Conceptual Modeling of Proteins Based on UniProt

Ana León Palacio[1][0000-0003-3516-8893] and Óscar Pastor [0000-0002-1320-8471]

[1] Universitat Politècnica de València, Valencia 46022, Spain
{aleon|opastor}@pros.upv.es

Abstract. 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 to facilitate an understanding of the disease mechanisms as well as to guide the development and application of therapies. This work focuses on describing how a shared understanding of the domain can be achieved by analyzing the conceptual precision of the main concepts that should constitute the ontological commitment 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 systems. The contribution of this paper is to show the conceptual complexity of the UniProtKB database, and to let users face and manage that complexity by providing a sound and well-grounded conceptual background to achieve the shared understanding of the domain, a crucial aspect to allow the design of any fruitful data analytics-based strategy. A conceptual model for proteins is carefully developed taking the UniProtKB database as data source, explaining in detail the problems that have been faced together with their corresponding solutions.

Keywords: Conceptual Modeling, Proteins, UniProtKB.

1 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 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.). 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 proteins case, that 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. Get 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 AI techniques that deliver results.
5. Generate good explanations.
6. Evolve the solution over time as more knowledge is acquired.

To accomplish the first step correctly is crucial to assess the quality of the whole process. This work focuses on describing how that first step can be achieved by analyzing the conceptual precision of the main concepts that should constitute the ontological commitment 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 systems.

Get a shared understanding of the domain under consideration requires to have a precise conceptual model of reference, ontologically well-grounded (identifying precisely the relevant concepts of the domain) and interpreting accurately the data that are managed in 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. 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. The contribution of this paper is to show the conceptual complexity of the UniProt, and to let users face and manage that complexity by providing a sound and well-grounded conceptual background to achieve the shared understanding of the domain, a crucial aspect to allow the design of any fruitful data analytics-based strategy. A conceptual model for proteins is carefully developed taking the UniProtKB database as data source, explaining in detail the problems that have been faced together with their corresponding solutions.

To this end this paper is structured as follows. In section 2, we introduce the concepts used to understand the structure, function, and involvement in disease of proteins. Using these concepts as 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.

2 Conceptual Modeling of Protein Information

The information needed to precisely characterize the concept of protein, in all 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 manage correctly all the available information about proteins, an immediate need emerges: having adequate, accurate, consistent, and manageable data sources. In this working domain, the UniProt database is a widely accepted

and used data source. We start our work by introducing its structure, in order the 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)[1], the SIB Swiss Institute of Bioinformatics[2] and the Protein Information Resource (PIR)[3]. The UniProt is supported by four main databases: the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef)[4], the Proteomes and the UniProt Archive (UniParc)[5]. The UniProtKB is the central database where functional information on proteins and annotation data (either manual or automatically produced) 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. The 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 4 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 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 get, and to facilitate a valuable and fruitful data exploitation strategy. To achieve this goal, this work is concretely based on the information provided by the UniProtKB database.

---

1 EMBL-EBI: https://www.ebi.ac.uk/
2 SIB: https://www.sib.swiss/
3 PIR: https://proteininformationresource.org/
4 UniRef: https://www.uniprot.org/uniref/
5 UniParc: https://www.uniprot.org/uniparc/

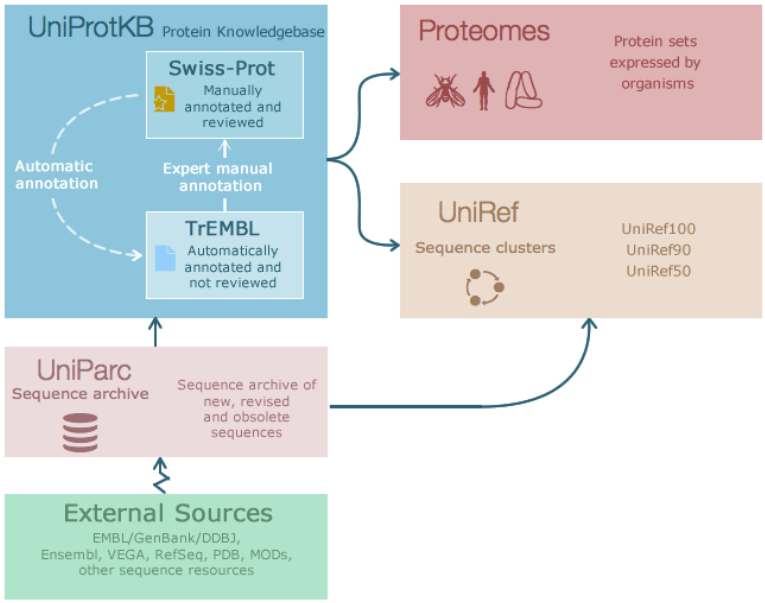

Fig. 1. Structure of the UniProt repository [5].

