IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 7, 2021, accepted November 13, 2021, date of publication November 16, 2021,
date of current version November 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3128757

A Conceptual Model-Based Approach to Improve the Representation and Management of Omics Data in Precision Medicine

ALBERTO GARCÍA S., ANA LEÓN PALACIO, JOSE FABIÁN REYES ROMÁN, JUAN CARLOS CASAMAYOR, AND OSCAR PASTOR
PROS Research Center, Universitat Politècnica de València, 46022 Valecia, Spain

Corresponding author: Alberto García S. (algarsi3@vrain.upv.es)

This work was supported in part by the Spanish State Research Agency and the Generalitat Valenciana under Project PROMETEO/2018/176, Project PDC2021-121243-I00, and Project INNEST/2021/57; and in part by the European Regional Development Fund (ERDF) and the European Union NextGenerationEU/PRTR.

ABSTRACT Precision medicine has emerged as a disrupting medical model to transform a historically reactive medicine into a proactive one that focuses on delivering individualized treatment. A relevant challenge of precision medicine is to integrate the large amount of omics data that exists. This data has a high degree of heterogeneity, dispersion, and isolation. In addition, there is a lack of a solid ontological commitment regarding domain concepts and definitions, and a unified guideline of how to transform data into knowledge is missing. In this work, we report our experience applying conceptual modeling to deal with these challenges in a specific genomics dimension, i.e., Precision Medicine. To do so, we have applied conceptual modeling techniques. The use of these techniques allows us to create representations of the world (i.e., conceptual schemes) that can be used for the purposes of understanding, communicating, and problem-solving. They also help to establish a common ontological framework to facilitate both communication and knowledge evolution in complex domains. We identify a set of limitations that emerged after working in a precision medicine context, and we describe how we have solved them using conceptual modeling. Thus, the main contribution of this work is to present the subsequent Conceptual Schema that allowed us to overcome these limitations, which provides a better representation of proteomics data and eases its integration.

INDEX TERMS Conceptual modeling, CSHG, CSG, evolution, genomics, human genome, genome.

I. INTRODUCTION

Getting a precise understanding of how life works is a titanic task, which is as complex as exciting. One of the fields that focuses on achieving this purpose is genomics, a novel interdisciplinary field of biology. Although our knowledge of this field is still quite limited, we have already begun to take advantage of this knowledge in several domains, such as agrifood and precision medicine.

The practice of medicine is currently undergoing a paradigm shift thanks to precision medicine, transforming a reactive medicine into a proactive one that focuses on individualized disease diagnosis, treatment, and prevention. To this aim, precision medicine can be tackled using a computational approach that attempts to integrate and interpret genomics data. Despite the high use of the term precision medicine in the literature, its exact meaning remains unclear [31]. Precision medicine is a concept that is closely related to ''personalized medicine''. Initially, precision medicine was defined as *how molecular diagnostics allows doctors to diagnose the cause of a disease unambiguously* [64]. Both terms (i.e., personalized medicine and precision medicine) have so much overlap that they were often used indistinctly. Over time, precision medicine has gained more popularity because of multiple reasons. First, ''personalized'' means to devise a different treatment for each individual patient, which is how medicine is practiced every day [27].

More recent approaches have tried to provide a more comprehensive definition of precision medicine. These approaches treat precision medicine as a standardized process [31]. Indeed, the personalized approach is inherent to the doctor-patient relationship, but being able to exploit new

The associate editor coordinating the review of this manuscript and approving it for publication was Wen-Sheng Zhao.

data adds a considerable amount of information that is beyond observable signs and symptoms. Precision medicine incorporates clinical data, lifestyle, or genomics information, among others. Commonly used definitions of precision medicine can be summarized as *the stratification of patients using novel approaches* [3]. The novelty of this new stratification is that it is not based on classical symptoms. Instead, it is based on identified "treatable" traits.[1] The result of this stratification is the identification of subgroups of patients whose treatment can be improved because of more precise drugs related to the recognition of these treatable traits.

As we have mentioned above, plenty of new data is taken into account in precision medicine, being genomics data the one the we are interested in. Genomics is a field that focuses on studying the genomes of organisms. Genomics includes, among other fields, the study of i) DNA sequences [37], ii) the functionality of genes [28], or iii) how our genome contributes to the development of the nervous system [6]. Genomics is part of what is called omics. Omics is a term that includes several fields of biology whose names end with the suffix omics and focus on specific fields of life. For instance, genomics focuses on the study of the genome, proteomics focuses on the study of proteins, metabolomics focuses on the study of the metabolism, etc. All these fields are relevant and provide precision medicine with very valuable data that can be converted into knowledge and, consequently, into better treatments. In this work, we have focused on how to improve the representation of the three fields of omics mentioned above, namely, proteomics (Section III-A), genomics (Section III-B), and metabolomics (Section III-C).

The rise of precision medicine is closely related to our ability to generate large amounts of omics data: we can sequence a whole human genome for one thousand dollars, but the cost was twenty times bigger a decade ago. In line with that, the speed at which we identify and sequence omics data has risen with the advent of Next Generation Sequencing (NGS) techniques. NGS is also called second-generation sequencing, and they dropped the sequencing costs [5] and increased sequencing speed [21] dramatically. The reason for that is parallelism: NGS techniques can parallelize many of the reactions carried out in sequencing, increasing throughput [24], [36]. Another difference is that NGS techniques use light detection methods whereas previous sequencing techniques used chemical methods.

The lack of a solid ontological commitment that is shared across the scientific community and the existing data problems that are inherent to genomics complicate the achievement of this goal. Some of the causes of the problems mentioned above include the following: the increasing sequencing capacity thanks to NGS techniques [54]; the lack of a unified guideline of how to transform data into knowledge [58]; and the high degree of heterogeneity, dispersion, and isolation that exist in genomic data [41]. The issue of domain heterogeneity stems from the fact that there are many different standards and formats. The issue of dispersion is due to the fact that there are hundreds of different relevant genomic databases, many of which are created or removed every year [48]. The isolation issue arises because it is difficult to integrate or interconnect genomics data because of the two problems mentioned above. These problems are globally referred to as genomic data chaos.

We face these challenges by applying Conceptual Modeling (CM) techniques and using their advantages to solve most of the problems that hinder the possibility of making precision medicine a reality in clinical practice. CM is defined as the activity of creating mental representations of the world (i.e., conceptual models) to be used for the purposes of understanding, communicating, and problem-solving [38]. Conceptual models are intuitive, direct, and natural representations of a domain that allows us to answer fundamental questions by identifying relevant concepts and their relationships with each other. Thus, they help to establish a common ontological framework that facilitates both communication and knowledge evolution in complex domains [10].

The use of CM has provided us with a solution to deal with one of the most complex tasks of current genomics research, the understanding of the human genome from a holistic perspective. This solution is the Conceptual Schema of the Human Genome (CSHG). The initial approach of the CSHG, which was centered only on the human genome, has been re-engineered to be transformed into a species-independent Conceptual Schema (CSG). More details regarding this re-engineering work and its validation can be found in [50]. Our CSG has been used in several real-world use cases, which can be grouped into two contexts. The first context is the genetic improvement of crops. The second context is precision medicine. We have identified a set of limitations of the CSG when we used it in the most recent use cases of precision medicine. These limitations need to be addressed in order to provide stakeholders with the best experience when using the CSG. Besides, the complexity and dynamism of genomics require the CSG to be in constant evolution and improvement.

In this work, we describe the identified limitations, and we show how we have expanded the CSG to solve them. This expansion of the CSG includes improving the representation of proteins, the addition of the clinical actionability concept, and improving the representation of pathways and their biological entities. In addition, we report the results of the subsequent discussions, focusing on the ontological commitments established. Thus, the main contribution of the work is to present an extended version of the CSG. This version has been improved and expanded enough to work with our most recent real-world use cases, which are associated with precision medicine practices.

The rest of the paper is structured as follows: Section II-A describes our previous work regarding the design of a holistic Conceptual Schema of the Human Genome. This section presents the CSHG and its evolution, from its initial version to the final one. Section II-B briefly describes the re-engineering

[1]This process is also known as deep phenotyping.

process that was carried out to obtain a Conceptual Schema of the Genome that is species-independent. Section III discusses the changes that we have made to the CSG to obtain the latest working version. Section IV ends with our conclusions and addresses further work.

II. BACKGROUND

In this Section, we report our previous work when modeling the Genome. First, we describe how the CSHG has been updated over the years, from its first working version, to the last one: version 3. Second, we motivate the need for creating a conceptual schema that is species-independent.

A. EVOLUTION OF THE CONCEPTUAL SCHEMA OF THE HUMAN GENOME

For years, our CSHG has provided us with a holistic perspective of the human genome. It constitutes the conceptual background needed to define and describe the relevant genomic components, and it makes the understanding of the human genome a viable task. Understanding the genome is a complex task because it is a domain that is under constant evolution, and even the most basic concepts are under discussion. In addition, there are plenty of polysemic terms, and selecting the appropriate meaning for a given context is not trivial. The decisions taken have a direct impact on the data analysis strategies that are associated with precision medicine.

The CSHG has evolved along with the advances of the ontological commitments that are associated with genomics. Initially, the main goal of the CSHG was to decipher life and understand the internals of genomics. Although this goal still exists, we have expanded our vision with the rise of precision medicine. The first versions were developed to understand the inners of the genome. But the later versions have been developed to understand the inner of the genome **and** improve the representation and management of omics data in precision medicine.

