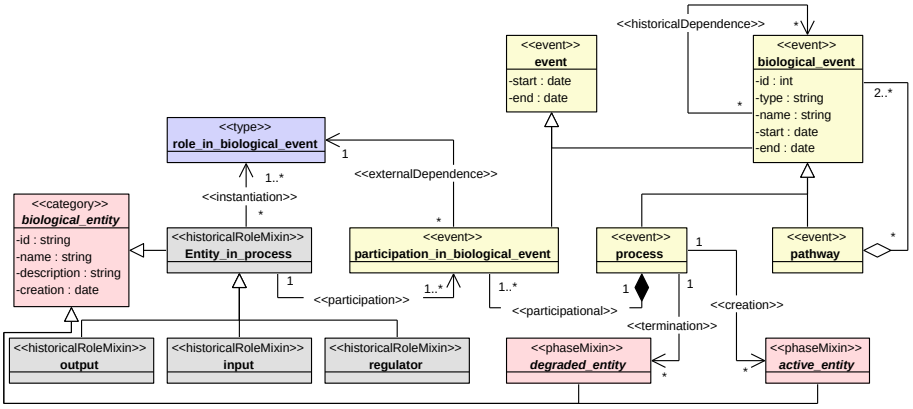

Fig. 2. Ontologically enriched version of the analyzed concepts.

Objects bear a number of *aspects*, some of which are intrinsic (existentially depend solely on them): QUALITIES or MODES. Qualities can be directly associated with structured value spaces (e.g., color or temperature); MODES are full-fledged object-like entities with their own aspects but still existentially dependent on some bearer. Besides intrinsic aspects, are relational ones: entities that are existentially dependent on a multitude of individuals, binding them. These are RELATORS, which are the truth-makers of material relations. For instance, the "participation in trial" relator connects a patient with a clinical trial.

OntoUML has perdurants to represent events [1]. Events are characterized with the «event» stereotype. They have their own properties and can be decomposed. Events are immutable because they only exist in the past. Endurants and perdurants interact in several ways. For example, endurants *participate* in events, are *created* by events, and are *terminated* by events. Finally, since events are particularized instances that only exist in the past, roles played by objects in an event (i.e., while an event was occurring) are termed HISTORICAL ROLES (or HISTORICAL ROLE MIXINS, depending whether they are sortals).

4 Ontological Analysis and Redesign

We review the original conceptualization underlying CSHG using an ontological analysis mediated by OntoUML. The results lead to an improved CSHG, whose sound and precise ontological commitment fulfills a conceptual clarification (Fig. 2). This analysis focuses on clarifying the notions of entity and event and how they relate to each other. The conceptual characterization of the Pathways view of the CSHG begins with the identification of its two main concepts, "entity" and "event". In the original UML class diagram of Fig. 1, are represented as simple classes. However, their exact conceptual characterization can be made explicit using OntoUML's finer-grained class and association constructs (reflecting UFO's distinctions among endurant and event types and relations).

The entity concept (renamed "biological entity") is used to define every physical entity that may have a role in one or more processes. Therefore, we annotated the concept of entity and the concept of simple with the «category» stereotype because categories aggregate essential properties to individuals that follow different identity principles (belong to different kinds). The "event" concept helps represent the set of biological events that occur in a body and are part of the human metabolism. In OntoUML, the stereotype «event» represents ontological entities that unfold over time, accumulating temporal parts and mapping the world from situation to situation [1]. By mapping our original class "event" to that notion in OntoUML [1], we add two new attributes, «begin» and «end», because events in OntoUML are framed by specific time intervals.

The original model had one type-reflexive relation connecting the Event class with itself and the rolenames "pre" and "post". That modeling choice left ambiguous whether this relation represented a mere temporal precedence between occurrences or a stronger causal connection. To make it explicit that the intended semantics referred to the latter, we used OntoUML's «historicalDependence» stereotype [1]. If an event of type A is historically dependent on a event of type B, then instances of A must necessarily be preceded by instances of type B. Historical dependence implies temporal precedence, but not vice versa.

Events can be composed of a set of other events, forming partonomies. This can be illuminated by the mereology of events underlying OntoUML [8]. Mereology accounts for two orthogonal dimensions of decomposition of events: a *structural dimension* and a *participation dimension*. For CSHG, following the structural dimension, there are two types of events: the process and the pathway. A process is an atomic, simple event of a given type that can not be decomposed into smaller parts, whereas a pathway is a complex event that can be decomposed into smaller events, which are either processes or pathways. This dimension is represented through an aggregation relationship with the event class. Following the language's imposed mereological theory, complex entities must be composed of at least two disjoint parts (the *Weak Supplementation Axiom* [8]) with minimum cardinality constraints on the relations. This revised part of the model is a direct instantiation of UFO's structural partonomy pattern [8].

The participation dimension is characterized by representing the role that biological entities play in processes. This was originally modeled by the "takes_part" class in the UML schema, where we showed that an entity can act as an input, an output, or a regulator in a process. This representation has been expanded in the OntoUML version of the schema. First, we created a set of classes stereotyped with «historicalRoleMixin» to indicate playing roles, which biological entities have participated in, as an event. Unlike the UML schema, the minimum cardinality of the association between the historical role and the process is one. For a biological entity to play the role, it must have mandatorily participated in an event. Historical roles explicitly describe the variety of roles that biological entities may play in the processes.