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, function and the sequence changes that can lead to disease.

2.1 Proteomes and Basic Information about Proteins

The entire set of proteins that can be expressed by the genome of an organism is called proteome. 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[6], that belongs to the UniProt repository. The basic concepts required to start the conceptualization process are organism, gene, and protein.

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, that is a hierarchy that represents the relative level of a group of organisms. If the organism described is a virus, the specific organism

[6] Proteomes: https://www.uniprot.org/proteomes/

or taxonomic group that are susceptible to be infected (called host) must be also specified because viruses only exist in association with a host.

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.). For naming the genes, the UniProtKB uses the acronym or official symbol (e.g. PAH). Genes are also represented using an identifier, provided by an external source (e.g. the HUGO Gene Nomenclature Committee), 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[7] by a sequencing project (known as ORF Names).

Protein. The information about proteins is basically composed by a unique identifier, provided by the UniProtKB (e.g. P00439), and a name (e.g. Phenylalanine-4-hydroxylase). The names of the proteins have evolved along time and some of them have become obsolete. Nevertheless, in some scientific literature and databases these names are still in use and the 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 be also 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 that can be seen in Fig. 2. The resulting conceptual model has been described using a UML Class Diagram that includes the classes, with their attributes and relationships, that model all the relevant information that has been introduced.

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

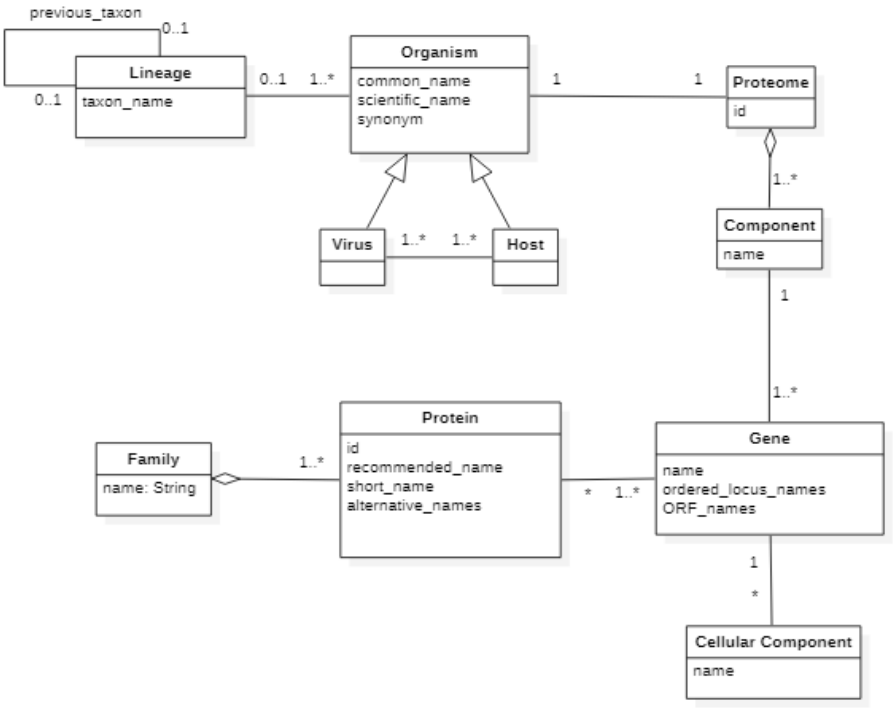

Fig. 2. Conceptual model that represents the basic information about proteomes, genes, proteins, and organisms.

Proteins rarely act alone, as 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 to be analyzed in our work.

2.2 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 [6]. The complete 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, as the occurrence of aberrant PPIs is the basis of multiple aggregation-related diseases, such as Creutzfeldt–Jakob and Alzheimer's diseases.

The UniProtKB database provides information about binary-protein interactions[8] (extracted from the IntAct[9] database), host-pathogen interactions, and protein-complex

[8] Binary-protein interactions: direct physical interactions between proteins.
[9] IntAct: https://www.ebi.ac.uk/intact/

interactions[10]. In the conceptual model shown in Fig. 3, the PPIs are represented by the association class Interaction. 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 considered). For host-pathogen interactions, the name of the virus is represented by an association with the corresponding Virus and Host classes. For protein-complex interactions, the association with the Complex class provides the name of all the proteins that are part of the complex.