The evolution of the CSHG to adapt to novel discoveries is necessary in order to ensure that it is both a valid and helpful artifact. Thus, our CSHG has undergone four major updates to adapt to these discoveries. A description of each version of the CSHG is found below, which provides clear insight into how relevant the update of the characterization of the human genome is.

1) CSHG VERSION 1

The goal of this preliminary version was to model the most basic concepts that play a crucial role in genomics from a unified and holistic perspective. In this version, we proposed a gene-centered vision. In that vision, the notion of gene is the central and most important concept, and its DNA sequence constitutes the structural unit of description. We focused on studying individual genes, the mutations that may affect them, and the consequences of these mutations from a holistic perspective.

Version 1 was divided into three main views: the ''Gene-mutation view'', the ''Genome view'', and the ''Transcription view''. The Gene-mutation view represents the structure of the genes and their allelic variations. The ''Genome view'' describes the variations that may be present in individual genomes when compared to the reference genome (i.e., individual-specific variations). The ''Transcription view'' models two dimensions. The first dimension is the set of basic components that participate in the protein synthesis process. The second dimension is the way they are affected by allelic variations in the genes that codify these proteins. More details regarding Version 1 can be found in [42].

2) CSHG VERSION 1.1

The main contribution of this version was the ''Phenotype view''. We characterized phenotypic information in order to provide more consistency and completeness to the model. It reinforced the holistic perspective of our CSG by explicitly establishing the link between phenotype and genotype. The ''Phenotype view'' represents the different phenotypes, their classification, and their severity. It also connects them to the variations that are responsible for their expression. More details regarding Version 1.1 can be found in [49].

3) CSHG VERSION 2

The third update of the CSHG completely changed how the genome is represented. While the previous versions were focused on gene sequences, this version focused on analyzing chromosome sequences. There are three reasons for this decision. The first reason is that domain experts work with DNA sequences whose origin is other genome structures that are different from genes (e.g., promoters, intergenic regions, etc.) The second reason is that the study of RNA sequences and amino acid sequences is as frequent and relevant as studying DNA sequences. The third reason is that there is a lack of consensus regarding the precise definition of the gene.

Consequently, Version 2 changed from a gene-centered to a chromosome-centered perspective. A new concept, called *chromosome element* was defined to easily model any relevant genome component as a part of a chromosome, providing its sequence. More details regarding Version 2 can be found in [45].

4) CSHG VERSION 3

The most recent update of the CSHG comprised a large number of updates. The first update consisted of conceptually refactoring a set of concepts that were too tied to specific solutions (i.e., databases or data types) in order to increase the genericity of the model. The second update consisted of including the notion of assembly and its implications regarding chromosomes and chromosome elements. The third update consisted of expanding the representation of the transcription process. The fourth update expanded the ''Variation view'' by updating how variations are characterized hierarchically. The fifth update was centered on

representing the structural changes that variants cause at the DNA, RNA, and amino acid levels.

Version 3 is divided into five views: i) the ''Structural view'', which describes the structure of the human genome; ii) the ''Transcription view'' which models protein synthesis; iii) the ''Variation view'' which characterizes changes in the sequence of the human genome regarding a reference sequence; iv) the ''Bibliography view'' which details information and sources related to the elements of the conceptual schema; and v) a new view, called ''Pathway view'', which represents human metabolic pathways, thereby increasing the holistic perspective of the conceptual schema. More details regarding Version 3 can be found in [16]

B. TOWARDS A SPECIES-INDEPENDENT CONCEPTUAL SCHEMA OF THE GENOME: THE CSG

After updating and extending the CSHG to Version 3, we have used it in real-world use cases related to precision medicine. However, since we have expanded the scope of our research to new genomic domains, we studied the suitability of applying the Conceptual Modeling techniques used to develop the CSHG to design a conceptual schema that could be useful in representing the genome of other species. Two characteristics are shared by every species. The first characteristic is that every organism is made of cells, and these cells contain genetic information in the form of DNA o RNA. The second characteristic is that life is just a set of chemical reactions at the molecular level.

Thus, an opportunity arose to check the validity of our assumptions. We collaborated with a research institute that studies the genome of the genus citrus. In this collaboration, we created a new Conceptual Schema that precisely captures the relevant concepts of this particular domain and their relationships with eachother. We called it the Conceptual Schema of the Citrus Genome (CSCG). A Genome Information System was developed using the CSCG as the ontological foundation [17].

After validating our previous insights, we ended up with two different conceptual schemes: the CSHG and the CSCG. Although there are differences between the two conceptual schemes, there are strong similarities in their concepts (e.g., the structural parts of the DNA sequence or the synthesis of proteins). Therefore, we proposed a novel hypothesis: if the genome at its most elemental level is the same, then the species from which the genome is being studied in a specific use case does not matter. This hypothesis was validated by creating the Conceptual Schema of the Genome (CSG) to precisely capture the genome components that are common across species. The CSG is CS that is generic enough from which specific conceptual views can be generated. These views are tailored to the particularities and idiosyncrasies of the use cases that we might face, making the working species just another particularity. Several challenges had to be faced to create the CSG, such as modeling the concepts of species and ortholog group. These additions allowed us to interconnect the different genomic components between species. The resulting conceptual schema was empirically validated to ensure its quality and correctness. More details regarding the generation of the CSG and its validation can be found in [15], [50]

III. EXPANDING THE CONCEPTUAL SCHEMA OF THE GENOME TO IMPROVE ITS USE IN PRECISION MEDICINE

As we mentioned in Section II-B, the most recent use cases where we have used the CSG required expanding some parts. Specifically, the transcription view, the variation view, and the pathway view. For the transcription view, we found that the representation of proteins required adding some new concepts. Three topics that are a matter of study in precision medicine are missing in the CSG: i) the description of protein function and the identification of protein-protein interactions [25], [33]; ii) the existence of protein isoforms [35]; and iii) the description of protein structure [51].

For the variation view, determining the clinical importance of variants is a crucial aspect. This process is known as variant interpretation. It consists of the classification of the variants according to standards and guidelines, such as the ACMG/AMP [47] or Sherloc [40]. However, the way some genomic databases handle variant interpretations can be improved. For example, some databases integrate all of the interpretations a variant has for different phenotypes to create an aggregate value that can lead to misinterpretations, which worses precision medicine diagnoses [34].

For the pathway view, there is a need to improve the characterization of the different physical entities that take part in pathways. In addition, the view needs to be expanded to allow stakeholders to i) locate the places where pathways are carried outm and ii) describe the changes that occur in the cell functionality because of these pathways.

A. EXTENDING AND IMPROVING THE REPRESENTATION OF PROTEINS

Proteins are macromolecules that play a fundamental role within every cell and metabolic reaction. For example, protein kinases are known to enable learning and memory [19], high concentrations of C-reactive protein (CRP) are associated with an increased risk of heart diseases [7], etc. The previous version of the CSG represented proteins as a type of entity that is characterized by an identifier, a name, and a description. Some proteins can be enzymes, which have a commission number. Proteins can take part in processes as an input, an output, or a regulator, and the combination of these processes represents biological pathways.

Although this approach captures the role of proteins in metabolic reactions with an appropriate level of detail, four challenges arose that are related to lower-level protein structure and functionality.

The first challenge is related to the existence of protein isoforms and precursor proteins. Protein isoforms are generated as a consequence of four events: alternative promoter usage [30], alternative splicing [28], alternative initiation [56], and ribosomal frameshifting [8]. In the previous

version of the CSG, protein isoforms were not modeled. Consequently, there was a loss of knowledge. For example, the Epidermal Growth Factor Receptor (EGFR) protein has four isoforms, but isoform 2 has relevant differences compared to the other ones. These specific isoforms may act as an antagonist of the epidermal growth factor. They are expressed in ovarian cancers [22], and their length is 405 amino acids instead of the 1,210 amino acids of the canonical isoform [44]. Another example is the Vascular Endothelial Growth Factor A (VEGFA) protein, which has seventeen isoforms. While some VEGFA isoforms are expressed widely, others are not [53]. A precursor protein is a protein that requires additional processing to become mature. This processing is called Post-translational Modifications (PTMs). PTMs include the extent of peptides, the cleavage of the initiator methionine, or the covalent binding of a lipid group. For example, the Apolipoprotein E (APOE) protein undergoes multiple glycosylations [39], although they are not required for proper expression and secretion [60]. The reported information could not be represented in the CSG; thus, the characterization and addition of these concepts are needed.

The second challenge is related to the physical structure of proteins. The amino acid sequence of proteins has a three-dimensional arrangement in space. This arrangement is stabilized by polar hydrophilic hydrogen, ionic bond interactions and internal hydrophobic interactions between non-polar amino acid side chains [13]. A protein's three-dimensional structure dictates its biological function. Knowing a protein structure provides a better understanding of how it works, which, in turn, allows us to hypothesize about how to affect it or modify it. This knowledge is of crucial importance in pharmacoproteomics [61]. However, experimentally identifying the structure of a protein is a complex process that requires a large amount of money, time, and expertise. The rate at which proteins are discovered and sequenced is much higher when compared to the discovery of protein structures. Currently, there are 1,500 times more protein sequences than structures. All in all, the availability of new protein sequence data continues to outpace the availability of experimental protein structure data by far, only increasing the need for accurate protein modeling tools [61]. Therefore, we determined that the CSG should model protein structure, including the secondary structure, the super-secondary structure, and the tertiary structure. This addition allows us to identify potential changes in protein structure due to the presence of variations in coding sequences of the DNA. Besides, Conceptual Modeling (CM) also helps to establish common ontological frameworks that facilitate both communication and knowledge evolution in complex domains such as genomics and proteomics because it answers fundamental questions regardless of the research area by identifying what concepts are relevant and their relationships with each other [10]. It also provides a means to make mental representations of the world explicit. Therefore, we believe that the use of CM will help develop more accurate protein modeling tools, which is a relevant addition to the existing body of knowledge.

