IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 7, 2021, accepted May 10, 2021, date of publication May 13, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3080040

Towards a Shared, Conceptual Model-Based Understanding of Proteins and Their Interactions

ANA LEON AND OSCAR PASTOR, (Member, IEEE)
Research Center on Software Production Methods (PROS), Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Ana Leon (aleon@pros.upv.es)

This work was supported in part by the Spanish State Research Agency under Grant TIN2016-80811-P, and in part by the Generalidad Valenciana under Grant PROMETEO/2018/176, co-financed with ERDF.

ABSTRACT Understanding the human genome is a big research challenge. The huge complexity and amount of genome data require extremely effective and efficient data management policies. A first crucial point is to obtain a shared understanding of the domain, which becomes a very hard task considering the number of different genome data sources. To make things more complicated, those data sources deal with different parts of genome-based information: we not only need to understand them well, but also to integrate and intercommunicate all the relevant information. The protein perspective is a good example: rich, well-known repositories such as UniProt provide a lot of valuable information that it is not easy to interpret and manage when we want to generate useful results. Proteomes and basic information, protein-protein interaction, protein structure, protein processing events, protein function, etc. provide a lot of information is that needs to be conceptually characterized and delimited. To facilitate the essential common understanding of the domain, this paper uses the case of proteins to analyze the data provided by Uniprot in order to make a sound conceptualization work for identifying the relevant domain concepts. A conceptual model of proteins is the result of this conceptualization process, explained in detail in this work. This holistic conceptual model of proteins presented in this paper is the result of achieving a precise ontological commitment. It establishes concepts and their relationships that are significant in order to have a solid basis to efficiently manage relevant genome data related to proteins.

INDEX TERMS Conceptual modeling, genomics, proteins.

I. INTRODUCTION

Clinical disease states reflect the interaction of a myriad of genetic and environmental contributions. In this context, a major challenge is to develop information systems and algorithms that can describe this complexity in order to facilitate an understanding of the disease mechanisms as well as to guide the development and application of therapies. Unfortunately, current research mainly focuses only on very specific parts of the domain (genes, variants, pathways, proteins, phenotypes, etc.). When individually considered, their complexity is clear when accessing real world data provided by their associated data sources of reference (such as the UniProt for the protein case, which is the working domain analyzed in this paper).

The ability to perform integrated analysis making use of multiple forms of complex data to uncover patterns and trends in ways that traditional methods cannot is an issue of critical relevance that can transform biology from an observational molecular science to a data-intensive quantitative genomic science [1]. To this end, a set of steps must be precisely accomplished according to [2]:

1) Obtain a shared understanding of the domain under consideration.
2) Understand what task is to be done and select the right scope.
3) Collect the right data.
4) Select the analysis techniques that deliver results (e.g., Artificial Intelligence or Data Science techniques).
5) Generate good explanations.
6) Evolve the solution over time as more knowledge is acquired.

This work focuses on describing how the first step can be achieved by analyzing the conceptual precision of the main concepts that should constitute the ontological commitment

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

that is strictly required when studying an important area of research: the role that proteins play in the different functions carried out within the cell of any living system. To such aim, we present a conceptual schema that focuses the structure of the proteins and their function, according to the main concepts considered in the UniProtKB database. We used this database because it is one of the most reputable and complete sources about proteins, and widely used by the research community. Since the work focuses on the proteins once they have been formed, concepts associated to gene expression (e.g., methylation and regulatory elements) are not described in this work. Such concepts are considered in a wider view of the schema under development where the structure of the DNA, in addition to the elements that take part on the transcription process and the gene expression are described.

Obtaining a shared understanding of the domain under consideration requires having a precise conceptual model of reference that is ontologically well-grounded (precisely identifying the relevant concepts of the domain) and accurately interpreting the data that are managed at the real, biological practical scope, in order to achieve an adequate, effective, and useful data representation.

A correct interpretation of how proteins work is essential to advance in the challenge of understanding the human genome. Proteins are the result of a complex process, named gene expression, in which the DNA sequence of an organism is transcribed into RNA and translated into a functional product. During the gene expression, the information in the DNA of every cell is converted into small, portable RNA messages. These messages travel from the cell nucleus to the ribosomes where they are ''read'' to make specific proteins. These proteins have specific functions key for the correct function of the cell, such as regulators of chemical reactions, interaction with other proteins to produce complexes, or transport chemical elements inside and outside the cell. When an alteration in the DNA sequence occurs, the transcription and translation processes can be affected and consequently a nonfunctional protein can be produced. In the worst case, even the entire protein can be missing. This situation alters the equilibrium of the processes that occur within the cell and lead to disease. Understanding how proteins are produced and the specific functions in which they are involved can help researchers to understand the mechanisms of disease and consequently improve the diagnosis and treatment. Genome sequencing can identify the variants carried out by an individual that make them susceptible to disease, but it does not reveal how the disease is caused. The protein perspective is strictly required to understand it, and this work describes how to get a solid, ontologically well-grounded understanding of such complex domain. A conceptual model for proteins has been carefully developed taking the UniProtKB database as data source and explaining in detail the problems that have been faced and their corresponding solutions.

This paper is structured as follows. In Section 2, we introduce the concepts that are used to understand the structure, function, and involvement in disease of proteins. Using these concepts as a basis, we explain the conceptualization process that results in the description of the corresponding conceptual model. In Section 3, we describe the importance of having a sound ontological commitment and the challenges found during the conceptualization process. We conclude with a discussion of future research directions in Section 4.

II. CONCEPTUAL MODELING OF PROTEIN INFORMATION

The information needed to precisely characterize the concept of protein in all of its relevant dimensions is very complex. Proteins are molecules made up of amino acids, linked together in a very specific sequence, that carry out different functions within the cell. These molecules can catalyze chemical reactions, be part of the structure of the cell, or act as signals [3]. To correctly manage all the available information about proteins, an immediate need emerges: having adequate, accurate, consistent, and manageable data sources. In this working domain, UniProt is a widely accepted and used repository. We start our work by introducing its structure in order to delimit the conceptual context of our modeling task.

The Universal Protein Resource (UniProt) [4] is a repository of protein sequences and annotation data from different organisms that emerged from the collaboration between the European Bioinformatics Institute (EMBL-EBI)(https://www.ebi.ac.uk/), the SIB Swiss Institute of Bioinformatics (https://www.sib.swiss/), and the Protein Information Resource (PIR)(https://proteininformationresource.org/). The UniProt repository is supported by four main databases: the UniProt Knowledgebase (UniProtKB) [5], the UniProt Reference Clusters (UniRef) [6], the Proteomes(https://www.uniprot.org/proteomes/) database, and the UniProt Archive (UniParc) [7]. The UniProtKB is the central database where functional information on proteins and annotation data (produced either manual or automatically) are collected. The UniRef database provides clustered sets of sequences (including isoforms). The Proteomes database collects information about sets of proteins whose genomes have been completely sequenced. UniParc is the sequence archive that contains most of the publicly available protein sequences in the world. These four databases provide a complete coverage of the sequence space (see Fig.1). The fact of having these four dimensions or databases is a clear indicator of the high level of complexity that is associated to the concept of protein.

The conceptual characterization and interconnection of these four databases is a significant challenge that should be faced in order to assess potential inconsistencies, redundancies, obsolete information, and other quality data problems that could seriously affect any data management process. To achieve this longer-term goal, the very first step is to characterize the fundamental, basic protein information. This initial, sound conceptual characterization task is the main goal of this work. It is indeed the only way to ensure the shared understanding of the domain that we want to obtain, and to facilitate a valuable and fruitful data exploitation strategy.

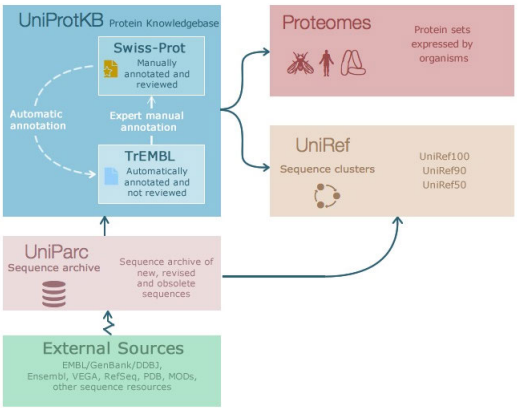

FIGURE 1. Structure of the UniProt repository [4].

To achieve this goal, this work is specifically based on the information provided by the UniProtKB database.

The conceptual model of proteins that we are going to elaborate will facilitate sharing a common, holistic, semantically precise perspective of the data provided by this database. The following sections focus on introducing the main concepts used to define proteins, including their structure, their function, and the sequence changes that can lead to disease.

A. PROTEOMES AND BASIC INFORMATION ABOUT PROTEINS

The entire set of proteins that can be expressed by the genome of an organism is called proteome [8]. Each proteome is made of a set of components that contain the genes in charge of codifying the different proteins. Additional information about proteomes can be found in the Proteomes database. The basic concepts that are required to start the conceptualization process are organism, gene, and protein.

1) ORGANISM

The organism is the source of the protein sequences that are part of the proteome. It is usually described by a Latin scientific name followed (optionally) by the English common name and a synonym if available (e.g., *Cardamine pratensis* (Cuckoo flower) whose synonym is Alpine bitter cress). The UniProtKB database also provides the taxonomic classification (lineage), which is a hierarchy that represents the relative level of a group of organisms. If the organism described is a virus, the specific organism or taxonomic group that is susceptible to infection (called host) must also be specified.