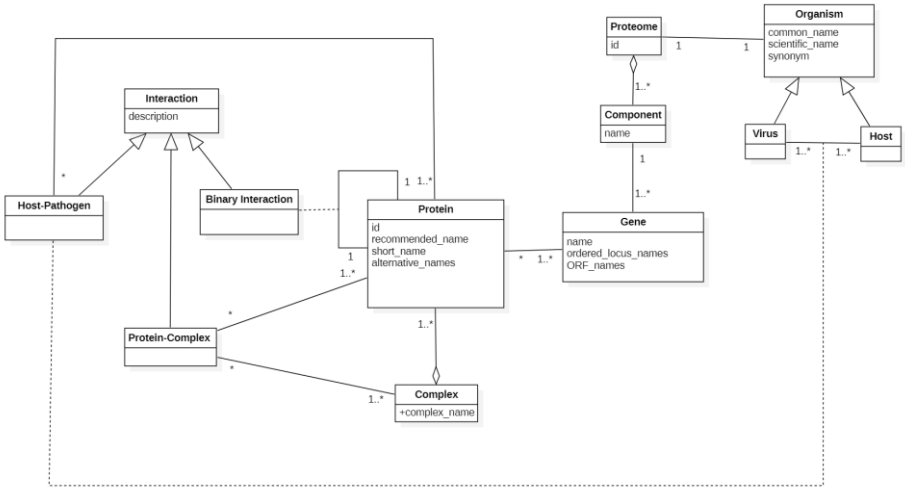

Fig. 3. Conceptual model that represents the interactions among proteins.

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 [7], IntAct, and MINT [8]. The detailed description of these repositories is considered as future work due to its importance to achieve an appropriate understanding of PPIs and to design better ways for analyzing and interpreting interactions.

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

[10] Protein-complex interactions: physical interactions among groups of proteins, without pairwise determination of protein partners.

2.3 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:

1. The primary structure is the sequence of amino acids that make up the protein.
2. The secondary structure arises from interactions between near-by 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 the secondary structure elements have folded together among each other.
4. The quaternary structure represents how its subunits are oriented and arranged with respect to one another.

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

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 primary structure. Due to different biological 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. Each protein sequence is characterized by a unique 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, called canonical sequence. The UniProtKB also provides information about the completeness of the canonical sequence (sequence status), describing if 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*, that 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, that 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 produces the different isoforms are described by the association class Event, as can be seen in Fig. 4.

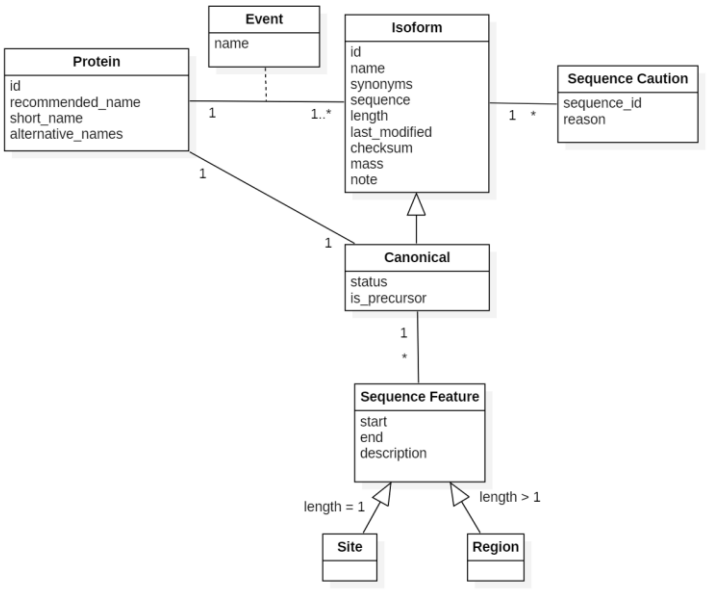

Fig. 4. Conceptual model that represents the canonical sequence and the different isoforms of a protein.

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 attribute *is precursor* of the *Canonical* class allows the description of the canonical sequence as a precursor. The PTMs are explained in detail in section 2.5.

Along the protein sequence there are interesting locations known as *sequence features*, that are classified as sites and regions depending on their length. Sequence features are represented in the conceptual model by the *Sequence Feature* class that 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 be the same. Due to the complexity and importance of these features, they will be explained in detail in section 2.4.