The third challenge is related to the biophysical and chemical properties of proteins. These properties also alter how they interact with the environment. In the previous version of the CSG, we modeled proteins as components that took part in processes in a static way (i.e., they will always perform the same way); but this approach is incomplete since they are not isolated entities. They are affected by the environment. There is a set of protein properties that alters their functionality: the wavelength at which photo-reactive proteins show maximal light absorption; the optimum pH at which proteins perform their activity; and the enzyme's maximal velocity (i.e., the rate attained when the site of an enzyme saturates with a substrate). As examples, the optimum pH for the Phosphatidylserine lipase (ABHD16A) enzyme is between 7.2 and 8.0 [52], and the Pyruvate kinase (PKM) enzyme has a Michaelis constant (KM) value of 2.7 mM for phosphoenolpyruvate at 32 degrees Celsius [12]. We concluded that these properties and context-dependent properties are required in the CSG in order to improve its holistic representation of proteomics.

To achieve our goal of integrating this information into our CSG, we started by studying the knowledge that is currently available in the domain. To do this, we decided to perform an in-depth study of the Universal Protein Resource (UniProt) database [55]. Uniprot provided us with valuable information regarding protein-related knowledge and how data is stored. For the challenge described above, UniProt contains detailed information on protein isoforms, precursor proteins, and the PTMs that transform a precursor protein into a mature one. It also describes the secondary and tertiary structure of proteins, including motifs and functional domains, and their location in the amino acid sequence. Uniprotalso offers contextual information on proteins. It provides data regarding protein light absorption, kinetic parameters, pH dependence, redox potential, and temperature dependence. Great effort was required to obtain the level of knowledge from Uniprot that is needed to expand our CSG, which resulted in the generation of a conceptual schema of UniProt. More details regarding the obtention of such a schema can be found in [32], where the conceptualization process is reported in detail.

Once we obtained a deep understanding of how protein knowledge is represented in UniProt, we started the next step: integrating this knowledge into the CSG in order to effectively overcome the challenges described above. The resulting conceptual schema is shown in Figure 1. For the first challenge, we created the concept of ISOFORM, which is defined as one of the sets of different versions that a PROTEIN can have. Each isoform is characterized by an identifier, a name, its sequence, its status (which is used to indicate whether the reported information has been manually validated by expert curators), and its level of existence (which is used to indicate to what extent its existence has been confirmed). Isoforms can either be MATURE or PRECURSOR ones. While precursor isoforms require further processing to

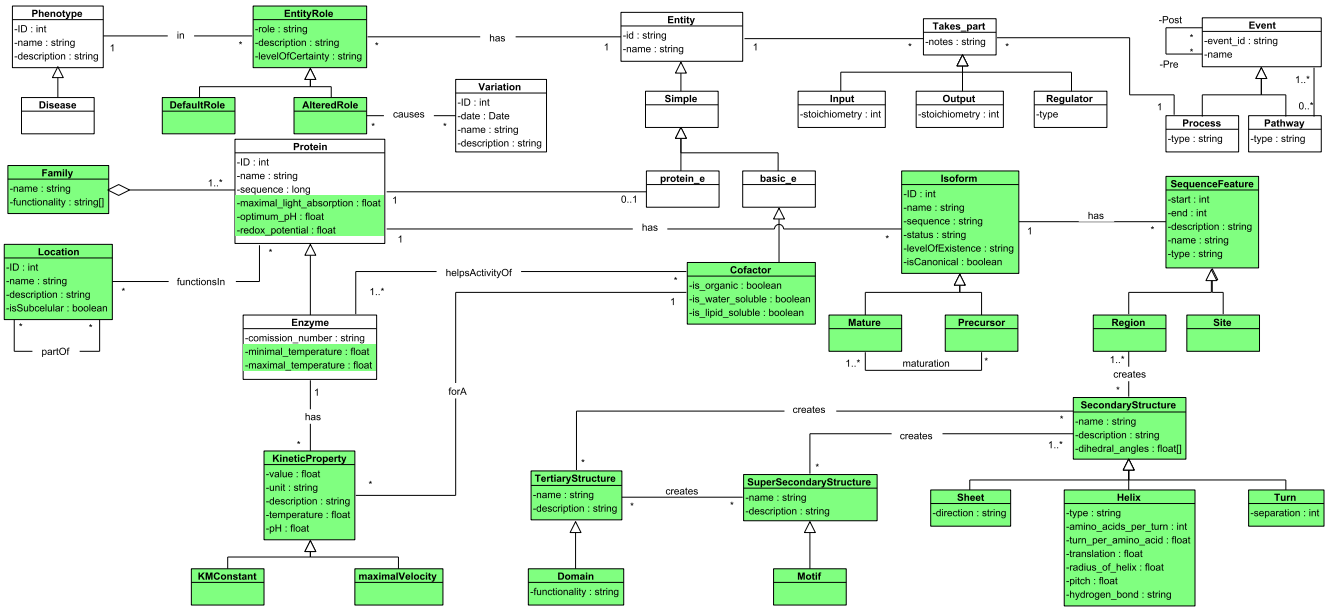

FIGURE 1. Representation of proteins in the Conceptual Schema of the Genome. The image shows a simplification of the CSG where only the modified concepts and those directly related to them are shown. Additions are represented in green.

become mature, not every mature isoform is derived from a precursor protein. Finally, one of the existing isoforms of a protein (either mature or precursor) is the canonical isoform, which is defined as the most prevalent one.

For the second challenge, we modeled the physical arrangement of proteins in space through their primary, secondary, and tertiary structures. The primary structure of a protein is its sequence of amino acids, which is an attribute of the isoform class. Also, the primary structure of the canonical isoform of a protein can contain a set of SEQUENCE FEATURES. They are defined as specific parts of the amino acid sequence that are considered to be of interest. Two types of sequence features can be identified: SITES and REGIONS. Sites describe single amino acid sequences such as cleavage, inhibitory, or breakpoint sites. Regions describe sequences of more than one amino acid with a functional or biological interest, e.g., a region that mediates transcriptional activity.

Then, we characterized the concepts of SECONDARY STRUCTURE and SUPER SECONDARY STRUCTURE. Secondary structures are recurring arrangements in space of adjacent amino acids [20]. They arise from the identified regions of interest in the sequence. This allows connecting the primary and the secondary structures in the model since secondary structures are the three-dimensional form of regions that are identified in the primary structure. Secondary structures are characterized by a name, a description, and their dihedral angles. The dihedral angle is the internal angle of an amino acid sequence at which two adjacent planes meet. The ϕ and the ψ angles are used. They have a limited number of possible values due to the existing chemical effects inside the protein structure (the possible values are identified in Ramachandran plots). The main secondary structures are sheets, turns, and helices, with helices being the most complex ones to characterize. While the sheet is characterized by its direction and the turn by the separation between the last two residues, the helix is characterized by its type, amino acids per turn, translation, radius, pitch, and hydrogen bond. The type indicates whether it is right-handed or left-handed. The value for amino acids per turn indicates the number of amino acids needed per turn of helix, and the turn per amino acid (i.e., the turn in degrees caused per amino acid) is derived from this value. The translation indicates the translation distance (in nm) per amino acid. The radius indicates the radius (in nm), of the helix. The pitch indicates the vertical distance (in nm) between two turns. The hydrogen bond indicates the type of hydrogen bond of the helix. Table 1 shows a set of helices as an example of how secondary structures can be instantiated using this characterization.

A super secondary structure is a combination of multiple secondary structures [20]. For example, a β-barrel is a super secondary structure that is obtained from a tandem repeat of β-sheets. Some super secondary structures are motifs, which are three-dimensional structures shared by a variety of different and evolutionarily unrelated proteins. For example, a helix-loop-helix is a motif that is composed of two α-helices connected by a turn.

Finally, we characterized the concept of TERTIARY STRUCTURE, which refers to the overall three-dimensional arrangement of a protein in space [20]. A tertiary structure is characterized by a name and a description. Tertiary structures are defined by the secondary and super secondary structures that compose them. Protein domain identification is relevant because the presence of a specific set of secondary or super secondary structures in a protein does not necessarily

TABLE 1. Example of helix structures. Aa refers to amino acid. The description field is empty for space reasons. R/L means that the helix can be either right-handed or left-handed.

Name	Description	Dihedral angles	Type	Aa per turn	Turn per Aa	Translation	Radius of helix	Pitch	Hydrogen bond
Alpha	-	[-60, -45]	R/L	3.6	100º	.15	.23	.54	i + 4 -> i
310	-	[-49, 26]	R/L	3	120º	.2	.19	.06	i + 3 -> i
pi	-	[55, 70]	R/L	4.4	87º	.11	.28	.48	i + 5 -> i

indicate a specific functionality. Thus, two proteins may share common secondary structures with unrelated functionality. A tertiary structure can also be specialized into a DOMAIN. Protein domains are the smallest functional unit of a protein. They are self-stabilizing regions that fold independently and have specific functionality. Domains are the building blocks of the proteins, and they can be part of several of them. They are also used to infer protein-protein interactions (PPIs) because domains are directly involved in intermolecular interactions. A huge number of domains have been identified, each with a specific cellular function [57]. This approach allows us to model the three-dimensional arrangement of proteins in space and to locate variations in the structural and functional units that shape proteins.