2) GENE

Each protein can be codified by one or more genes and produced by different cellular components (organelles) in the cell (e.g., the hydrogenosome, the mitochondrion, the nucleus, etc.). To name the genes, the UniProtKB database uses the acronym or official symbol (e.g., PAH). Genes are also represented by the naming systems used to sequentially assign an identifier to each gene of a chromosome (known as Ordered Locus Names) and the list of names that are temporarily attributed to an Open Reading Frame[1] by a sequencing project (known as ORF Names).

3) PROTEIN

The information about proteins is basically composed by a unique identifier, which is specific to the UniProtKB database (e.g., P00439), and a name (e.g., Phenyl-alanine-4-hydroxylase). The names of the proteins have evolved over time and some of them have become obsolete. Nevertheless, in some scientific literature and databases these names are still in use and UniProtKB provides a complete list for the unambiguous identification of the associated proteins. These names include a recommended name, a short name, and a list of alternative names. Proteins can also be grouped into families that descend from a common ancestor and typically have similar three-dimensional structures, functions, and significant sequence similarity.

After the identification of the concepts that make up the basic information about proteomes, genes, and proteins, the subsequent conceptualization task leads to the representation shown in Fig. 2. The resulting conceptual model has been described using a UML Class Diagram that includes the classes, with their attributes and relationships. It models all of the relevant information that has been introduced.

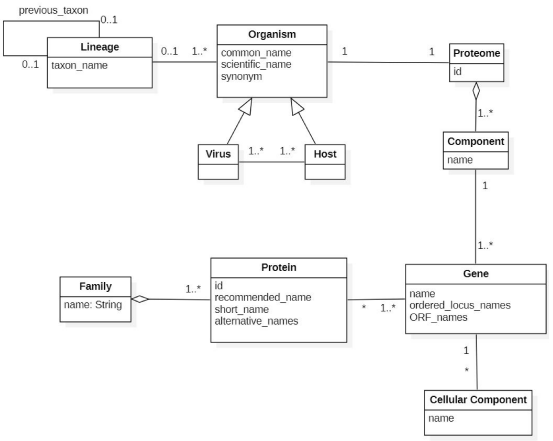

FIGURE 2. Conceptual model that represents the basic information about proteomes, genes, proteins, and organisms.

Proteins rarely act alone since many molecular processes within a cell are carried out by complex components thanks to the ability of proteins to interact with each other. Therefore, Protein-Protein interactions are the next conceptual step that have been analyzed in our work.

B. PROTEIN-PROTEIN INTERACTIONS

Protein–Protein interactions (PPIs) are the specific physical contacts between proteins that occur by selective molecular docking in a specific biological context [9]. The complete

[1]Open Reading Frame: A continuous stretch of codons (including start and end codons) that can be translated.

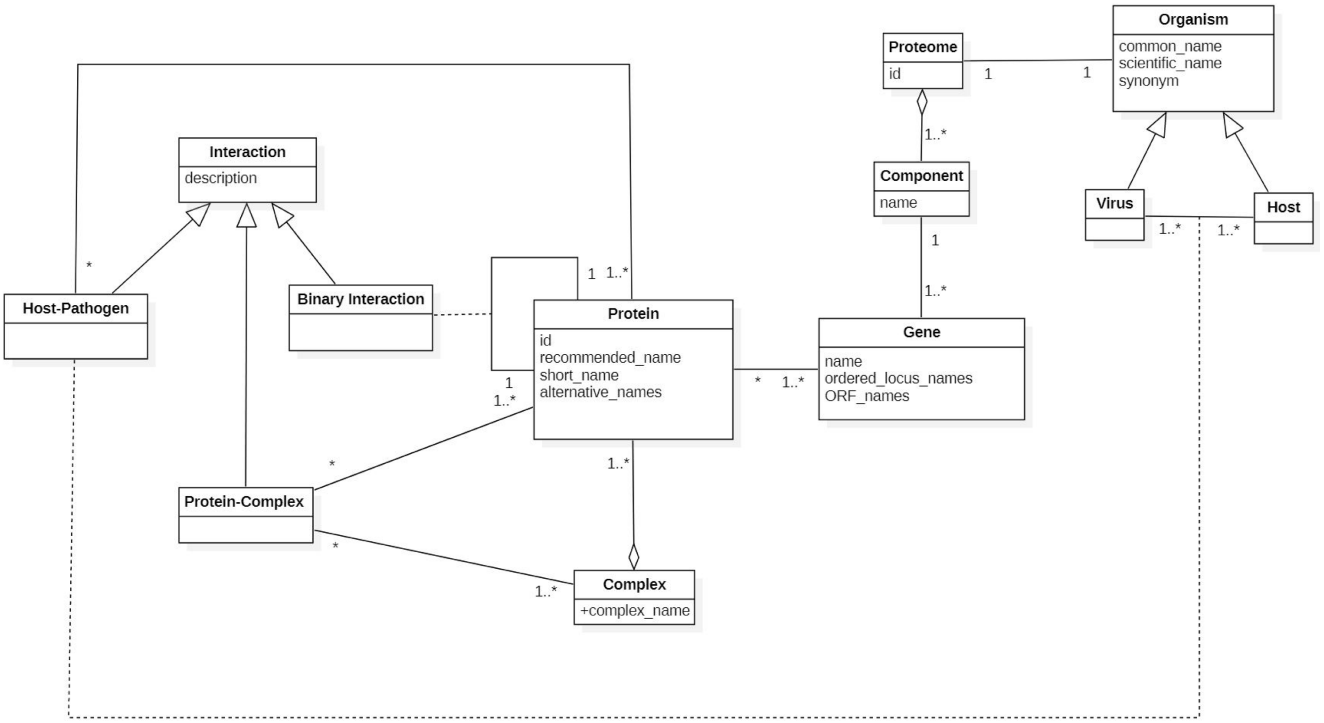

FIGURE 3. Conceptual model that represents the interactions among proteins.

map of protein interactions that can occur in a living organism is called interactome. Extending the initial conceptual model introduced in Fig. 2 with this information is the next modeling step. It is a crucial step since the occurrence of aberrant PPIs is the basis of multiple aggregation-related diseases, such as Creutzfeldt–Jakob and Alzheimer's.

The UniProtKB database provides information about binary-protein interactions[2] (extracted from the IntAct database [10]), host-pathogen interactions, and protein-complex interactions.[3] In the conceptual model shown in Fig. 3, the PPIs are represented by the Interaction association class. The description of the interaction is represented by the corresponding attribute. This class specializes into the Binary, the Host-Pathogen, and the Protein-Complex classes (the three types of interactions that are considered). For host-pathogen interactions, the virus and the host that are related are represented through the association class connected to the association defined between the corresponding Virus and Host classes. For protein-complex interactions, the association with the Complex class provides the name of all of the proteins that are part of the complex that participates in the interaction being modeled.

Another essential concept for defining PPIs is the biological context. The interactions depend on cell type, developmental stage, environmental conditions, protein modifications, etc. This information is provided by specialized databases and repositories such as DIP [11], IntAct, and MINT [12]. The detailed description of these repositories is the natural next step, which is viewed as future work to be done once the structural conceptual model presented in this work is complete. This future work will be the way to achieve a complete understanding of PPIs and to design better ways for analyzing and interpreting interactions.

The shape of a protein is essential in order to understand its function because it determines whether the protein can interact with other molecules. Characterizing the protein structure is the next modeling step.

C. PROTEIN STRUCTURE

This section extends the conceptual model by describing the information associated to the different structural levels of a protein and how it can be modeled. Proteins are complex and irregular structures that can be described using four levels [13]:

1) The primary structure is the sequence of amino acids that make up the protein.
2) The secondary structure arises from interactions between nearby amino acids as the primary structure starts to fold into its functional three-dimensional form.
3) The tertiary structure is the overall three-dimensional shape once all of the secondary structure elements have folded together with each other.
4) The quaternary structure represents how its sub-units are oriented and arranged with respect to one another.

[2]Binary-protein interactions: direct physical interactions between proteins.

[3]Protein-complex interactions: physical interactions among groups of proteins, without pair-wise determination of protein partners

A sound knowledge of protein structure is essential to understand their function and how to design, inhibit, and activate proteins.

1) PRIMARY STRUCTURE

Proteins are made of a linear sequence of amino acids that is the result of the translation of the DNA. This sequence is called the primary structure. Due to different bio-logical events (alternative promoter usage, alternative splicing, alternative initiation, and ribosomal frameshifting), a gene can be translated into similar amino acid sequences, leading to the presence of different versions of the same protein. These versions are called isoforms [14]. Each protein sequence is characterized by an identifier (the primary accession number of the protein, followed by a dash and a number), a name, and a set of synonyms. Additional relevant properties include its length, its molecular mass in Daltons, the last update, a checksum used to track sequence updates, and other additional information. The protein sequence displayed by default on the UniProtKB website is the isoform to which all positional annotations refer to, which is called the canonical sequence. The UniProtKB also provides information about the completeness of the canonical sequence (sequence status), describing whether it is complete or fragmented. Any severe discrepancy between the canonical sequence and other available sequences (e.g., the ones reported in a paper or predicted somewhere else) are described in a note, called sequence caution, which includes the reason that justifies its existence along with the identifier of the discrepant sequence.