Secondary and Tertiary Structure. The basic elements that constitute the secondary structure of a protein are:

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

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

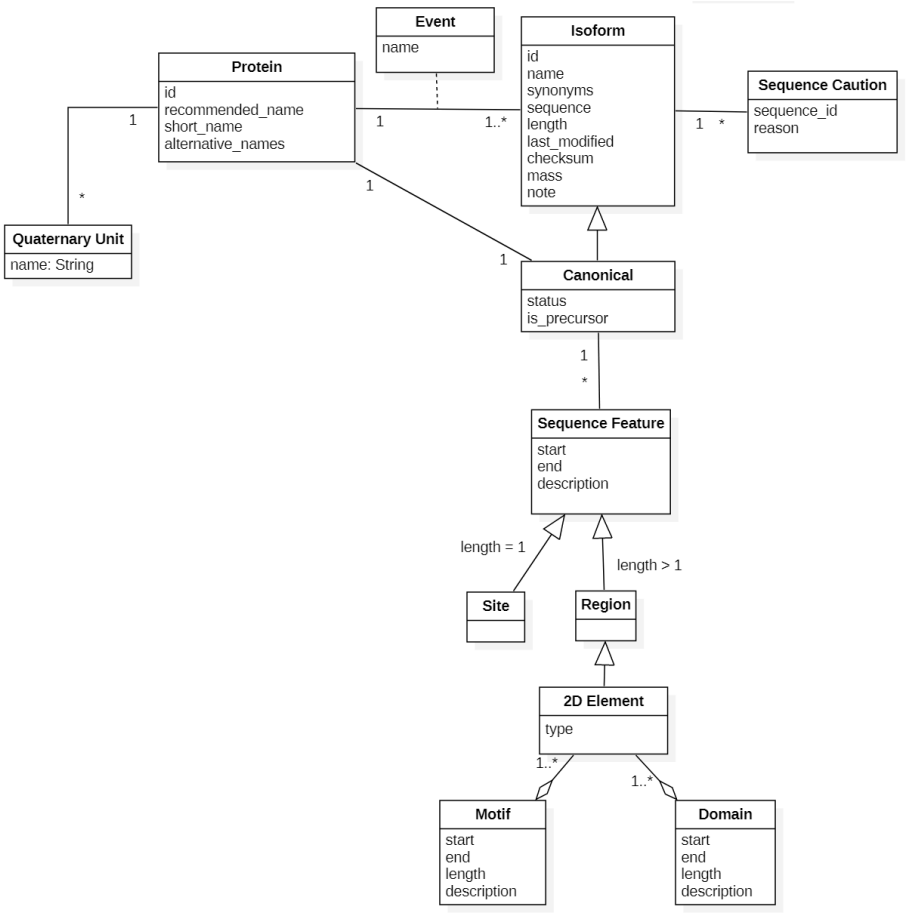

Fig. 5. Conceptual model that represents the elements that make up the 2D and 3D structure of a protein.

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 section 2.4.

Quaternary Structure. The quaternary structure of a protein represents the spatial arrangement of multiple folded protein subunits in complexes that can range from simple dimers to large homooligomers. The different subunits are represented in the conceptual model by the Quaternary Unit class that 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).

2.4 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. As it has already been mentioned in the previous section, these positions are known as *sequence features* and can be specialized into *sites* and *regions* depending on their length.

Sites. A site is described as a relevant single amino acid sequence 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. 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). 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 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 directly involved in chemical reactions. Active sites can have specific roles, represented by the role attribute, such as charge relay system (charge movement), electrophile (electron acceptor), nucleophile (electron donor), proton donor and proton acceptor. If known, the name of the specific enzymatic activity is provided (enzymatic_actibity attribute). Nucleophiles give rise to short-lived covalent intermediates whose name is also provided if available (intermediate attribute). The concept of enzyme is explained in detail at the end of this section.

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 folds and function independently of the rest of the protein [9]. 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) shared among different proteins. Motifs are unable to fold independently and often do not perform a specific function [9]. 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 subtype of coiled coil in which the amino acid leucine is predominant.

The result of the conceptualization process can be seen in Fig. 7.