Even though motifs and domains are made of secondary elements, they have a completely different nature. Motifs are formed by the connection of helices and sheets through turns. They mainly have a structural function, although sometimes they can perform similar biological functions in a particular protein family, and they are not independently stable. Domains can contain both secondary and super secondary structures and are formed by disulfide bridges, ionic bonds, and hydrogen bonds. Each domain has a unique function, and they are independently stable.

For the third challenge, we included the identified parameters into the protein and enzyme entities, and we defined the concept of kinetic context. In proteins, we added the following attributes:

- Maximal light absorption: the wavelength (in nm) at which photo-reactive proteins show their maximal light absorption.
- Optimum pH: the optimum pH for protein activity.
- Redox potential: the tendency of a protein (in mV) to gain or lose electrons.

In enzymes, we included the minimal and maximal temperature, which indicates the upper and lower limit of temperature (in Celsius) where an enzyme can perform its action. Enzymes also have two kinetic properties, the Michaelis–Menten kinetics (KM CONSTANT) and the MAXIMAL VELOCITY. The KM constant represents the substrate concentration at which half of the enzyme's active sites are occupied (this constant is used to measure the affinity of an enzyme for a substrate). It is the value of the substrate concentration at half of the maximal velocity. The maximal velocity of a reaction is reached when the enzyme sites are saturated by a substrate.

The values of the kinetic properties of the enzyme can change and are dependent on an external context that is dynamic. This context refers to the substrate that saturates the sites of an enzyme and two conditions that are optional: the temperature and the pH. Our approach models the default maximal velocity and KM constant of an enzyme for a specific substrate and modifies these values depending on the temperature and the pH. Substrates are represented in the model as COFACTORS, a specific type of the *basic* element entity. Cofactors can be organic or non-organic, water-soluble, or lipid-soluble.

Apart from the three additions mentioned above, we also defined the concept of ROLE. According to the data available at the UniProt database, proteins can play a specific role in a phenotype, and this role is altered when variations in the DNA appear. For example, there is evidence suggesting that protein misfolding causes several degenerative and neurodegenerative disorders such as Alzheimer's disease or Parkinson's disease [9]. We generalized this concept in the CSG so that every biological entity can have a role in a given phenotype. Roles are characterized by the role type, a description, and a level of certainty (which indicates to what extent this link is valid). There are two types of roles, the DEFAULT ROLE, and the ALTERED ROLE. The default role indicates the expected role of an entity for a phenotype. The altered role emerges when changes in the DNA appear. The altered role can be linked to a set of variations that cause it. An example of an altered role is a protein with decreased efficiency due to a premature stop codon. To illustrate, the Optineurin protein is coded by the OPTN gene, and it plays an important role in the maintenance of the Golgi complex, in membrane trafficking, and in exocytosis [14]. The expression of the Optineurin protein is regulated by intraocular pressure [59], which is an example of the default role of the protein; however, altered roles may arise because of different variants in the genome. For example, Optineurin may selectively promote cell death of retinal ganglion [46] or increase the risk of suffering Normal Pressure Glaucoma (NPG) [46], which are examples of altered roles of the protein.

These changes allowed us to overcome the three challenges reported above. We can describe the existence of precursor proteins and protein isoforms with a high level of detail and clarity. The physical structure of proteins has been included in the model; the primary, secondary, and tertiary structure of proteins have been identified and characterized. Finally, the most relevant biophysical and chemical properties of proteins have been identified and represented in the CSG.

In summary, the number of modifications applied to the model are the following: 23 classes have been created, and five attributes have been added to existing classes

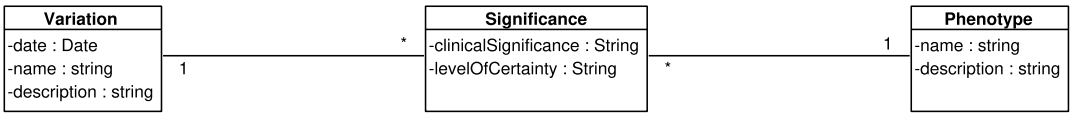

FIGURE 2. Clinical Significance in the Conceptual Schema of the Human Genome.

B. CREATING THE CONCEPT OF CLINICAL ACTIONABILITY

One of the pillars of the Precision Medicine (PM) approach is genetic diagnosis. It is based on determining the practical importance of each DNA variant according to its role in the development of a disease, i.e., identifying if a variant is disease-causing or increases the risk of suffering it. The role of a variant for a given phenotype is known as ''clinical significance''. The interpretation of the clinical significance of a variant is a challenging process that requires gathering and assessing the available evidence. Several publicly available databases, such as ClinVar, Ensembl, ClinGen, or CIViC, provide the clinical significance of variants.

Inspired by the approach that is provided by these data sources, the CSHG (and the CSG) modeled variants so that they can be associated with multiple clinical interpretations, and each clinical interpretation links a variation to a phenotype (see Fig. 2). Each variant (represented by the Variation class) may have multiple clinical interpretations provided by the scientific community (represented by the Significance class) for each Phenotype. The interpretations are described by the ''ClinicalSignificance'' and the ''levelOfCertainty'' attributes. The ''ClinicalSignificance'' attribute determines the practical importance of the variant (i.e., the variant's pathogenicity) while the ''levelOfCertainty'' represents the relevance of the evidence used by each expert to assess that importance (i.e., the reliability of the clinical significance). The level of certainty is crucial in genetic diagnosis since it provides clinicians with a means to determine whether to include or discard variants.

Clinical significance is a helpful tool in genetic diagnosis, and it allows the scientific community to transform existing data into relevant information. However, we have identified a number of challenges related to the use of the clinical significance that is provided by these repositories (i.e., ClinVar, Ensembl, etc.). These challenges can be summarized as follows:

- The clinical significance of variants tends to be reported as a whole rather than at a phenotype level. This representation has a direct impact on the selection of variants. For example, a variant whose clinical significance is pathogenic for a given phenotype but benign for another will be reported as a conflicting variant. This interpretation is correct if both clinical significances refer to the same phenotype, but it is incorrect if they refer to different phenotypes. Thus, the clinical significance must be analyzed at a phenotype level.
- The management of conflicts between different interpretations for a variant is imprecise and deficient. This management is, in part, a consequence of the item described above. The existing conflicts are not well-treated because they are annotated as conflicting without trying to solve them.

As a consequence of these problems, clinical experts are forced to review and analyze each interpretation in order to identify the correct role of variants. This review process frequently conforms to a tedious, manual, and error-prone working process that diminishes the added value of Information Systems in developing efficient precision medicine. The most problematic situation occurs when domain experts disagree about the role of a variant in the development of disease (i.e., one expert considers the variant as pathogenic and another one benign). This has led to the need to provide a new and improved solution.

We started by making a precise characterization of clinical significance. We characterized the existing types of clinical significance, and we determined how the conflicts that might arise among experts are managed. For the existing types of clinical significance, we have identified thirteen possible values:

1) **Pathogenic**: increases the susceptibility of predisposition to a certain Mendelian disorder.
2) **Benign**: reduces the susceptibility of predisposition to a certain Mendelian disorder.
3) **Likely Benign**: strong evidence in favor of reducing the susceptibility of predisposition to a certain Mendelian disorder.
4) **Likely Pathogenic**: strong evidence in favor of increasing the susceptibility of predisposition to a certain Mendelian disorder.
5) **Affects**: causes a non-disease phenotype, such as lactose intolerance.
6) **Drug Response**: alters a specific drug response in some way.
7) **Confers Sensitivity**: confers Sensitivity to a specific drug.
8) **Association**: identified the association to a disorder in a GWAS study.
9) **Uncertain Significance**: limited evidence regarding pathogenicity.
10) **Protective**: Decrease the risk of suffering a disoder.
11) **Conflicting data from submitters**: groups within a consortium have conflicting interpretations of a variant.
12) **Not Provided**: no clinical significance reported.
13) **Other**: any other possible value.

The existing clinical significances can be simplified and grouped according to their likelihood of causing a potentially

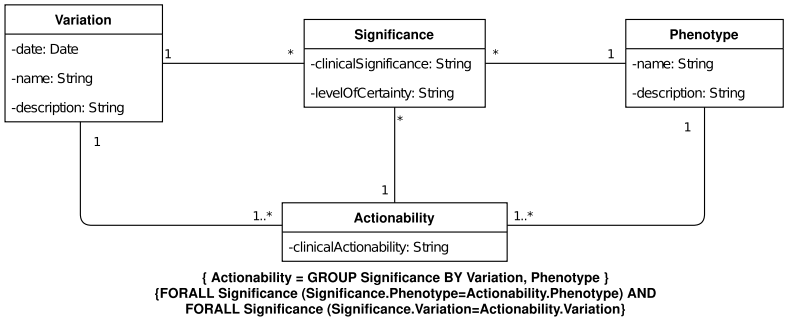

FIGURE 3. Conceptual schema to represent clinical actionability.

damaging phenotype. We have used this approach as a basis to create an aggregate value for each phenotype that is associated with a variant. We call it ''Clinical Actionability''. It is obtained by grouping the different interpretations that exist into a single term that integrates them. Clinical actionability provides a more precise assessment of the clinical significance of variants by integrating them and providing clinical actionability on a per-phenotype basis. It also precisely manages conflicts regarding the clinical significance of variants.