Protein sequences are represented in the conceptual model using the Isoform class, which specializes into the Canonical class. Each protein is associated to only one canonical sequence and additionally can be associated to many isoforms. The mechanism(s) that produce(s) the different isoforms are described by the association class Event, as shown in Fig. 4.

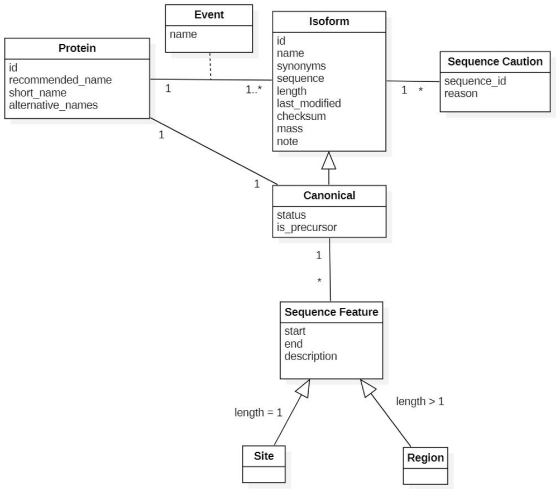

FIGURE 4. Conceptual model that represents the canonical sequence and the different isoforms of a protein.

Sometimes, the canonical sequence requires processing or post-translational modifications (PTMs) to become mature. In this case, the canonical sequence is known as precursor. In the conceptual model, the boolean attribute *is precursor* of the Canonical class allows the description of the canonical sequence as a precursor. The PTMs are explained in detail in Subsection E.

Along the protein sequence, there are interesting locations known as sequence features, which are classified as sites and regions depending on their length. Sequence features are represented in the conceptual model by the Sequence Feature class, which has three main attributes: description, start, and end. This class specializes into the Site and the Region classes, with a restriction of length. For sequences of one amino acid, the start and the end must obviously be the same. Due to the complexity and importance of these sequence features, they are explained in detail in Subsection D.

2) SECONDARY AND TERTIARY STRUCTURE

The basic elements that constitute the secondary structure of a protein are:

1) Turn: the part of the protein sequence that reverses its overall direction.
2) Beta strand: the part of the protein sequence that is almost fully extended.
3) Helix: the part of the sequence that forms a helix.

These elements are defined as regions and are represented in the conceptual model by the 2D Element class, where the *type* attribute is used to differentiate between the different elements, as shown in Fig. 5.

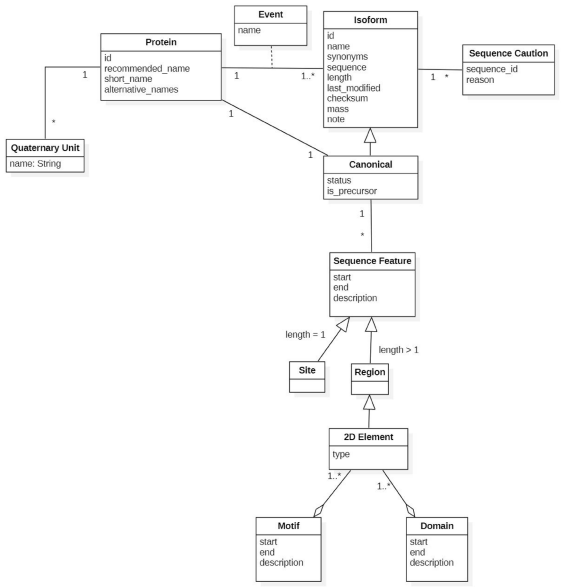

FIGURE 5. Conceptual model that represents the elements that make up the 2D and 3D structure of a protein.

Some secondary structures can be combined and organized into characteristic three-dimensional structures known

as domains and motifs. More details about them are provided in Subsection D.

3) QUATERNARY STRUCTURE

The quaternary structure of a protein represents the spatial arrangement of multiple folded protein sub-units in complexes that can range from simple dimers to large homo-oligomers. The different sub-units are represented in the conceptual model by the Quaternary Unit class which includes a name attribute (see Fig. 5).

The next modeling step is the precise characterization of the sequence features and the 3D elements that make up the three-dimensional structure (motifs and domains).

D. SEQUENCE FEATURES, MOTIFS, AND DOMAINS

Along a protein sequence, there are positions with interesting functional properties or where other compounds can bind and perform actions over the protein. These positions are known as sequence features and can be specialized into sites and regions depending on their length.

1) SITES

A site is described as a relevant single amino acid sequence that is characterized by its position and a description. Sites can be divided into three different types: cleavage sites, binding sites, and active sites. A cleavage site is a specific location at the sequence where site-specific proteases cut the protein [15]. When the protease is known, its name is represented by the protease attribute. A binding site describes the interaction between a single amino acid and another chemical entity (ligand) [16]. Ligands usually bind to the protein using weak forces (non-covalent bonding), but sometimes covalent interactions may occur. If the ligand is a metal ion, the binding site is called metal binding. If available, additional relevant information is provided, such as the nitrogen atom of the histidine side chain involved (*pros* or *tele*) and the via of the interaction (e.g., amide nitrogen and carbonyl oxygen). An active site is a position of an enzyme that is directly involved in chemical reactions [17]. Active sites can have specific roles, which are represented by the role attribute, such as charge relay system (charge movement), electrophile (electron acceptor), nucleophile (electron donor), proton donor, and proton acceptor. Nucleophiles give rise to short-lived covalent intermediates whose name is also provided if available (intermediate attribute). This characteristic is represented by an integrity constraint that states that the intermediate attribute only has value when instantiating nucleophiles. The concept of enzyme is explained in detail at the end of this section.

2) REGIONS

A region is a part of the protein sequence that describes a sequence range of interest in a general way. Special regions are those that conform the three-dimensional structure of a protein: domains and motifs. A domain is a specific combination of secondary structures that fold and function independently of the rest of the protein [18]. Special types of domains are:

1) Transmembrane domain: extent of a membrane-spanning region.
2) Intramembrane domain: extent of a region that is buried within a membrane but does not cross it.
3) Topological domain: subcellular compartment where each non-membrane region of a membrane-spanning protein is found.
4) DNA-binding domains: region that can recognize a specific DNA sequence or have a general affinity to DNA. Examples of DNA-binding domains are the AP2/ERF domain, the ETS domain, the Fork-Head domain, the HMG box, and the Myb domain.

A motif is a short structure (usually not more than 20 amino acids) that can be present in different proteins. Motifs are unable to fold independently and often do not perform a specific function [18]. Common types of motifs are:

1) Calcium binding: motif that coordinates calcium ions. One common calcium-binding motif is the EF-hand, but other calcium-binding motifs also exist.
2) Zinc finger: motif that coordinates one or more zinc ions to stabilize its structure. They are structurally diverse, and there are more than 40 types annotated in UniProtKB. The most frequent ones are the C2H2-type, the CCHC-type, the PHD-type, and the RING-type.
3) Coiled coil: motif built by two or more alpha helices that wind around each other to form a supercoil. Leucine-zippers constitute a sub type of coiled coil in which the amino acid leucine is predominant.

The result of the conceptualization process is shown in Fig. 6.

Other interesting regions in the protein sequence are:

1) Compositional bias: local shift in amino acid or nucleotide sequences that can occur as an adaptation of an organism to an extreme ecological niche, or as the signature of a specific function or localization of the corresponding protein. Types of compositionally biased regions are homopolymeric stretches and large regions of compositional bias.
2) Signal peptide: short region involved in the transport of the protein to or through the cell.
3) Propeptide: part of a protein that is cleaved during maturation or activation.
4) Transit peptide: region responsible for the transport of a protein encoded by a nuclear gene to a specific organelle (mitochondrion, chloroplast, etc.).

Signal peptides, propeptides, and transit peptides are usually removed from the mature protein due to post-translational modifications. More information about these modifications is provided in Subsection E. Four specialized classes are created (with Region as the parent class) to represent the four type of regions that are relevant.