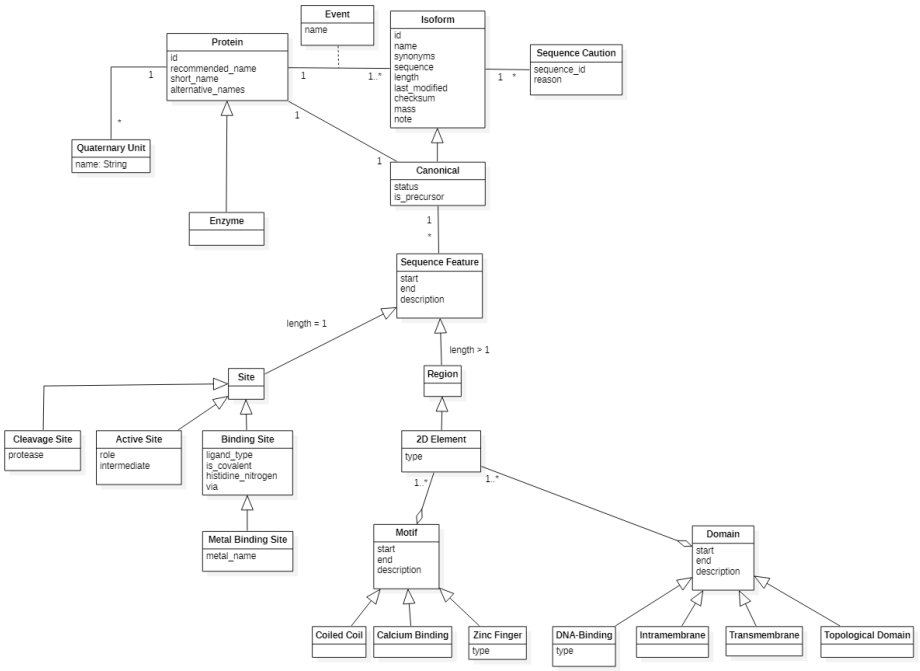

Fig. 7. Conceptual model that represents the sequence features and the different elements that constitute the three-dimensional structure of the proteins.

Another interesting and important concept that appears in this section is the notion of enzyme. Enzymes are a special type of proteins and carry out most of the chemical reactions that take place in a cell. Consequently, there are functions and characteristics (such as active sites) specific of enzymes that are not present in other types of proteins. A new class (Enzyme) has been created in the conceptual model as a specialization of the Protein class to represent enzymes.

Other interesting regions in the protein sequence are:

4. 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 (at least 4 residues in length) and large regions of compositional bias.
5. Signal peptide: Short region involved in the transport of the protein to or through the cell.
6. Propeptide: Part of a protein that is cleaved during maturation or activation.
7. 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 are provided in section 2.5. The representation of these regions, that extend the conceptual model, can be seen in Fig. 8.

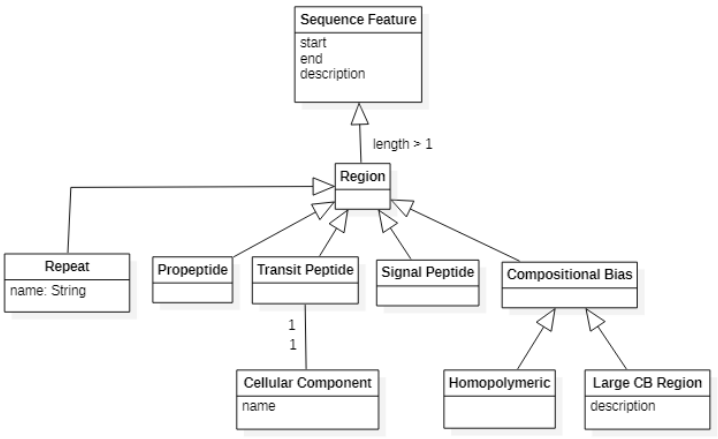

Fig. 8. Conceptual model that represents other regions of interest, some of them usually affected by post-translational modifications. The organelle where the protein is transported is represented by an association with the Cellular Component class (defined in section 2.1 to represent the organelles where the gene is produced).

Along the protein sequence there can also be repeats, that vary from short amino acid sequences to large repetitions 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 can be seen in Fig. 8.

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

2.5 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 the most common modifications occur co- and post-translationally:

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, known as Initiator Methionine.
2. Post-translational modifications (PTMs) occur once the protein has been translated and released from the ribosome. 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 PTMs includes the form of the protein that undergoes the modification, the enzyme that carries out the modification, the host (if the protein belongs to an infectious organism), the frequency of the modification and the type of relationship with any another feature (i.e. partial, alternate or transient).

Because of these protein processing events 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, that 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 provide 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. 10 shows the result of the conceptualization of the protein processing events.

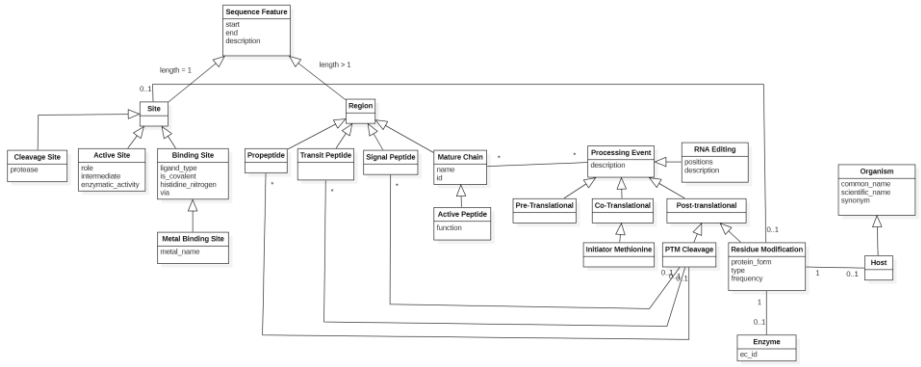

Fig. 10. Conceptual model that represents the processing events that produce a mature protein.