As shows in Fig. 3, in this version of the CSG, we have included the new ''Actionability'' class. This class is associated to a Variation, a Phenotype, and a set of Significances. The clinicalActionability attribute is the practical importance of the variant effect, i.e., it determines whether a variant is truly *actionable* for a phenotype based on the existing clinical significance. For each phenotype of a variant, the clinical actionability is calculated as an aggregate of the different clinical significance provided by experts. Only one clinical actionability is allowed for each Variation-Phenotype pair, which is represented as a data constraint to ensure data integrity and correctness. The clinicalActionability attribute can take five different values:

1) Disorder causing or risk factor: The variant is the cause of the phenotype or increases the likelihood of presenting it. This group includes the interpretations whose clinical significance is pathogenic, likely pathogenic, affects, risk factor, or association.
2) Uncertain role: The role of the variant in the development of the phenotype is not clear. This group includes the interpretations whose clinical significance is uncertain significance, or when conflicts between interpretations are present.
3) Not disorder causing or protective effect: The variant is not the cause of the phenotype nor does it provides a protective effect against it. This group includes the interpretations whose clinical significance is benign, likely benign, association not found, or protective.
4) Affects drugs or treatment response: The variant affects the sensitivity or response to the specified drug or treatment. This group includes the interpretations whose clinical significance is drug response or confers sensitivity.
5) Not provided: The variant does not have interpretations, and, as a consequence the clinical significance is unknown. This group includes the interpretations whose clinical significances is either unknown or not provided.

TABLE 2. The list of clinical significances associated to the c.2843G>A variation in ClinVar.

Interpretation	Phenotype
Pathogenic	Leber congenital amaurosis 8 Retinal dystrophy Retinitis pigmentosa 12 Pigmented paravenous chorioretinal atrophy CRB1-Related Disorders Abnormality of the eye Retinitis pigmentosa
Benign	Pigmented paravenous chorioretinal atrophy

The impact of the changes described above (i.e., including the new actionability class and its relationships in the CSG) can be summarized as:

- Abstraction of the different variant effects according to their likelihood of being disease causing or protective.
- Being able to evaluate the clinical impact of variants for each associated phenotype.
- Decreasing the effort required for the evaluation of conflicts between interpretations.
- Decreasing the effort required to add new data sources that use different terms to classify the clinical significance.

To illustrate, the c.2843G>A variant [1] is considered a conflicting variation, but this situation is caused because of the missing clinical actionability concept. When this variation is analyzed at a phenotype level, we see that there is no conflict regarding the interpretations of the variant (see Table 2).

This situation has relevant consequences in precision medicine and in the diagnosis of genetic diseases since variations with conflicting interpretations are usually discarded [23]. This variation is not a rare case. We performed an analysis of the variations stored in the ClinVar database (which can be accessed in [2]) in order to understand the magnitude of the problem. The result is that 41,433 out of 776,454 variants (i.e., 5%) have a conflicting clinical significance. Although it is a small percentage, these variants are located in 2,589 genes and affect 5,067 phenotypes. The most

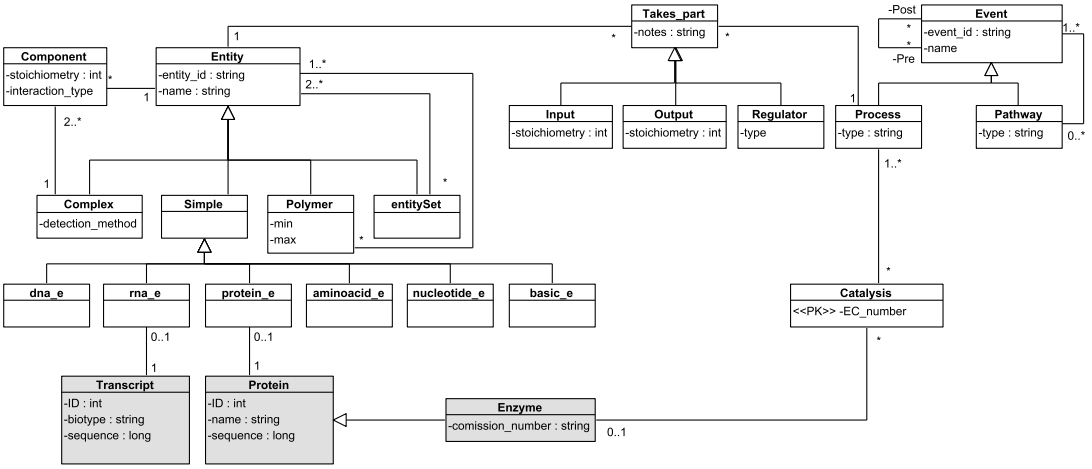

FIGURE 4. The previous representation of the pathway view in the Conceptual Schema of the Human Genome and the Conceptual Schema of the Genome. The classes that are shaded in light gray do not pertain to this view, but they are linked to one of the classes of the pathway view. Thus, they are represented in order to provide additional context and information.

affected phenotypes are cardiopathies, cancer, and different types of muscular dystrophy. These variants are more likely to be discarded from genetic diagnosis because they are labeled as conflicting even though their exclusion could lead to missing important information that could have a high impact on disease diagnosis and treatment.

In summary, the number of modifications applied to the model are the following: one class, three relationships, and one integrity constraint have been created.

C. METABOLIC PATHWAYS AND THEIR BIOLOGICAL ENTITIES

The pathway view of the CSG represents the metabolic reactions that take part in the organism of eukaryotic species. These reactions constitute a vast network of molecular interactions, where molecules are created, modified, combined, and destroyed. The pathway view represents all of these interactions as events in which input entities are transformed into output entities. The events can be regulated by regulator entities since they act as enhancers, promoters, catalysts, etc. Events are temporarily dependent on each other, which allows us to determine their temporal structure (i.e., the previous and later events of a given event). There are two types of events. The first type is the Process, which represents an atomic event that cannot be decomposed into more basic steps. The second type is the Pathway, which represents a complex event that is composed of a sequence of other events (either processes or pathways).

What is an entity? What kind of entities are defined? Figure 4 show our previous characterization of entities. We defined four types of entities: Simple, Complex, Polymer, and Entity Set. Simple entities represent the elementary entities that take part in processes. We have characterized simple entities based on the biological molecules that compose them. There are simple entities that are made of DNA (dna_e), RNA (rna_e), or amino acids (protein_e). Also, we have characterized the building blocks of these entities, i.e., the amino acid (aminoacid_e) and nucleotide (nucleotide_e) entities. Finally, there is another simple entity that is used to identify chemical compounds such as water or ATP (basic_e). The classes with "_e" are used as a dictionary to represent real biological elements. Therefore, the protein_e class is a dictionary class that represents the biological concept of the protein.

Simple entities can act together to perform biological activities, creating entities that can be Complex, Entity Set, or Polymer. Complex entities are created when at least two entities of any type are combined, such as protein complexes. Since each Entity plays a specific role in the complex, the role of a specific Entity in a Complex is represented through the Component class, which associates them and specifies the interaction type.

Entity Set entities are created when a set of entities of any type can be used indistinctly because they play an equivalent role. This type of entity is used to create aggregates of entities to reduce the granularity of pathways.

Polymer entities are created when at least two entities are concatenated. Unlike Complex and Entity Set entities, a Polymer must be composed of only one type of entity.

We conducted a workshop with a group of biologists and bioinformaticians to check the suitability of this representation and identify conceptual weaknesses. The result of this workshop can be summarized in the following points:

- There are some design decisions regarding the characterization of the Entity that requires further clarification. Are polymer entities a type of complex entity? Should the catalysis be a subtype of the regulator class? Do we really need to model the concepts of transcript and protein two times (using the rna_e and protein_e classes)? Why are the dna_e, rna_e, and the elements that compose them (i.e., the nucleotide_e class) represented at the same level of hierarchy? Should oligopeptides and