Along the protein sequence there can also be repeats, which vary from short amino acid sequences to large repetitions

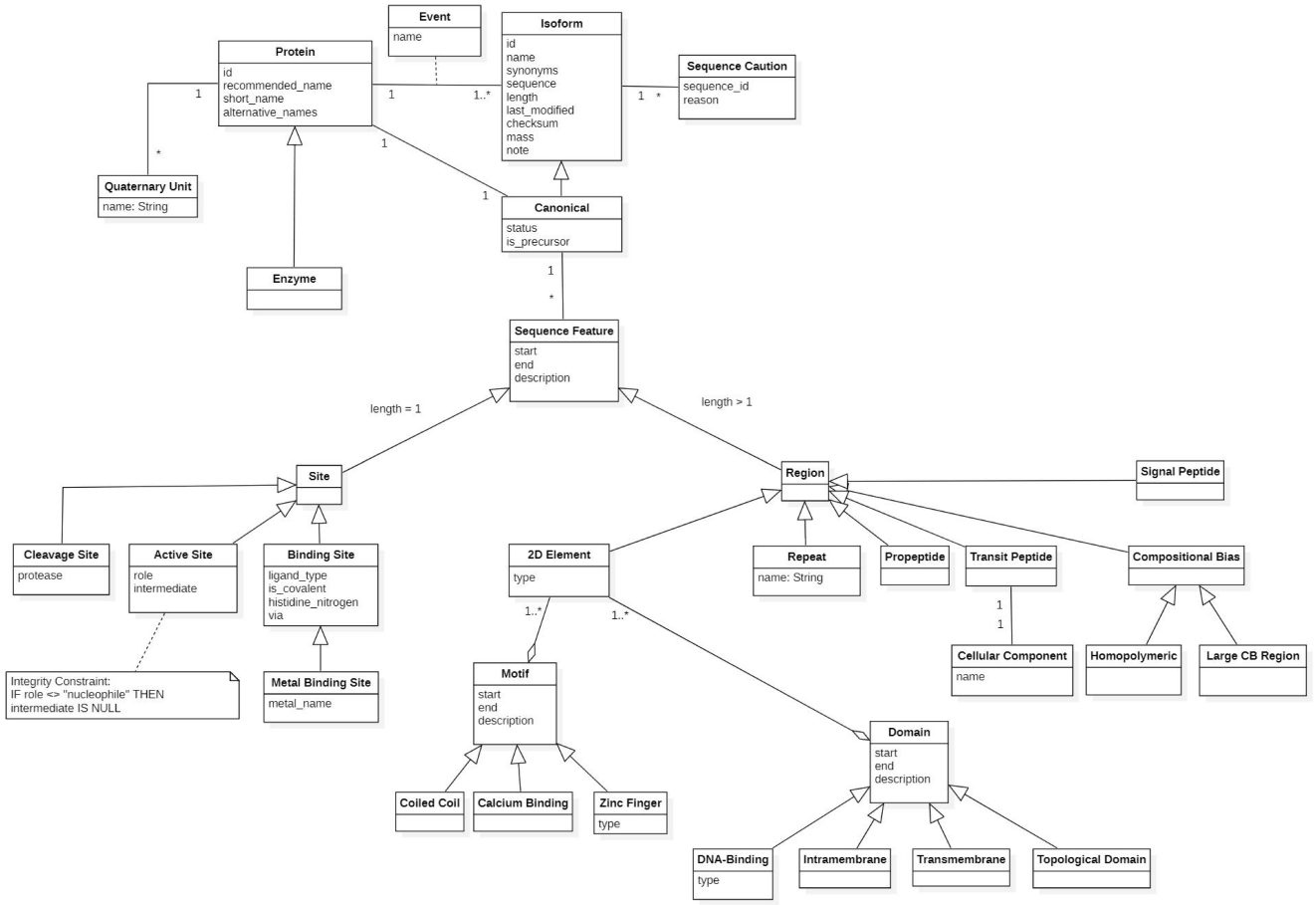

FIGURE 6. Conceptual model that represents the sequence features and the different elements that constitute the three-dimensional structure of the proteins. When describing transit peptides, the organelle where the protein is transported is represented by an association with the Cellular Component class (defined in Subsection A to represent the organelles where the gene is produced).

containing multiple domains. These repeats can contain elements that belong to different structural components and they are represented in the conceptual model using the Repeat class, as shown in Fig. 7.

Another interesting and important concept that appears in this Section is the notion of enzyme. Enzymes are a special type of protein that carry out most of the chemical reactions taking place in a cell. More information about these reactions is provided in Subsection F. A new class (Enzyme) has been created in the conceptual model as a specialization of the Protein class to represent enzymes.

Once the basic concepts that characterize the structure of proteins have been described and modeled, the next step is to conceptualize another fundamental aspect: protein processing events.

E. PROTEIN PROCESSING EVENTS

Sometimes, proteins require modifications to generate a stable structure and perform an appropriate function. These structural modifications result in a proteolytic cleavage of certain regions of the protein sequence or in the addition of a modifying group to an amino acid. They are known as protein processing events, and they may occur pre-, co-, and post-translationally. The most common ones are the co- and post-translational modifications:

1) Co-translational modifications are produced after translation has begun but before the protein is released from the ribosome. A well-known and frequent co-translation modification is the cleavage of the amino acid that commonly initiates the synthesis of proteins, which is known as Initiator Methionine.
2) Post-translational modifications (PTMs) occur once the protein has been translated and released from the ribosome [19]. Common PTMs are the removal of signal peptides, propeptides, and transit peptides. The PTMs that produce a modified residue include phosphorylation, methylation, acetylation, amidation, formation of pyrrolidone carboxylic acid, isomerization, hydroxylation, sulfation, flavin-binding, cysteine oxidation, and nitrosylation. The information that characterizes this type of PTMs includes the enzyme that carries out the modification, the host (if the protein belongs to an infectious organism), the frequency of

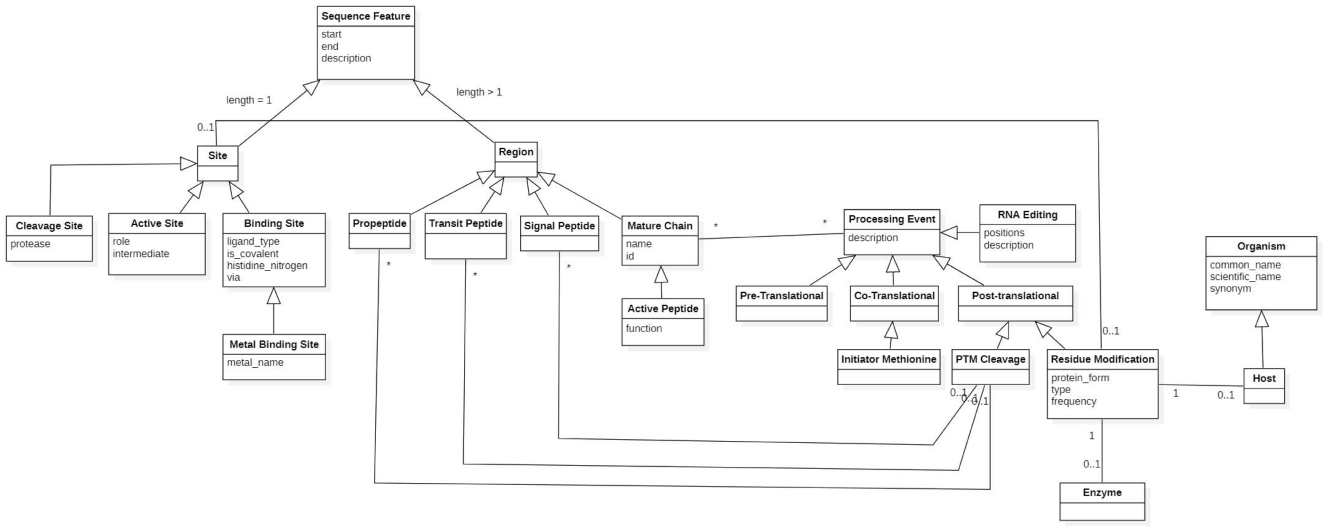

FIGURE 7. Conceptual model that represents the processing events that produce a mature protein.

the modification, and the type of relationship with any another feature (i.e., partial, alternate, or transient).

The existence of these protein processing events explains that diverse mature chains can be produced. If the mature chains have a well-defined biological activity, they are known as active peptides. The processing events are represented in the conceptual model by the Processing Event class, which specializes into the different types that we have discussed (the three Pre-, Co-, and Post-translational specialized classes). For PTM cleavages, the region that is removed is represented by an association with the corresponding class (Signal Peptide, Propeptide, or Transit Peptide). The result of the modifications is represented by the Mature Chain class and its subsequent specialization into the Active Peptide class. Each mature chain has its own identifier provided by the UniProtKB. To represent residue modifications, a new class is introduced (Residue Modification) as well as the corresponding associations with the enzyme and the host involved in the events. Fig. 7 shows the result of the conceptualization of the protein processing events.

Other post-translational modifications are described below. As they convey relevant information that has emerging properties, in order to obtain an adequate understanding of the domain, they are explicitly represented in the conceptual model as different specializations of the Post-translational class:

1) Lipidation: process that consists in the covalent binding of a lipid group to a peptide chain [20]. Common types of lipidation are N-Myristoylation, palmitoylation, GPI-anchor addition, prenylation, and lipidation of bacterial proteins (S-diacylglycerol).
2) Glycosylation: process that consists in the covalent attachment of a glycan group (mono-, di-, or polysaccharide) [21]. Glycosylation types are classified according to the identity of the atom of the amino acid which binds the carbohydrate chain, i.e., C-linked, N-linked, O-linked, or S-linked. In N-linked glycosylation, the type of glycan is provided if available, and it is represented by the corresponding attribute of the N-Linked class in the conceptual model.
3) Disulfide bond: Many proteins are stabilized by disulfide bonds. It involves a reaction between the sulfhydryl (SH) side chains of two cysteine residues [22]. Disulfide bonds are of two types: intrachain (within a polypeptide chain) and interchain (between separate protein chains). For intrachain disulfide bonds, specific information regarding the properties or the function is indicated (if provided). For interchain disulfide bonds, the name of the second protein is provided as well as the position of the second cysteine within that protein or chain. This information is represented by the Cysteine Position association class which connects the Protein and the Interchain classes.
4) Cross-link: Process that describes covalent linkages of various types that are formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links) [23]. For intrachain cross-links, the amino acids involved are explicitly mentioned. For interchain cross-links, the second amino acid corresponds to the second protein, whose name is also provided. This information is represented in the conceptual model by the Cross-link class, and the different types are represented by the *Type* attribute.