Events depend on biological entities. Since atomic events (i.e., processes) are directly existentially dependent on biological entities, we can use the extension-

ality principle of the event mereology to derive the existential dependency of complex processes. The defined role mixins enables the creation of "portions" to describe the specific participation of an entity. The participation class, stereotyped as «event», divides an event into the individual participation of biological entities. Every instance of the "participation" class is derived from parthood and existential dependence, and bound to a specific subtype of a historical role mixin. Making explicit the notion of participation is of great importance from an ontological point of view. For instance, the process by which proteins are synthesized (translation) can be decomposed into atomic steps (e.g., initiation, elongation, and termination) to model the "constructed" dimension by creating segments using temporal schemes as external references. It can also be decomposed into portions that encapsulate the participation of biological entities in the whole process (e.g., participation of the ribosome and the mRNA strand).

Another capability, enabled by the use of the «event» stereotype, is modeling the creation and termination of biological entities. Millions of molecules are created and destroyed by different events that occur in our body, which is a special type of participation of endurants (i.e., biological entities) in events. To represent this situation, we modeled two phases to represent whether an entity exists or has been destroyed. The «phaseMixin» stereotype is used to represent changes in intrinsic properties of kinds (destroyed or not). If a biological entity is related to an event using an association stereotyped with «creation», that entity is created in that event. Similarly, for the «termination» stereotype.

5 Results

Several changes resulted from the ontological analysis, the most important of which is the clarification that enforces conceptual transparency on the initial representation. The ontological analysis identified and changed several aspects of the model to better grasp domain semantics, improving the representation of events and how biological entities change over time. The use of the «phase» stereotype in the OntoUML model enriches the representation of the effects caused by events. In the UML version, an entity can act as an input, an output, or a regulator. In the OntoUML version, an additional dimension allows us to indicate whether the entity has been degraded. An entity can be modeled that is: i) degraded as a result of a process; ii) created as a result of a process; iii) modified as a result of a process; or iv) degraded as a result of regulating a process. This change in the state of an entity (degraded or not) could not be modeled without the phase stereotype. This clarifies that the changes of biological entities in our bodies result from processes, but it is not clear how to model the degradation of entities within the UML model. The creation of the active_entity and degraded_entity phases provides additional mechanisms to ensure the correctness of the model. This prevents introducing errors when instantiating and populating the model. Such constraints are difficult to identify in the UML model.

6 Conclusion

The modeling of the human genome is an effort to understand life through the development of a conceptual model. This research has several implications. First, recognizing the complexity of this domain shows the importance of representing the human genome by a model that supports a shared understanding. Second, by making the ontological clarity of the conceptual model explicit, it is possible for the model to have a solid foundation. Further work will include the addition of notions of situation and disposition. These concepts are important because they enable the representation of diseases and pathways using situations and altered functions of modified proteins as dispositions. Various parts of CSHG will be instantiated and the ontological exercise applied to the remaining views. Thus, conceptual models are a practical way for domain experts and computer scientists to share the knowledge needed to develop systems to support processing of the huge amount of genomics data that now exists.

References

1. Almeida, J.P.A., et al.: Events as entities in ontology-driven conceptual modeling. In: International Conference on Conceptual Modeling. pp. 469–483. Springer (2019)
2. García S., A., et al.: Towards the understanding of the human genome: a holistic conceptual modeling approach. IEEE Access **8**, 197111–197123 (2020)
3. García S., A., et al.: A Conceptual Model-based approach to improve the representation and management of omics data in Precision Medicine. IEEE Access (2021)
4. Gaudet, P., et al.: Gene ontology: pitfalls, biases, and remedies. In: The gene ontology handbook, pp. 189–205. Humana Press, New York, NY (2017)
5. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. Ph.D. thesis, University of Twente (Jan 2005)
6. Guizzardi, G., et al.: Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model. In: ER 2021. pp. 356–366. Springer (2021)
7. Guizzardi, G., et al.: Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story. Applied ontology **10**(3-4) (2015)
8. Guizzardi, G., et al.: Towards ontological foundations for the conceptual modeling of events. In: ER 2013. pp. 327–341. Springer (2013)
9. Palacio, A.L., et al.: Genomic Information Systems applied to Precision Medicine: Genomic Data Management for Alzheimer's Disease Treatment. In: International Conference on Information Systems Development (2018)
10. Palacio, A.L., et al.: Toward an Effective Medicine of Precision by Using Conceptual Modelling of the Genome. In: IEEE/ACM International Workshop on Software Engineering in Healthcare Systems. pp. 14–17 (2018)
11. Pastor, O.: Conceptual modeling of life: beyond the homo sapiens. In: ER 2016. pp. 18–31. Springer (2016)
12. Pastor, O., et al.: Using conceptual modeling to improve genome data management. Briefings in Bioinformatics **22**(1), 45–54 (2021)
13. Pastor, O., et al.: Model-driven architecture in practice: a software production environment based on conceptual modeling. Springer (2007)

14. Román, J.F.R., et al.: Use of GeIS for Early Diagnosis of Alcohol Sensitivity. In: Bioinformatics 2016. pp. 284–289 (2016)
15. Smith, B., et al.: The ontology of the gene ontology. In: AMIA Annual Symposium Proceedings. vol. 2003, p. 609. American Medical Informatics Association (2003)