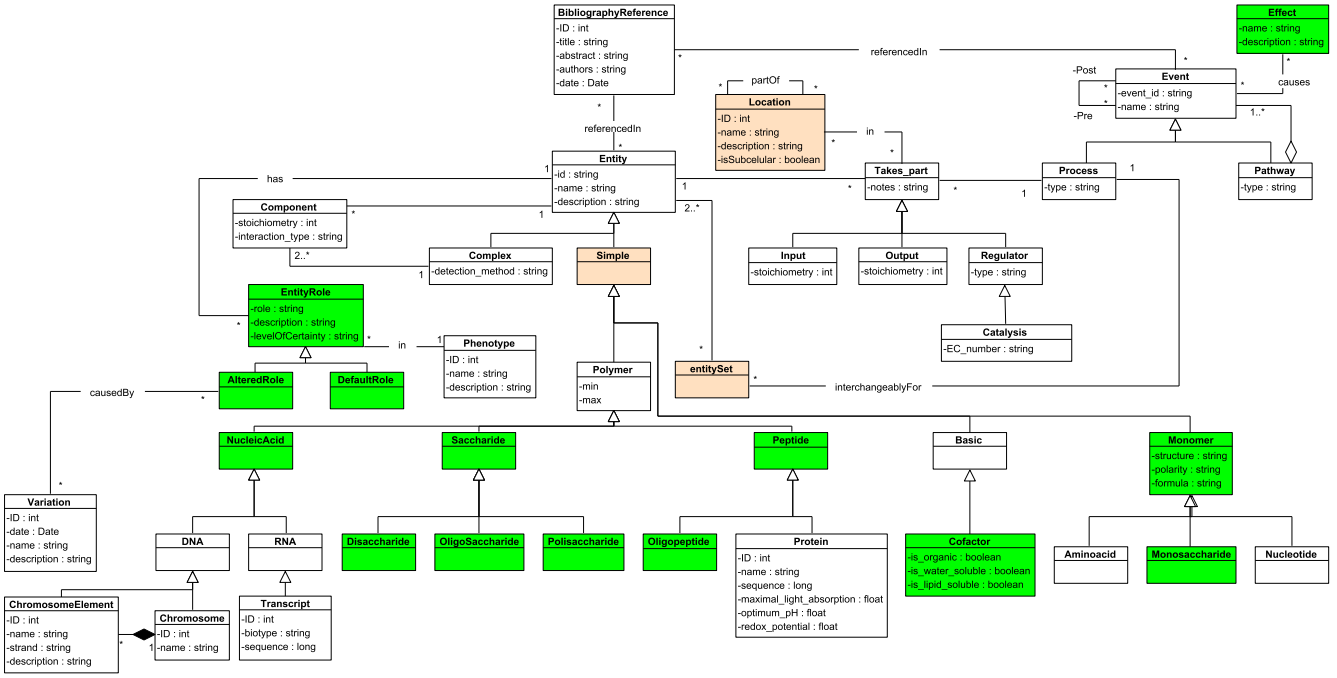

FIGURE 5. **Representation of the Pathway view of the Conceptual Schema of the Genome. The image shows a simplification of the CSG where only the modified concepts and those directly related to them are shown. Additions are represented in green. Modified concepts are shown in light orange.**

polipeptides be modeled also? Should saccharides be modeled also?

- There is no way to determine the location where the events are produced. This also refers to distinguishing an entity that can be in different locations (e.g., how to differentiate between extracellular glucose and cytosolic glucose).
- This characterization can only represent changes that modify entities, but it cannot represent other changes such us the movement of an entity in a cell or changes in intracellular pH.
- It would be interesting to provide a means to represent the high-level role of entities in phenotypes because the reasons at the level of the chemical reaction might be unknown.
- Gene Ontology (GO) [4], [18] is one of the most commonly used databases for annotating the functionality of genes and their products. Workshop members indicated that they use GO extensively, so integrating GO data into the CSG should be one of our priorities. In line with that, it would be interesting to provide a means to determine the role that genes and their products play in a phenotype.

To deal with these points, we started by studying Reactome,[2] since it is one of the most commonly used databases for working with reactions, proteins, and pathways. We downloaded the full Reactome data-set of information.[3] The data is stored in a graph-oriented database that is implemented using Neo4J.

After a thorough study of Reactome data, we started updating our CSG to address the points summarized above. Fig. 5 shows the resulting conceptual schema. The most notable change is that we have generalized the concept of ENTITY, and now every entity of the structural view and the transcription view is an entity. The number of subtypes of entities has been reduced to two: COMPLEX and SIMPLE. The notion of complex entity remains unchanged. It is defined as a class that represents the combination of two entities (either complex or simple). The notion of simple entity is more generic, and there are three types of simple entities: MONOMER, POLYMER, and BASIC.

A monomer entity represents the molecules that can react together to form larger entities (i.e., polymer entities). For example, nucleotides (i.e., a monomer) form nucleic acid (polymer). Monomers are characterized by a physical structure (i.e., cyclic or linear), a polarity (i.e., polar or non-polar), and a skeletal formula. There are three monomer entities in our model since they are the most representatives in life forms [26]. These are the amino acid, the monosaccharide, and the nucleotide.

A polymer entity represents the larger entities that are composed of a set of monomer entities of the same type. There are three types of polymer entities since we have three monomer entities. The first type of polymer is the PEPTIDE, which is made of amino acids. A peptide can be an OLIGOPEPTIDE if it made of at most fifty amino acids; otherwise, it is a PROTEIN [62]. The second type of polymer

[2]https://reactome.org
[3]https://reactome.org/download-data

is the SACCHARIDE, which is made of monosaccharides. A saccharide can be a disaccharide if it is made of two monosaccharides, an oligosaccharide if it is made of between two and twenty monosaccharides [11]; otherwise, it is a polysaccharide. The third type of polymer is the NUCLEIC ACID, which is made of nucleotides. A nucleic acid can be DNA if it is double-stranded and contains Thymine, or RNA if it is single-stranded and contains Uracil.

A basic entity represents chemical compounds that can be used in biological processes, such as water or Adenosine Triphosphate (ATP). Some basic entities can be required by enzymes to perform their catalytic activity. They are called COFACTORS. Cofactors can be water soluble/ lippid soluble, and they can be organic/inorganic. For example, Vitamin C is an organic, water-soluble cofactor; Vitamin A is an organic, lipid-soluble cofactor; and Zinc is an inorganic cofactor.

The ENTITY SET concept has been reevaluated, and it is no longer a type of entity. Conceptually speaking, it groups a set of entities that can be used indistinctly in a process because they play an equivalent role, but it is not a subtype of entity. We have also linked it to the event in which the entities are interchangeable.

We have characterized the concept of LOCATION to indicate where a protein functions (e.g., the CFTR protein, which is known to cause Cystic fibrosis [29], functions in the pancreas, the lung, and the cell membrane, among other locations). Our initial approach was to change the link between the location and the protein. We moved it upwards in the hierarchy so that the location is linked to the entity. However, we detected some situations that forced us to reevaluate this approach. There are entities that have i) different roles in the same location, ii) the same role in different locations, iii) different roles in different locations, or iv) the same role in different locations. Moreover, the role of an entity changes depending on the process where it intervenes. Consequently, we decided to link the location to the takes_part class. Now, entities play a given role for a process in a specific location, which is a much more precise conceptualization of how the biological components of life work.

When we dive into the pathway view, we observe that the representation of the result of an event is limited to the output entities of the processes that compose it, which is a useful and correct approach, but it is not complete. There are events whose result is not limited to a set of entities, but that also include a change in our body. For example, in the transportation of bicarbonate through the ion channels event, we would model *bicarbonate* as a basic entity that has an *output* role in the *extracellular medium* location. However, this event is related to additional changes in our bodies that are not linked to entities, like maintaining a normal pH regulation.

In the CSG, we have included the EFFECT class, which is characterized by a name and a description. This class allows the pathway view to be enriched in order to represent the changes produced by the events that do not involve entities. To clarify, maintaining a normal pH regulation is an effect of the transportation of bicarbonate through the ion channels event with this approach. Effects complement the changes of entities in the processes that compose an event.

Our last concern is the role that entities (especially genes and their products) play regarding a phenotype. For example, the CD22 gene codes the B-cell receptor CD22 protein. The default role of this protein is to disrupt microglia cell function, but an altered role causes detritus accumulation in the brain, which produces Alzheimer's [43], [63]. Being able to annotate such roles of entities is crucial in establishing genotype-phenotype associations and in being able to precisely characterize how our body works. It also has important implications in precision medicine.

Our approach for characterizing these entity roles consists of defining the ENTITYROLE concept, which is characterized by a name and a description. Additionally, we indicate the level of certainty with which the role has been established. This attribute allows us to discriminate computationally predicted roles from manually revised ones. The notion of ENTITYROLE establishes entity-phenotype pairs with a descriptor. There are two types of roles. The first role is the DEFAULTROLE, e.g., the disruption of microglial cells. The second role is the ALTEREDROLE, e.g., the accumulation of detritus. Altered roles can be linked to the variants that cause them if this knowledge is available. In other words, we can identify the genetic variants that disrupt the normal function of an entity regarding a phenotype, and these disruptions include a level of certainty that provides additional context.

These changes have allowed us to correct the five conceptual weaknesses that were identified by the group of biologists and bioinformaticians. The concept of the entity and its specialization classes has been completely reworked. We have complemented the representation of events by including the location where they are produced. The representation of events has been expanded to represent additional changes in the organism that are not related to entities. The high-level role that entities have in phenotypes has been included to provide an additional layer of knowledge and contextual information.

In summary, the number of modifications applied to the model are the following: ten classes have been created, and three classes have been reevaluated.

IV. CONCLUSION AND FUTURE WORK

Precision medicine is crucial in providing a more correct and individualized diagnosis, treatment, and prevention of human disease. The challenges that precision medicine presents can be overcome (and their outcomes greatly improved) by using techniques that help experts to acquire a deep understanding of the context as well as the ability to communicate effectively. We have shown that CM techniques are positioned as one of the most appropriate tools for improving precision medicine. In this work, we demonstrate the benefits of using CM. First, it allows us to communicate effectively with the stakeholders. Second, it provides an ontological commitment that is shared by stakeholders. Consequently, more efficient

exploitation of the information can be achieved, which leads to more precise and helpful results.