Fig. 8 shows the representation of the lipidation, glycosylation, disulfide bond, and cross-link events.

There is a special type of processing event that occurs post-transcriptionally, named RNA editing. In this process, nucleotide changes (conversions, insertions, or deletion of nucleotides) are introduced into an RNA sequence leading to one or more amino acid changes [24]. In the UniProtKB,

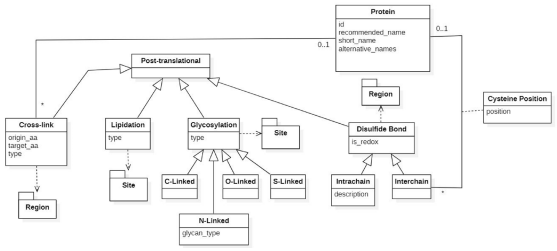

FIGURE 8. Conceptual model that represents the lipidation, glycosylation, disulfide bond, and cross-link events. To facilitate the visualization, the regions and sites associated to each class are represented as packages.

these changes are described as a list of positions and a global description that contains details about the editing process or the effect on the protein function. Conceptually speaking, this is not a precise way of representing these changes because it is not possible to specify: i) what type of change is exactly produced (insertion, conversion, or deletion of nucleotides), and ii) what consequence each change has. In any case, for our modeling purpose, the conceptual model is enriched with the RNA editing class, which is represented as an additional specialization of the Processing Event class (see Fig. 8). This leaves the door open for future semantic improvement if the relevant biological data could be more precisely represented in the corresponding data sources.

Proteins usually perform important functions such as the control of chemical reactions, the transport of ions through the cell membrane, the transformation of cell products, etc. Therefore, the next step of this work is the conceptualization of these functions.

F. THE FUNCTION OF PROTEINS

Proteins are complex molecules that perform the multitude of functions required by the cells to maintain the structure, function, and regulation of tissues and organs [25]. This information can be structured in different topics such as the general function of the protein, specific functions performed by enzymes, activity regulation, biophysicochemical properties, and pathways where the protein is involved.

1) GENERAL FUNCTION

The general function of a protein provides a general idea about the function(s) that the protein carries out, along with the supporting evidence. For example, the Phenylalanine-4-hydroxylase protein catalyzes the hydroxylation of L-phenylalanine to L-tyrosine, and this assertion has been performed manually based on two experiments whose corresponding publications can be accessed in PubMed using the identifiers 18460651 and 18835579, respectively. In some cases, each protein isoform performs a different function. Therefore, in the conceptual model, the concept of function is represented with a class that is associated to the Isoform class and not to the Protein class.

2) SPECIFIC FUNCTIONS PERFORMED BY ENZYMES

The main function of enzymes is to catalyze the chemical reactions that occur in the cell. The information is extracted from the Rhea database [26] whenever possible or described as free text. These chemical reactions are usually associated to an identifier provided by the Enzyme Commission number (*ec_number*). As the detailed description of a chemical reaction is out of the scope of this work, only a general *description* and the *ec number* are represented as attributes. Any reference to external sources that provide more information is represented using a cross-reference to the corresponding database. Once more, we want to emphasize that this basic conceptualization facilitates future semantic extensions done from a solid, conceptually well-grounded basis. To carry out their catalytic activity, enzymes require non-protein molecules called cofactors. The UniProtKB only represents cofactors that allow more than 50% of the maximum catalytic activity. Cofactors are described by a name (e.g., Fe2+) and an identifier according to the Chemical Entities of Biological Interest (ChEBI) database [27]. Any additional information is provided as a note. The conceptualization of the functions performed by proteins and enzymes as well as the cofactors are represented in Fig. 9.

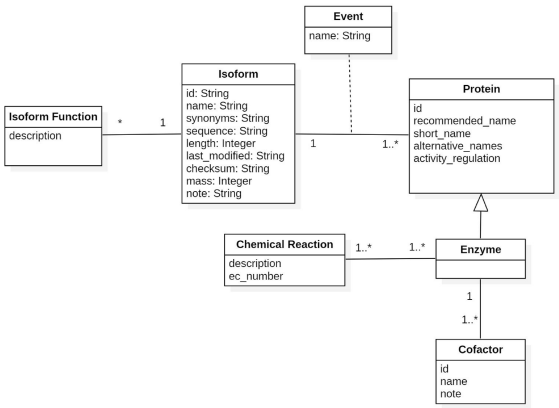

FIGURE 9. Conceptual model that represents the chemical reactions catalyzed by enzymes and the cofactors required by them.

3) ACTIVITY REGULATION

There are regulatory mechanisms that control (activate or inhibit) the functions performed by the protein. For example, phosphorylation leads to an increase in the catalytic activity of the Tyrosine 3-monooxygenase protein. Prior to release 2018_08, the activity regulation was only associated to enzymes. Afterwards, the activity regulation was extended to transporters and microbial transcription factors. These mechanisms and the elements involved are described as free text and represented as an attribute of the Protein class in the conceptual model (see Fig. 10). Since the type of the protein is not represented in the UniProtKB, it is not possible to specify which types are affected by the activity regulation, we have

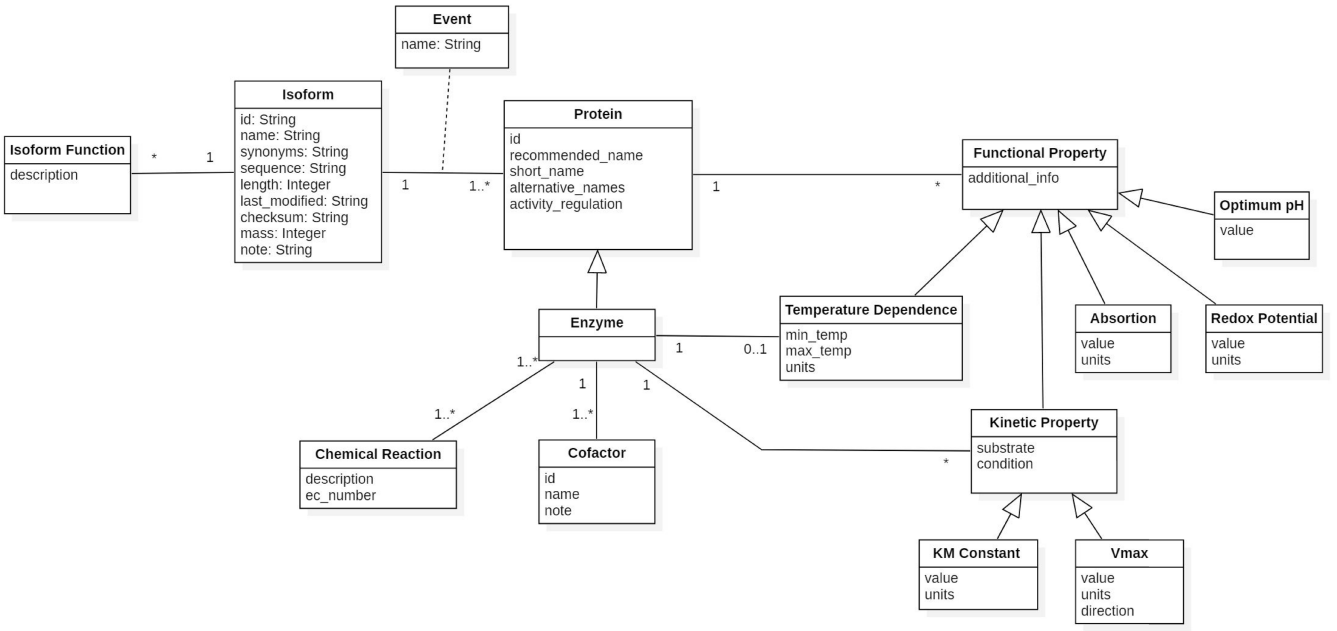

FIGURE 10. Conceptual model that represents the functional properties of proteins. Kinetic properties are specific to enzymes and extend the information with the substrate and the environmental conditions required for the reaction to take place.

identified this as a conceptual weakness to be considered for further quality improvement of the UniProtKB data.

4) BIOPHYSICAL AND CHEMICAL PROPERTIES

Proteins present a set of biophysical and chemical properties that are directly related to their functional capacity. The UniProtKB database provides information about the following ones:

1) Maximal light absorption: This property indicates the wavelength at which photoreactive proteins show their maximal light absorption (e.g., 353 nm). The Michaelis-Menten constant (KM) and maximal velocity (Vmax): These kinetic properties are used to study the chemical reactions that are catalyzed by enzymes. The KM constant indicates the affinity of an enzyme for a substrate (e.g., the KM value of Deoxynucleoside kinase for thymidine is 0.9 μM). The Vmax of the reaction is the rate reached when the enzyme sites are saturated with the substrate (e.g., the Vmax of Deoxynucleoside kinase for thymidine is 29.4 mmol/min/mg). Both parameters depend on environmental conditions. If the enzyme is multifunctional or if the reaction is reversible, different KM and Vmax values can be measured.
2) pH dependence: This property is used to describe the optimum pH for protein activity.
3) Redox potential: The redox potential is specific to electron transport proteins and measures the tendency of the protein to gain or lose electrons (e.g., the redox potential of TMX3 protein is 157 mV).
4) Temperature potential: The temperature potential indicates the optimal temperature range at which an enzyme performs its activity (e.g., the optimal temperature for the XTH22 enzyme is from 12 to 18 degrees Celsius).