Other post-translational modifications are:

1. Lipidation: Process that consists in the covalent binding of a lipid group to a peptide chain. 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 attach of a glycan group (mono-, di-, or polysaccharide). 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 represented by the corresponding attribute of the N-Linked class in the conceptual model.
3. Disulfide bond: Many proteins are stabilized by disulfide bonds and involves a reaction between the sulfhydryl (SH) side chains of two cysteine residues. 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 class that connects the Protein and the Interchain classes.
4. Cross-link: Process that describes covalent linkages of various types formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links). 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.

In Fig. 11, the representation of the lipidation, glycosylation, disulfide bond and cross-link events can be seen.

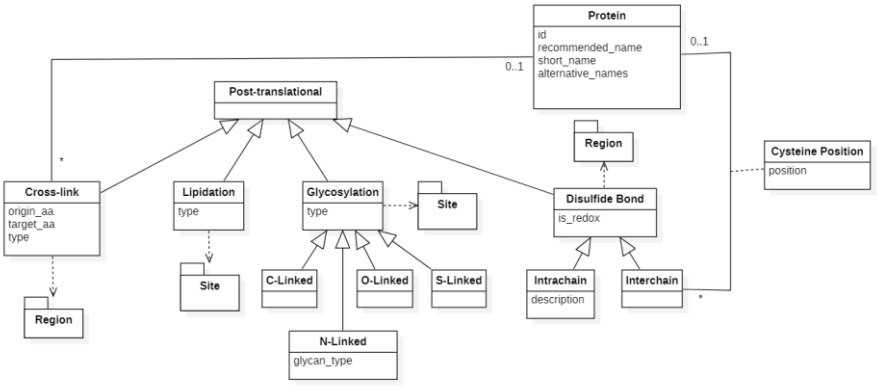

Fig. 11. Conceptual model that represents the lipidation, glycosylation, disulfide bond and cross-link events. To ease the visualization, the regions and sites associated to each class are represented as packages.

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. In the UniProtKB, 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 very precise way of representing these changes because it is not possible to specify what type of change is produced (insertion, conversion, or deletion of nucleotides) and what consequence each change has. In the conceptual model, the RNA editing is represented as a specialization of the Processing Event class (see Fig. 10).

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.

2.6 The Function of Proteins

Proteins are complex molecules that play the multitude of functions required by the cells to maintain the structure, function and regulation of tissues and organs [10]. 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.

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 resp. In some cases, each protein isoform performs a different function. Therefore, in the conceptual model the function is associated to the Isoform class and not to the Protein class.

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[11] database 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, the ec number and the external identifier associated to the reaction (a cross-reference to the corresponding database) are represented. 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. Fe^{2+}) and an identifier according to the Chemical Entities of Biological Interest (ChEBI)[12] database. 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. 12.

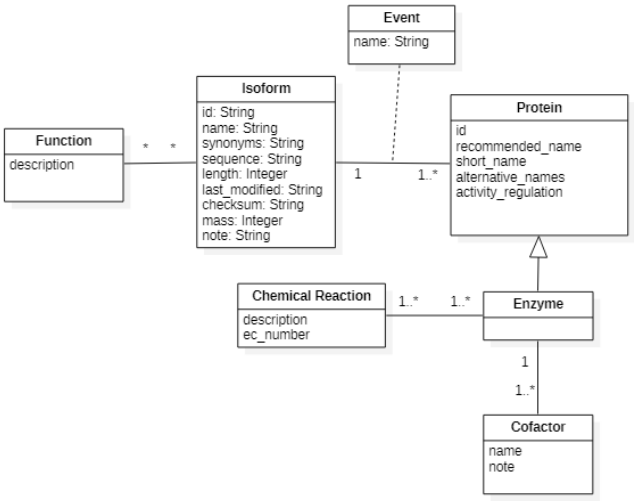

Fig. 12. Conceptual model that represents the chemical reactions catalyzed by enzymes and the cofactors required by them. The identifiers for chemical reactions and cofactors are represented by a cross-reference that points to the corresponding database. To simplify the schema visualization, cross-references are represented as packages.

[11] Rhea: https://www.rhea-db.org/
[12] ChEBI: https://www.ebi.ac.uk/chebi/

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. 12). As 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, what we consider a "conceptual weakness".

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

1. Maximal light absorption: This property indicates the wavelength at which photoreactive proteins show their maximal light absorption (e.g. 353 nm).
2. 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.
3. pH dependence: This property is used to describe the optimum pH for protein activity.
4. Redox potential: The redox potential is specific of 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).
5. 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.) what allows to add new properties if required or represent the measures 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 associated to enzymes that extends the 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 can be seen in Fig. 13. As happened with the activity regulation, the absence of protein types in the UniProtKB database forces to represent the redox potential as a generic functional property.