Our CSG must be updated continuously in order to deal with the complexity and variability of the domain. In this work, we have presented the experience and the knowledge that we have accumulated during the elaboration of the different versions of the CSG. The initial version, which was centered on the human genome (CSHG), focused on creating a semantic and content description of the most relevant concepts of the domain based on a gene-centered vision. Version 2 changed to a chromosome-centered one to simplify the CS and to provide a more flexible approach. Version 3 increased its flexibility by expanding the interactions among the different parts of the CS and including new, sound domain information that was missing. The first version of the CSG provided us with a new CS that can work with any eukaryotic species. With the second version of the CSG, we have improved and expanded several views in order to boost its potential application in precision medicine.

We also want to emphasize that these changes respond to real domain-user needs that were requested. The changes described in Section III-A allow proteins to be studied in more detail, including their physical structure and their interactions with the body. The changes described in Section III-B improve how variants pathogenicity is handled, which eases data management and potentially improves the selection of relevant variants. Finally, the changes described in Section III-C provide a more detailed representation of human pathways, which opens a wide range of analyses regarding their function and how they are linked to diseases.

The previous versions of the CSHG and the CSG have been validated by i) domain expert validation and ii) design and implementation of model-driven development (MDD) tools. We aim to follow the same approach with this new version of the CSG. On the one hand, we have met several times with domain experts (i.e., physicians and bioinformatics) while generating this new working version. In these meetings, we gathered their requirements and discussed our proposed solutions. Thus, we can assert that the CSG has been validated from a theoretical point of view. On the other hand, we plan to develop a platform that implements a genomic pipeline [41] following an MDD approach with this conceptual schema as future work.

Other future work is oriented towards enriching the model semantics and introducing new relevant concepts. First, we will expand and characterize the concept of location further to differentiate between more general locations, such as organs, and more specific locations, such as the components of a cell. Second, we will start studying how to include somatic variants in the CSG and their link to cancer.

ACKNOWLEDGMENT

The authors would like to thank the members of the PROS Research Center Genome group for fruitful discussions regarding the application of Conceptual Modeling in the medical field.

REFERENCES

[1] *ClinVar Variant Details (VCV000039614.14)*. Accessed: Oct. 21, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/variation/39614/

[2] *ClinVar Variants (GRCh38 Assembly)*. Accessed: Jan. 1, 2021. [Online]. Available: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz

[3] A. Agusti, E. Bel, M. Thomas, C. Vogelmeier, G. Brusselle, S. Holgate, M. Humbert, P. Jones, P. G. Gibson, J. Vestbo, R. Beasley, and I. D. Pavord, "Treatable traits: Toward precision medicine of chronic airway diseases," *Eur. Respiratory J.*, vol. 47, no. 2, pp. 410–419, Feb. 2016.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, no. 1, pp. 25–29, 2000.

[5] P. J. Beurton, R. Falk, and H. J. Rheinberger, *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[6] M. S. Boguski and A. R. Jones, "Neurogenomics: At the intersection of neurobiology and genome sciences," *Nature Neurosci.*, vol. 7, no. 5, pp. 429–433, May 2004.

[7] J. P. Casas, T. Shah, A. D. Hingorani, J. Danesh, and M. B. Pepys, "C-reactive protein and coronary heart disease: A critical review," *J. Internal Med.*, vol. 264, no. 4, pp. 295–314, Oct. 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2796.2008.02015.x

[8] T. Chaijarasphong, R. J. Nichols, K. E. Kortright, C. F. Nixon, P. K. Teng, L. M. Oltrogge, and D. F. Savage, "Programmed ribosomal frameshifting mediates expression of the α-carboxysome," *J. Mol. Biol.*, vol. 428, no. 1, pp. 153–164, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283615006798

[9] T. K. Chaudhuri and S. Paul, "Protein-misfolding diseases and chaperone-based therapeutic approaches," *FEBS J.*, vol. 273, no. 7, pp. 1331–1349, Apr. 2006.

[10] L. M. L. Delcambre, S. W. Liddle, O. Pastor, and V. C. Storey, "A reference framework for conceptual modeling," in *Conceptual Modeling* (Lecture Notes in Computer Science), J. C. Trujillo, K. C. Davis, X. Du, Z. Li, T. W. Ling, G. Li, and M. L. Lee, Eds. Cham, Switzerland: Springer, 2018, pp. 27–42, doi: 10.1007/978-3-030-00847-5_4.

[11] N. M. Delzenne, "Oligosaccharides: State of the art," *Proc. Nutrition Soc.*, vol. 62, no. 1, pp. 177–182, Feb. 2003. [Online]. Available: https://www.cambridge.org/core/journals/proceedings-of-the-nutrition-society/article/oligosaccharides-state-of-the-art/68515DB878EB478FDF43B8522ED17CAF

[12] J. D. Dombrauckas, B. D. Santarsiero, and A. D. Mesecar, "Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis," *Biochemistry*, vol. 44, no. 27, pp. 9417–9429, Jul. 2005.

[13] L. R. Engelking, *Textbook of Veterinary Physiological Chemistry*. Amsterdam, The Netherlands: Elsevier, 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/C20100660470

[14] J. A. Fifita, K. L. Williams, V. Sundaramoorthy, E. P. Mccann, G. A. Nicholson, J. D. Atkin, and I. P. Blair, "A novel amyotrophic lateral sclerosis mutation in OPTN induces ER stress and Golgi fragmentation *in vitro*," *Amyotrophic Lateral Sclerosis Frontotemporal Degener.*, vol. 18, nos. 1–2, pp. 126–133, Jan. 2017.

[15] A. García S. and J. C. Casamayor, "On how to generalize species-specific conceptual schemes to generate a species-independent conceptual schema of the genome," *BMC Bioinf.*, vol. 22, no. 13, p. 353, Sep. 2021.

[16] S. A. García, A. L. Palacio, J. F. R. Román, J. C. Casamayor, and O. Pastor, "Towards the understanding of the human genome: A holistic conceptual modeling approach," *IEEE Access*, vol. 8, pp. 197111–197123, 2020.

[17] S. A. García, J. F. R. Román, J. C. Casamayor, and O. Pastor, "Towards an effective and efficient management of genome data: An information systems engineering perspective," in *Information Systems Engineering in Responsible Information Systems* (Lecture Notes in Business Information Processing), C. Cappiello and M. Ruiz, Eds. Cham, Switzerland: Springer, 2019, pp. 99–110.

[18] Gene Ontology Consortium, "The gene ontology resource: Enriching a GOld mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, 2021.

[19] K. P. Giese and K. Mizuno, "The roles of protein kinases in learning and memory," *Learn. Memory*, vol. 20, no. 10, pp. 540–552, Oct. 2013. [Online]. Available: http://learnmem.cshlp.org/content/20/10/540