In general, each property is conceptually described by its value and units (Celsius degrees, μM, etc.), which allows new properties to be added if required, or allows measures to be represented using different metrics. Any additional information can be also included as free text. Temperature dependence is described as a range (min and max temperature) and kinetic properties (KM constant and Vmax) are represented as a specialization class (Kinetic Property) associated to enzymes in order to extend their information with the substrate and the environmental conditions. The Vmax also considers the direction (forward or backward) in which the reaction takes place because the value can differ if the reaction is reversible. The details of this part of the conceptual model are shown in Fig. 10, where classes and attributes, associations with their corresponding cardinalities, and specialization have been introduced to represent all of the relevant domain properties that have been described. As occurred with the activity regulation, the absence of protein types in the UniProtKB database forces the redox potential to be represented as a generic functional property. All of the relevant concepts are in any case adequately represented in the holistic conceptual model that we have designed.

Throughout this section, the functional characteristics of proteins and enzymes have been described in detail to identify the building units that explain the elementary working procedure of proteins. However, things are not so simple. In real life, the individual functions carried out by each protein are sequentially linked to others, making complex reactions called metabolic pathways that are key for the correct

functioning of the cells. This is the last significant protein function cited at the beginning of Subsection F, which we analyze individually in its own subsection due to its relevance and its potential conceptual complexity.

G. BIOLOGICAL PATHWAYS AND SUBCELLULAR LOCATIONS

The chemical reactions that occur within a cell are sequentially linked in a series of steps called biological pathways. These pathways can be very complex and are usually made up of different subpathways. Therefore, they are commonly described as a hierarchy of superpathway, pathway and subpathway. For example, the pathway called L-phenylalanine degradation is part of a more complex pathway called Amino acid degradation. The proteins and the enzymes act as participants in these pathways and any modification in their structure can alter their function and thus the balance of the cell, potentially leading to disease. For this reason, it is very important to correctly determine in which pathways the proteins participate because it can help to understand the impact of any protein sequence alteration [28]. The part of the conceptual model that represents the information about pathways is shown in Fig. 11.

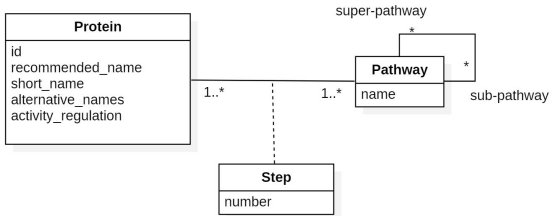

FIGURE 11. Conceptual model that describes the hierarchy of pathways in which a protein takes part. The number of the step is represented as an association class.

The chemical reactions and the role that the proteins play are described in specialized databases and repositories such as KEGG [29] and Reactome [30]. Details about the different steps that conform the functional structure of a pathway can be found in them (the current model only specifies a basic reference to the number of the pathway step where a protein participates). The integration of all of this detailed information with the conceptual model that we are elaborating is a very attractive objective of our future work, which would make it possible to increase the understanding of the processes that lead to disease. As we have commented previously, it is absolutely necessary to have a core conceptual model in order to guide the subsequent extension process. This conceptual backbone is the main contribution of this paper.

Proteins have evolved to function optimally in a specific subcellular localization (nucleus, cytosol, plasmatic membrane, etc.) [31]. Each isoform can act in a different location, and the correct identification of these locations can improve the understanding of protein functions and the discovery of new therapeutic targets. Locations are described using a controlled vocabulary with a name, a description, and a note

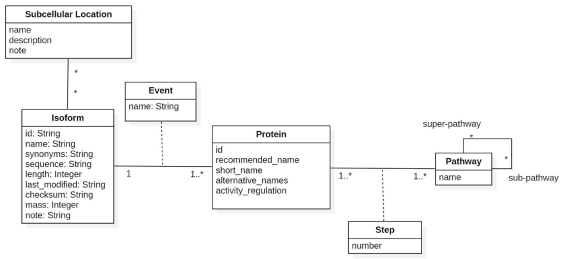

FIGURE 12. Conceptual model that describes the locations where proteins perform their function.

to add additional information if required. Fig. 12 shows how this information has been included in the conceptual model.

Since proteins are functionally related or are members of the same pathway or protein complex, the conceptual schema allows the representation of co-expression networks, a potent approach to gather biologically relevant information, e.g., for the identification of genes not yet associated with explicit biological questions, and for accelerating the interpretation of molecular mechanisms at the root of significant biological processes.

Alterations in the structure of the proteins can lead to malfunction and consequently to the development of a disease. Therefore, the next step of this work is the conceptual identification of the mechanisms that can produce this situation.

H. INVOLVEMENT IN DISEASE: VARIANTS AND POLYMORPHISMS

As explained in previous sections, proteins do not function in isolation, and their interactions with one another mediate metabolic and signaling pathways as well as complex cellular processes. Due to their central role in the biological function of cells, the changes in DNA that affect the structure of the proteins can produce folding and interaction problems leading to disease in the affected organisms. For example, protein misfolding is believed to be the primary cause of Alzheimer's disease, Parkinson's disease, Huntington's disease, and many other degenerative and neurodegenerative disorders [32]. Some proteins may also cause allergic reactions in certain organisms (e.g., mammals) or catalyze reactions that may cause multiple allergies. The Allergic Reaction specialization of the Disease class represents this aspect.

The information about the diseases associated to genetic variants are commonly described by a disease name, an abbreviation, and a description. For example, the Adrenocorticotropic hormone receptor is associated to Glucocorticoid deficiency 1, also known as GCCD1. Additionally, the role of proteins in disease pathogenesis (causative, susceptibility, modifier, etc.) is also provided if available. The representation of the diseases associated to a protein in the conceptual model is shown in Fig. 13.

The variants that occur in the protein sequence are represented by the amino acid change (e.g., S $\rightarrow$ I), the position in the sequence (start and end regarding the canonical

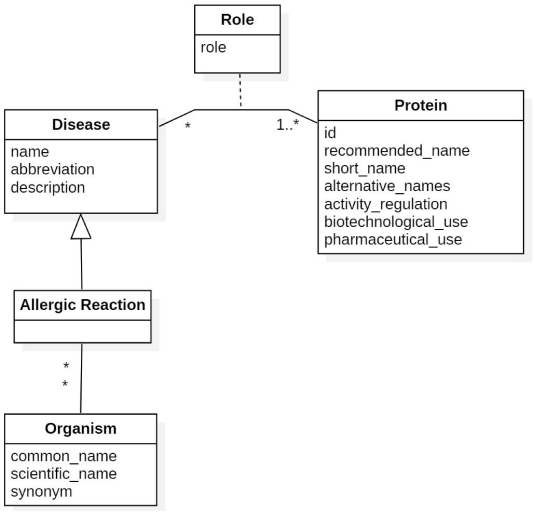

FIGURE 13. Conceptual model that describes the diseases associated to a protein. The role of the protein in the disease is represented by the association class Role.

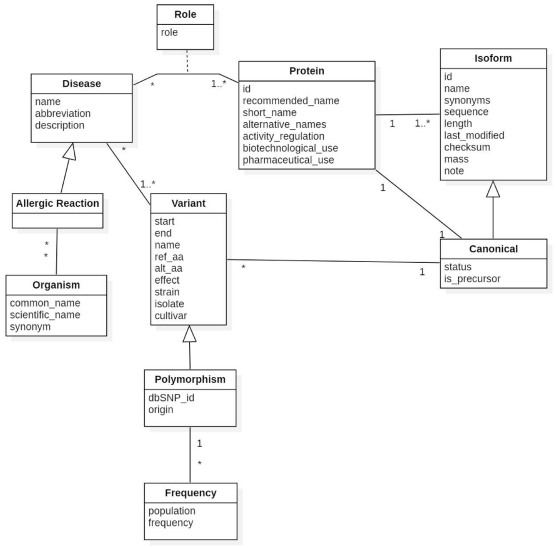

FIGURE 14. Conceptual model that describes the variants that may occur in the protein sequence.

sequence), the name of the variant (if known), and its effect on the protein, the cell, or the complete organism (e.g., the change of an Isoleucine, I, by a Valine, V, in position 79 of the Aldo-keto reductase family 1 member C2 protein sequence causes a partially impaired activity). If the variant is observed in specific strains, isolates, or cultivars, they are also represented. All of this information is modeled by including the Variant class in the conceptual model with its corresponding attributes.

Polymorphisms are a type of variant that commonly consists of a single nucleotide change (known as Single Nucleotide Polymorphism or SNP) at the codon level. Even when it is known that some polymorphisms can involve more than one amino acid, the SNP term is generally used to describe this type of small changes. If the SNPs have been annotated in the dbSNP database [33], the corresponding identifier is provided. Additional information that can be provided to describe polymorphisms is the cell type or tissue of origin of the variant (somatic or germline) and the distribution (frequency) of the SNP in a given population. To represent SNPs, a specialization of the Variant class is introduced in the conceptual model (the Polymorphism class) together with an association to the Frequency class, which allows expressing that a common polymorphism may have different frequencies in different populations. The result of the conceptualization process is depicted in Fig. 14.

It is important to highlight that some types of changes are not annotated in the UniProtKB database because their deleterious effects on the protein function are considered to be obvious. These include major changes such as frameshifts or premature stops. In addition, nucleotide indels are not described in detail since they are usually assumed to produce a nonfunctional protein.