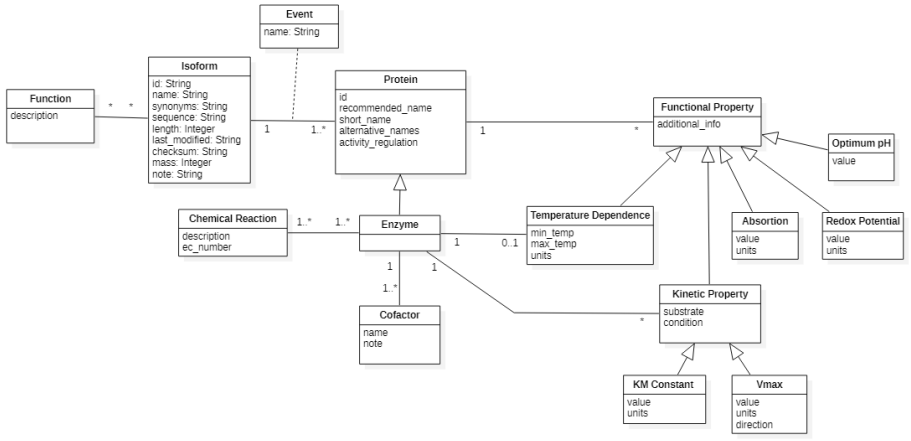

Fig. 13. Conceptual model that represents the functional properties of proteins. Kinetic properties are specific of enzymes and extends the information with the substrate and the environmental conditions required for the reaction to take place.

Along 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. But 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 work of the cells.

2.7 Biological Pathways and Subcellular Locations

The chemical reactions that occur within a cell are sequentially linked in series of steps called biological pathways. These pathways can be very complex and usually made up of different sub-pathways. Therefore, they are commonly described as a hierarchy of “super-pathway”, “pathway” and “sub-pathway”. 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 in consequence the balance of the cell, 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. The part of the conceptual model that represents the information about pathways can be seen in Fig. 14.

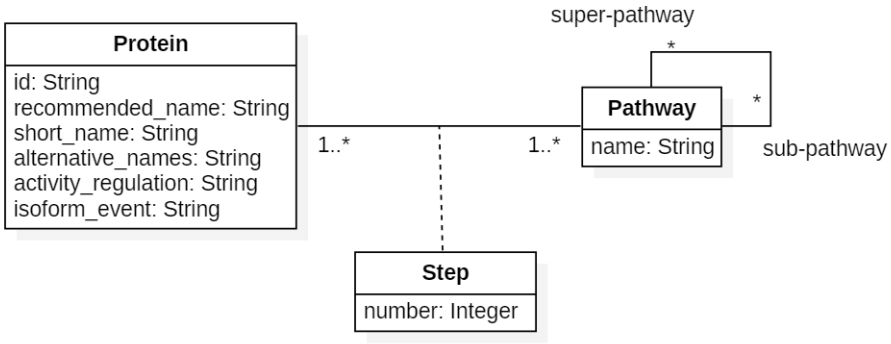

Fig. 14. Conceptual model that describes the hierarchy of pathways in which a protein takes part. The number of the step is represented as an association class.

The chemical reactions and the role that the proteins have are described in specialized databases and repositories such as KEGG[13] and Reactome[14]. The integration of all this detailed information with the conceptual model that we are elaborating is a very attractive further work that would make possible to increase the understanding of the processes that lead to disease.

Proteins have evolved to function optimally in a specific subcellular localization (nucleus, cytosol, plasmatic membrane, etc.)[11]. 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 to add additional information if required. Fig. 15 shows how this information has been modeled.

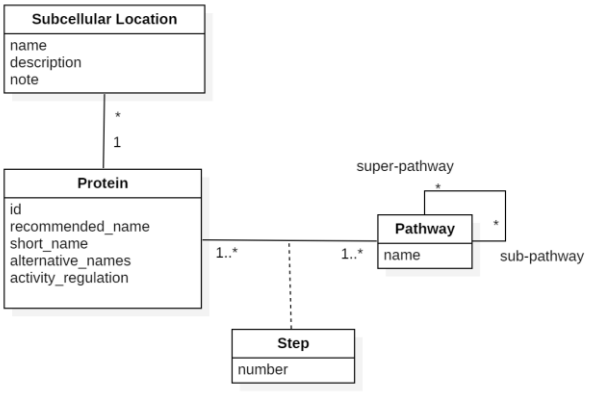

Fig. 15. Conceptual model that describes the locations where proteins perform their function.

[13] KEGG: https://www.genome.jp/kegg/
[14] Reactome: https://reactome.org/

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 conceptualization of the mechanisms that can produce such situation.