[20] W. T. Godbey, "Proteins," in *An Introduction to Biotechnology*. Sawston, U.K.: Woodhead Publishing, 2014, ch. 2, pp. 9–33. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781907568282000022
[21] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nature Rev. Genet.*, vol. 17, no. 6, pp. 333–351, Jun. 2016.
[22] W. M. T. Groenestege, S. Thébault, J. van der Wijst, D. van den Berg, R. Janssen, S. Tejpar, L. P. van den Heuvel, E. van Cutsem, J. G. Hoenderop, N. V. Knoers, and R. J. Bindels, "Impaired basolateral sorting of pro-EGF causes isolated recessive renal hypomagnesemia," *J. Clin. Invest.*, vol. 117, no. 8, pp. 2260–2267, Aug. 2007.
[23] S. M. Harrison, E. R. Riggs, D. R. Maglott, J. M. Lee, D. R. Azzariti, A. Niehaus, E. M. Ramos, C. L. Martin, M. J. Landrum, and H. L. Rehm, "Using ClinVar as a resource to support variant interpretation," *Current Protocols Hum. Genet.*, vol. 89, no. 1, pp. 8.16.1–8.16.23, Apr. 2016.
[24] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016.
[25] R. I. Doğan *et al.*, "Overview of the biocreative VI precision medicine track: Mining protein interactions and mutations for precision medicine," *Database*, vol. 2019, Jan. 2019, Art. no. bay147.
[26] G. John, S. Nagarajan, P. K. Vemula, J. R. Silverman, and C. K. S. Pillai, "Natural monomers: A mine for functional and sustainable materials—Occurrence, chemical modification and polymerization," *Prog. Polym. Sci.*, vol. 92, pp. 158–209, May 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0079670018300273
[27] A. Katsnelson, "Momentum grows to make 'personalized' medicine more 'precise,'" *Nature Med.*, vol. 19, no. 3, p. 249, Mar. 2013.
[28] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm, "Function of alternative splicing," *Gene*, vol. 514, no. 1, pp. 1–30, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037811191009791
[29] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm, "Function of alternative splicing," *Gene*, vol. 514, no. 1, pp. 1–30, Feb. 2013, doi: 10.1016/j.gene.2012.07.083.
[30] A. R. Kornblihtt, "Promoter usage and alternative splicing," *Current Opinion Cell Biol.*, vol. 17, no. 3, pp. 262–268, Jun. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0955067405000566
[31] I. R. König, O. Fuchs, G. Hansen, E. V. Mutius, and M. V. Kopp, "What is precision medicine?" *Eur. Respiratory J.*, vol. 50, no. 4, pp. 1–12, 2017.
[32] A. Leon and O. Pastor, "Towards a shared, conceptual model-based understanding of proteins and their interactions," *IEEE Access*, vol. 9, pp. 73608–73623, 2021.
[33] J. A. Leopold and J. Loscalzo, "Emerging role of precision medicine in cardiovascular disease," *Circulat. Res.*, vol. 122, no. 9, pp. 1302–1315, Apr. 2018.
[34] A. León, S. A. García, M. Costa, A. V. Ribelles, and O. Pastor, "Evolution of an adaptive information system for precision medicine," in *Intelligent Information Systems*, vol. 424, S. Nurcan and A. Korthaus, Eds. Cham, Switzerland: Springer, 2021, pp. 3–10.
[35] D. Li, F. L. Mastaglia, S. Fletcher, and S. D. Wilton, "Precision medicine through antisense oligonucleotide-mediated exon skipping," *Trends Pharmacol. Sci.*, vol. 39, no. 11, pp. 982–994, Nov. 2018.
[36] S. McGinn and I. G. Gut, "DNA sequencing—Spanning the generations," *New Biotechnol.*, vol. 30, no. 4, pp. 366–372, 2013.
[37] W. Miller, K. D. Makova, A. Nekrutenko, and R. C. Hardison, "Comparative genomics," *Annu. Rev. Genomics Hum. Genet.*, vol. 5, no. 1, pp. 15–56, 2004.
[38] J. Mylopoulos, "Conceptual modelling and Telos," Univ. Toronto, Toronto, OH, USA, Tech. Rep., 1992, pp. 49–68.
[39] J. Nilsson, U. Rüetschi, A. Halim, C. Hesse, E. Carlsohn, G. Brinkmalm, and G. Larson, "Enrichment of glycopeptides for glycan structure and attachment site identification," *Nature Methods*, vol. 6, no. 11, pp. 809–811, Nov. 2009.
[40] K. Nykamp, M. Anderson, M. Powers, J. Garcia, B. Herrera, Y. Y. Ho, Y. Kobayashi, N. Patil, J. Thusberg, M. Westbrook, and S. Topper, "Sherloc: A comprehensive refinement of the ACMG–AMP variant classification criteria," *Genet. Med.*, vol. 19, no. 10, pp. 1105–1117, 2017.
[41] A. L. Palacio and O. P. López, "Smart data for genomic information systems: The SILE method," *Complex Syst. Informat. Model. Quart.*, vol. 17, no. 17, pp. 1–23, Dec. 2018.
[42] O. Pastor, A. M. Levin, M. Celma, J. C. Casamayor, A. Virrueta, and L. E. Eraso, "Model-based engineering applied to the interpretation of the human genome," in *The Evolution of Conceptual Modeling* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6520. Berlin, Germany: Springer, 2011, pp. 306–330.
[43] J. V. Pluvinage, M. S. Haney, B. A. H. Smith, J. Sun, T. Iram, L. Bonanno, L. Li, D. P. Lee, D. W. Morgens, A. C. Yang, S. R. Shuken, D. Gate, M. Scott, P. Khatri, J. Luo, C. R. Bertozzi, M. C. Bassik, and T. Wyss-Coray, "CD22 blockade restores homeostatic microglial phagocytosis in ageing brains," *Nature*, vol. 568, no. 7751, pp. 187–192, Apr. 2019.
[44] J. L. Reiter and N. J. Maihle, "A 1.8 kb alternative transcript from the human epidermal growth factor receptor gene encodes a truncated form of the receptor," *Nucleic Acids Res.*, vol. 24, no. 20, pp. 4050–4056, Oct. 1996.
[45] J. F. R. Román, Ó. Pastor, J. C. Casamayor, and F. Valverde, "Applying conceptual modeling to better understand the human genome," in *Conceptual Modeling* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9974. Cham, Switzerland: Springer-Verlag, 2016, pp. 404–412.
[46] T. Rezaie, A. Child, R. Hitchings, G. Brice, L. Miller, M. Coca-Prados, E. Héon, T. Krupin, R. Ritch, D. Kreutzer, R. P. Crick, and M. Sarfarazi, "Adult-onset primary open-angle glaucoma caused by mutations in optineurin," *Science*, vol. 295, no. 5557, pp. 1077–1079, Feb. 2002.
[47] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology," *Genet. Med., Off. J. Amer. College Med. Genet.*, vol. 17, no. 5, pp. 405–423, May 2015.
[48] D. J. Rigden and X. M. Fernández, "The 27th annual nucleic acids research database issue and molecular biology database collection," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1–D8, Jan. 2020.
[49] R. Román and J. Fabián, "Design and development of a genomic information system based on a holistic conceptual model of the human genome," M.S. thesis, Dept. Comput. Syst. Comput., Polytech. Univ. Valencia, Valencia, Spain, 2018.
[50] S. A. García and J. C. Casamayor, "Towards the generation of a species-independent conceptual schema of the genome," in *Advances in Conceptual Modeling* (Lecture Notes in Computer Science), vol. 12584, G. Grossmann and S. Ram, Eds. Vienna, Austria: Springer, Nov. 2020, pp. 61–70.
[51] T. Sanavia, G. Birolo, L. Montanucci, P. Turina, E. Capriotti, and P. Fariselli, "Limitations and challenges in protein stability prediction upon genome variations: Towards future applications in precision medicine," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1968–1979, Jul. 2020.
[52] J. R. Savinainen, J. Z. Patel, T. Parkkari, D. Navia-Paldanius, J. J. T. Marjamaa, T. Laitinen, T. Nevalainen, and J. T. Laitinen, "Biochemical and pharmacological characterization of the human lymphocyte antigen B-associated transcript 5 (BAT5/ABHD16A)," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e109869.
[53] B. Shan, J. Gerez, M. Haedo, M. Fuertes, M. Theodoropoulou, M. Buchfelder, M. Losa, G. K. Stalla, E. Arzt, and U. Renner, "RSUME is implicated in HIF-1-induced VEGF-A production in pituitary tumour cells," *Endocrine-Related Cancer*, vol. 19, no. 1, pp. 13–27, Feb. 2012.
[54] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?" *PLOS Biol.*, vol. 13, no. 7, Jul. 2015, Art. no. e1002195.
[55] T. UniProt Consortium, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.
[56] C. Touriol, S. Bornes, S. Bonnal, S. Audigier, H. Prats, A.-C. Prats, and S. Vagner, "Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons," *Biol. Cell*, vol. 95, nos. 3–4, pp. 169–178, May 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1016/S0248-4900%2803%2900033-9
[57] L. P. Tripathi, Y. A. Chen, K. Mizuguchi, and Y. Murakami, "Network-based analysis for biological discovery," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford, U.K.: Academic, 2019, pp. 283–291. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128096338206742

[58] M. Vihinen, "Problems in variation interpretation guidelines and in their implementation in computational tools," *Mol. Genet. Genomic Med.*, vol. 8, no. 9, Sep. 2020, Art. no. e1206. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mgg3.1206
[59] J. L. Vittitow and T. Borrás, "Expression of optineurin, a glaucoma-linked gene, is influenced by elevated intraocular pressure," *Biochem. Biophys. Res. Commun.*, vol. 298, no. 1, pp. 67–74, Oct. 2002.
[60] M. E. Wernette-Hammond, S. J. Lauer, A. Corsini, D. Walker, J. M. Taylor, and S. C. Rall, "Glycosylation of human apolipoprotein E: The carbohydrate attachment site is threonine 194," *J. Biol. Chem.*, vol. 264, no. 15, pp. 9094–9101, 1989.
[61] F. A. Witzmann and R. A. Grant, "Pharmacoproteomics in drug development," *Pharmacogenomics J.*, vol. 3, no. 2, pp. 69–76, Jan. 2003. [Online]. Available: https://www.nature.com/articles/6500164
[62] A. A. Zamyatnin, A. S. Borchikov, M. G. Vladimirov, and O. L. Voronina, "The EROP-Moscow oligopeptide database," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D261–D266, Jan. 2006, doi: 10.1093/nar/gkj008.
[63] X. Zhang, R. Wang, D. Hu, X. Sun, H. Fujioka, K. Lundberg, E. R. Chan, Q. Wang, R. Xu, M. E. Flanagan, A. A. Pieper, and X. Qi, "Oligodendroglial glycolytic stress triggers inflammasome activation and neuropathology in Alzheimer's disease," *Sci. Adv.*, vol. 6, no. 49, Dec. 2020, Art. no. eabb8680.
[64] X. D. Zhang, "Precision medicine, personalized medicine, omics and big data: Concepts and relationships," *J. Pharmacogenomics Pharmacoproteomics*, vol. 6, no. 2, 2015, Art. no. e144.

ALBERTO GARCÍA S. is currently pursuing the Ph.D. degree with the VRAIN Research Institute, Universitat Politècnica de València, Spain. He is also studying how to improve genome data analysis.

ANA LEÓN PALACIO is currently a Postdoctoral Researcher with the VRAIN Research Institute, Universitat Politècnica de València, Spain. She works on how to provide a systematic approach to efficiently manage genomic data.

JOSE FABIÁN REYES ROMÁN is currently a Postdoctoral Researcher with the VRAIN Research Institute, Universitat Politècnica de València, Spain. He develops solutions on the genomic domain from a conceptual modeling perspective.

JUAN CARLOS CASAMAYOR is currently an Associate Professor and a Researcher with the VRAIN Research Institute, Universitat Politècnica de València, Spain. His research interests include databases and information systems design.

OSCAR PASTOR is currently a Full Professor and the Director of the VRAIN Research Institute, Universitat Politècnica de València, Spain. He is also leading a multidisciplinary project linking information systems and bioinformatics to designing and implementing tools for conceptual modeling-based interpretation of the human genome information.