Some proteins have a toxic effect that can be lethal when present in a certain dose or concentration. The UniProtKB database provides information about the organism and the mode of delivery (intraperitoneal, intravenous, intramuscular, subcutaneous, intracerebroventricular, intracranial, or intraabdominal injection) that produces a certain effect (lethal dose, paralytic dose, effect dose, or lethal concentration) in at least 50% of the tested organisms. The representation of this information in the conceptual model is shown in Fig. 15.

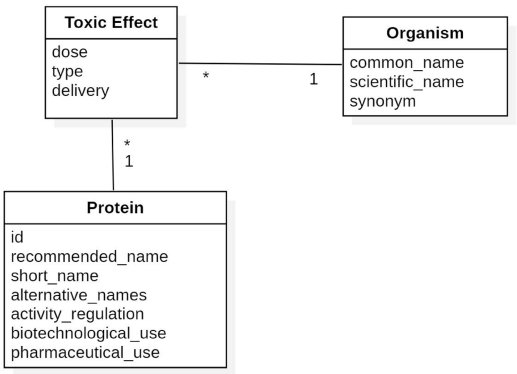

FIGURE 15. Conceptual model that describes the toxic effect that proteins can cause when they are present in a certain dose or concentration.

Besides the information that has been already described in the previous sections, the UniProtKB database also provides additional data associated to characteristics of the protein entry in the database (last update, status, etc.), similarities with other sequences, cross-references to external sources, and the possible industrial and pharmaceutical use of the protein.

I. ADDITIONAL INFORMATION

The UniProtKB database reports cross-references to other relevant data sources where more specialized data can be

found (sequence databases, chemistry databases, genome annotation, enzymes, and pathways, etc.). Additionally, the UniProtKB provides links to other proteins whose sequences are similar at different levels of identity thresholds (100%, 90%, and 50%). Caution notes are used to represent any possible error and/or cause of confusion that could be relevant for the interpretation of the information provided about the protein.

Other relevant information about the protein entry that can also be found in the UniProtKB database includes the last update, the annotation program, and the status. The status is a set of descriptors that summarizes the annotation content and the evidence about the protein. The status is composed of three main descriptors:

1) Entry status: indicates whether or not an entry in the database (in this case, the data about the protein) has been manually annotated and reviewed by the UniProtKB curators. Its possible values are Reviewed and Unreviewed.
2) Annotation score: provides a heuristic measure of the annotation content of a protein (protein names, functional annotations, sequence annotations, cross-references, etc.). The final score is computed in terms of the completeness of this content and is represented as a 5-point-system. Proteins with an annotation score of 1 have a rather basic annotation, and proteins with an annotation score of 5 are considered to be the best-annotated entries.
3) Protein existence: indicates the level of the evidence that supports the existence of the protein. The level of the evidence can range from uncertain (the existence of the protein is unsure) to experimental (there is clear experimental evidence for the existence of the protein). The values that can be assigned to this descriptor are: Protein uncertain, Protein predicted, Protein inferred from homology, Experimental evidence at transcript level, and Experimental evidence at protein level.

The conceptual representation of cross references, sequence similarities, caution notes, and entry-specific data is shown in Fig. 16.

Proteins can also be used in industrial biotechnological processes or as a pharmaceutical drug. These characteristics are described using the *biotechnological_use* and *pharmaceutical_use* attributes, which have been added to the Protein class.

The evidence that supports the assertions made on the characteristics of a protein is represented as a set of ''evidence tags'', which describes the source of the information (e.g., an experiment that has been published in the scientific literature). Each evidence tag has an evidence type (e.g., manual assertion based on experiment, inferred from electronic annotation, etc.), and the source(s) of the information, which are usually database records (e.g., articles from the scientific literature are represented as PubMed records). As it is not possible to precisely determine which attributes constitute

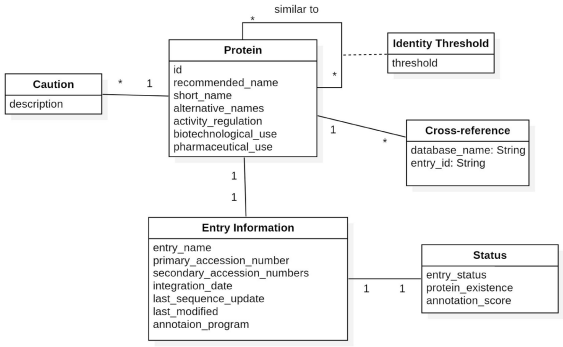

FIGURE 16. Conceptual model that describes additional information about the protein entry in the database, cross-references to other sources, and similarity with other proteins.

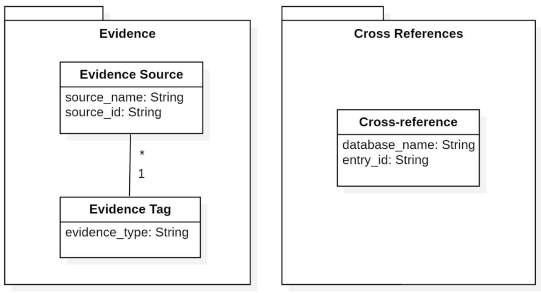

FIGURE 17. Conceptual model that represents the evidence supporting protein assertions and cross-references to external sources.

each record, only the name of the source and the identifier of the record are initially considered in the conceptual model. This allows the user to navigate to the specified source if more information is required. Fig. 17 shows how evidence and cross references are represented in the conceptual model, using the *Evidence Tag*, the *Evidence Source*, and the *Cross-reference* classes.

Any element of the conceptual model is susceptible to having the associated evidence tags and cross-references. As the representation of these elements could make the model very complex and difficult to read, they have been encapsulated into packages and omitted from the figures shown throughout this paper. Nevertheless, in the global model, all evidence tags, and cross-references are represented as dependencies in the corresponding packages. The complete model can be accessed in [34].

Throughout this section, the main concepts that characterize the complexity of protein structure, function, and association with disease have been described and represented through a conceptualization process that results in a holistic conceptual model. The concepts have been represented as they are described in a well-known, widely used database curated by experts in the domain, the UniProtKB database. During the process, the complexity of the information led to the identification of different issues that hinder the description of the underlying ontological commitment.

III. DISCUSSING THE UNDERLYING ONTOLOGICAL COMMITMENT

Having a sound ontological commitment to describe the relevant concepts of a domain provides its shared understanding and help to structure, share, collect, and analyze data in a precise way to derive meaningful conclusions. Throughout this work, the different concepts about protein structure, function, and association with disease used by the UniProtKB database have been analyzed to determine the underlying ontological commitment. During this conceptualization process, two issues or weaknesses have been identified.

The first issue is related to the types of proteins. It is known that some functions or properties are specific to certain types of proteins such as enzymes. Nevertheless, it is interesting to mention that the UniProtKB database does not explicitly differentiate between types of proteins (e.g., providing a *type* field). Therefore, it is not possible to determine how many protein types are considered and which specific characteristics or information are associated to each one. From a conceptual modeling perspective, this differentiation is very important. The specialization of proteins into different types would allow us to clearly determine which functions can be carried out by each type, providing a better understanding of the domain and the information that is going to be explored. This would also avoid mistakes in data collection and representation, allowing the development of sound information systems to manage all the increasing and complex knowledge.

The second issue found is that some data do not exactly correspond to the concept they represent. For example, sites are defined as ''interesting single amino acid sites on the sequence''. Using this definition, sites should correspond to only one amino acid position in the protein sequence. Nevertheless, when searching the UniProtKB database, it is possible to find sites with a length of two amino acids (e.g., in protein Q9UDY8), which contradicts the main definition provided by the documentation.

Despite these issues, the UniProtKB database can be considered a well-grounded and complete repository that represents all of the most significant information about proteins. This repository allows experts to collect data from many different aspects and to extend the knowledge with cross-references to other specialized sources, providing a detailed view of this complex and interesting domain. With the help of the conceptual modeling presented in this work, a shared understanding of the domain is facilitated, while further extensions and clarifications can be incorporated as new information is discovered and becomes available.

IV. CONCLUSION AND FUTURE WORK

Proteins are the working machines that perform essentially all functions in living systems. Therefore, it is crucial to have a good understanding of how proteins fold and which biological processes are involved in order to make predictions about their function and to comprehend how changes in the protein structure can lead to disease.

In this dynamic and changing context, the information required to achieve a proper understanding is very complex as there are many interconnected concepts that must be precisely defined to avoid misunderstandings. This complexity can be observed when accessing specialized repositories such as the UniProtKB, where the structure used to represent the information on the website can be overwhelming for any interested user.

Thoughout this work, a sound analysis of the concepts managed by the database have been performed to derive the underlying ontological commitment. The result of the conceptualization process is a conceptual model that is represented using an UML Class Diagram. During this process, it was also possible to identify some issues or conceptual weaknesses that hinder the understanding and the correct representation of important concepts such as the type of proteins and the sites in a protein sequence. Despite these issues, the UniProtKB is a well-grounded, and commonly used database that provides valuable knowledge about this complex domain.

Since the work has focused on analyzing the main concepts managed in the UniProtKB database, some details about more specific concepts remained out of its scope. These concepts are managed by other specialized databases to which the UniProtKB provides cross-references. Due to the importance of considering them in order to obtain a complete and solid understanding about the whole protein domain, a detailed analysis about the following concepts is considered the aim of one immediate future work:

- A more detailed analysis of protein-protein interactions (PPIs): The study of the interactome is crucial to understand the causes that lead to the development of certain diseases and the mechanisms by which pathogens such as viruses or bacteria are capable of producing an infection in other organisms. The PPIs depend on cell type, developmental stage, environmental conditions, protein modifications, etc., which is known as biological context. This information is provided by specialized databases and repositories such as DIP, IntAct, and MINT. We plan to extend and semantically enrich our conceptual model with the corresponding, relevant information.
- Analysis of biological pathways: The chemical reactions and processes that occur within a cell determine its correct or incorrect function and thus the healthy or unhealthy state of a living system. These reactions and processes can be very complex, and detailed information about them, and the role that the proteins have is described in specialized databases and repositories such as KEGG (https://www.genome.jp/kegg/) and Reactome (https://reactome.org/). We also plan to analyze these data sources in order to make the corresponding conceptual extensions to our initial conceptual model of proteins.

- Analysis of the Gene Ontology (GO): Another way of determining molecular functions, subcellular locations, and biological processes in which the proteins are involved is using a set of hierarchical terms defined by Gene Ontology (GO)(http://geneontology.org/). Even when these terms are similar but not exactly equivalent to pathways, they are widely used by different databases. Taking this terminology into consideration opens another significant future extension for this work.
- A more detailed analysis of the quaternary structure: The complexity and importance of the structure of the proteins require a deeper analysis of the elements that constitute it to precisely understand their function. We plan to extend the conceptual model with the corresponding information.

The analysis done in this work is crucial in stating the need of having a sound ontological commitment in a domain as complex as genomics. In this case, we have focused on the protein context, but the description of the genome structure and the understanding of how it works require a holistic perspective that must include much more information than that obtained only from proteins. Following this line of reasoning, the Conceptual Schema of the Human Genome (developed by the PROS Research center at the Universitat Politècnica de València) [35], [36] [37] represents a first stone in the building of the core of a solid and conceptually well-grounded description of the domain. The results of this work will serve to enrich the existing model, increasing its value and allowing a shared understanding among experts.

REFERENCES

[1] M. West, "Embracing the complexity of genomic data for personalized medicine," *Genome Res.*, vol. 16, no. 5, pp. 559–566, May 2006.
[2] S. Spreeuwenberg, P. Henao, and K. Hiroi, *AIX: Artificial Intelligence Needs EXplanation: Why and how Transparency Increases the Success of AI Solutions*. Amsterdam, The Netherlands: CB, 2019.
[3] *NIH Genetics Glossary (Protein)*. Accessed: May 20, 2020. [Online]. Available: https://www.genome.gov/genetics-glossary/Protein
[4] UniProt Consortium, "The universal protein resource (UniProt) in 2010," *Nucleic Acids Res.*, vol. 38, no. 1, pp. D142–D148, Jan. 2010.
[5] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, "UniProtKB/Swiss-prot," in *Plant Bioinformatics*, D. Edwards, Ed. Totowa, NJ, USA: Humana Press, 2007, pp. 89–112.
[6] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: Comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, May 2007.
[7] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler, "UniProt archive," *Bioinformatics*, vol. 20, no. 17, pp. 3236–3237, Nov. 2004.
[8] UniProt. *What Are Proteomes?* Accessed: May 20, 2020. [Online]. Available: https://www.uniprot.org/help/proteome
[9] G. Rabbani, M. H. Baig, K. Ahmad, and I. Choi, "Protein-protein interactions and their role in various diseases and their prediction techniques," *Current Protein Peptide Sci.*, vol. 19, no. 10, pp. 948–957, Aug. 2018.
[10] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Feuermann, U. Hinz, and C. Jandrasits, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D841–D846, Jan. 2012.
[11] L. Salwinski, "The database of interacting proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 449–451, Jan. 2004.
[12] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D857–D861, Jan. 2012.
[13] Nature. *Protein Structure*. Accessed: May 20, 2020. [Online]. Available: https://www.nature.com/scitable/topicpage/protein-structure-14122136/
[14] M. Stastna and J. E. Van Eyk, "Analysis of protein isoforms: Can we do it better?" *Proteomics*, vol. 12, nos. 19–20, pp. 2937–2948, Oct. 2012.
[15] J. Ni and M. Kanai, "Site-selective peptide/protein cleavage," in *Topics in Current Chemistry*, vol. 372. Cham, Switzerland: Springer, 2015, pp. 103–123.
[16] J. Si, R. Zhao, and R. Wu, "An overview of the prediction of protein DNA-binding sites," *Int. J. Mol. Sci.*, vol. 16, no. 12, pp. 5194–5215, Mar. 2015.
[17] Britannica. *The Role of the Active Site*. Accessed: May 20, 2020. [Online]. Available: https://www.britannica.com/science/protein/The-role-of-the-active-site
[18] W. R. P. Novak, "Tertiary structure domains, folds, and motifs," in *Molecular Life Sciences*. New York, NY, USA: Springer, 2014, pp. 1–5.
[19] V. Uversky, "Posttranslational modification," in *Brenner's Encyclopedia of Genetics*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego, CA, USA: Elsevier, 2013, pp. 425–430.
[20] B. Chen, Y. Sun, J. Niu, G. K. Jarugumilli, and X. Wu, "Protein lipidation in cell signaling and diseases: Function, regulation, and therapeutic opportunities," *Cell Chem. Biol.*, vol. 25, no. 7, pp. 817–831, Jul. 2018.
[21] J. Eichler, "Protein glycosylation," *Current Biol.*, vol. 29, no. 7, pp. R229–R231, Apr. 2019.
[22] M. J. Saaranen and L. W. Ruddock, "Disulfide bond formation in the cytoplasm," *Antioxidants Redox Signaling*, vol. 19, no. 1, pp. 46–53, Jul. 2013.
[23] UniProt. *Cross-Link*. Accessed: May 20, 2020. [Online]. Available: https://www.uniprot.org/help/crosslnk
[24] K. L. Witkin, S. E. Hanlon, J. A. Strasburger, J. M. Coffin, S. R. Jaffrey, T. K. Howcroft, P. C. Dedon, J. A. Steitz, P. J. Daschner, and E. Read-Connole, "RNA editing, epitranscriptomics, and processing in cancer progression," *Cancer Biol. Therapy*, vol. 16, no. 1, pp. 21–27, Jan. 2015.
[25] Genetics Home Reference. *What Are Proteins and What do They do?* Accessed: May 20, 2020. [Online]. Available: https://ghr.nlm.nih.gov/primer/howgeneswork/protein
[26] T. Lombardot, A. Morgat, K. B. Axelsen, L. Aimo, N. Hyka-Nouspikel, A. Niknejad, A. Ignatchenko, I. Xenarios, E. Coudert, N. Redaschi, and A. Bridge, "Updates in Rhea: SPARQLing biochemical reaction data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D596–D600, Jan. 2019.
[27] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1214–D1219, Jan. 2016.
[28] National Human Genome Research Institute. *Biological Pathways Fact Sheet*. Accessed: May 20, 2020. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet
[29] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.
[30] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, and F. Loney, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D498–D503, Nov. 2019.
[31] P. Dönnes and A. Höglund, "Predicting protein subcellular localization: Past, present, and future," *Genomics, Proteomics Bioinf.*, vol. 2, no. 4, pp. 209–215, Nov. 2004.
[32] 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.
[33] S. T. Sherry, "DbSNP: The NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001.
[34] A. Leon and O. Pastor, "Conceptual model of proteins," Universitat Politècnica de València, Valencia, Spain, Tech. Rep. 147884, 2020. [Online]. Available: http://hdl.handle.net/10251/147884

[35] J. F. R. Román, O. Pastor, J. C. Casamayor, and F. Valverde, "Applying conceptual modeling to better understand the human genome," in *ER 2016: Conceptual Modeling*, vol. 9974. Gifu, Japan: Springer, 2016, pp. 404–412.
[36] O. P. López, A. L. Palacio, J. F. R. Román, and J. C. Casamayor, "Modeling life: A conceptual schema-centric approach to understand the genome," in *Conceptual Modeling Perspectives*, J. Cabot, C. Gómez, O. Pastor, M. R. Sancho, and E. Teniente, Eds. Cham, Switzerland: Springer, 2017, pp. 25–40.
[37] O. Pastor, "Conceptual modeling meets the human genome," in *Conceptual Modeling—ER 2008. ER 2008* (Lecture Notes in Computer Science), vol. 5231, Q. Li, S. Spaccapietra, E. Yu, and A. Olivé, Eds. Berlin, Germany: Springer, 2008, pp. 1–11.

ANA LEON received the Ph.D. degree in computer science from the Universitat Politècnica de València (UPV), in 2019. She is currently a Researcher with the Research Center on Software Production Methods (PROS), UPV, where her research activity is focused on the use of conceptual models for the development of Genomic Information Systems, and the definition of a systematic process for the search, identification, load and exploitation of DNA variants in the context of Precision Medicine. She is also an University Expert in Medical Genetics and Genomics from the Universidad Católica de Murcia, Spain. Her research interests include conceptual modeling, genomic data science, explainable AI, data quality, and information systems.

OSCAR PASTOR (Member, IEEE) is currently a Full Professor and the Director of the PROS Research Center, 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.