2.8 Involvement in Disease: Variants and Polymorphisms

As has been 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 DNA changes 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 [12]. Some proteins may also cause allergic reactions in certain organisms (e.g. mammals) or catalyze reactions that may cause multiple allergies.

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 the protein in the disease pathogenesis (causative, susceptibility, modifier, etc.) and links to external sources such as OMIM are also provided if available. The representation of the diseases associated to a protein in the conceptual model can be seen in Fig. 16.

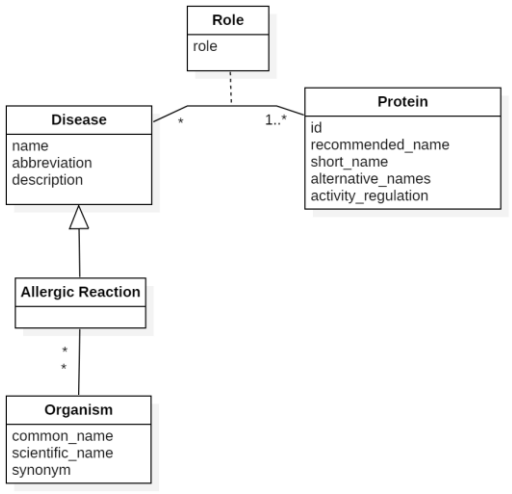

Fig. 16. Conceptual model that describes the diseases associated to a protein. The role of the protein is represented by the association class Role.

The variants that occur in the protein sequence are represented by the amino acid change (e.g. S → I), the position in the sequence (start and end regarding the canonical 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 this information is modeled by including the Variant class in the conceptual model.

Polymorphisms are a type of variant commonly due to 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[15] database, 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, what allows to express that a common polymorphism may have different frequencies in different populations. The result of the conceptualization process is depicted in Fig. 17.

[15] dbSNP: https://www.ncbi.nlm.nih.gov/snp/

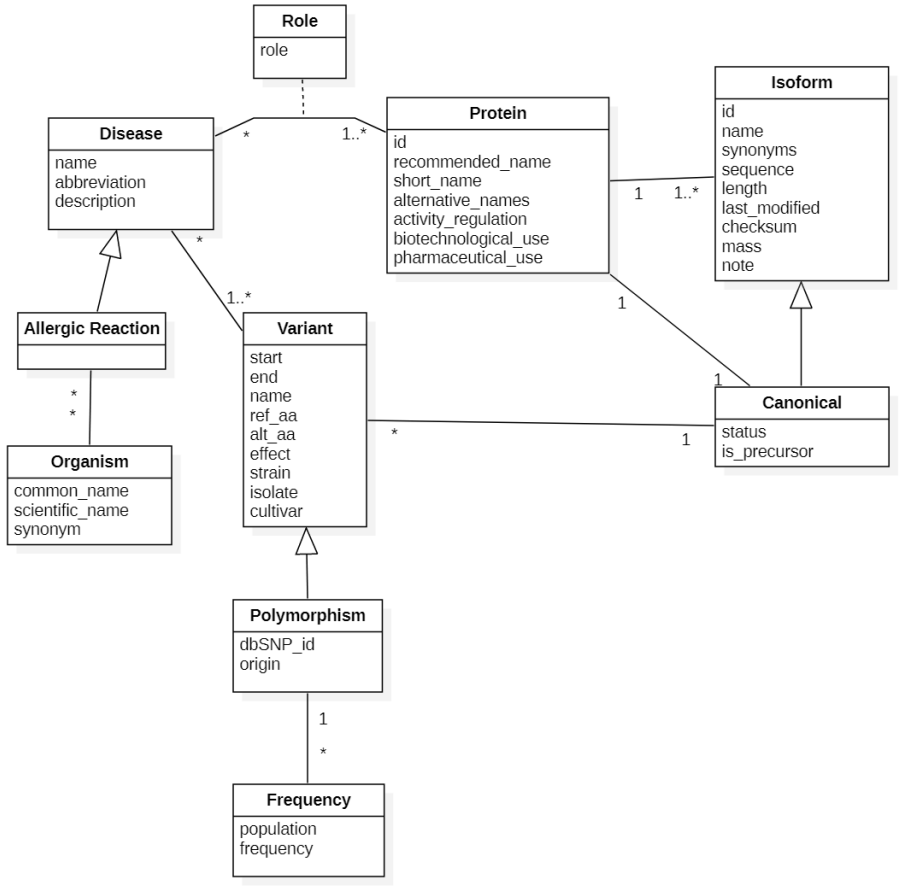

Fig. 17. Conceptual model that describes the variants that may occur in the protein sequence.

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

Some proteins have a toxic effect that can be lethal when present in a certain dose or concentration. The UniProtKB 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 can be seen in Fig. 18.

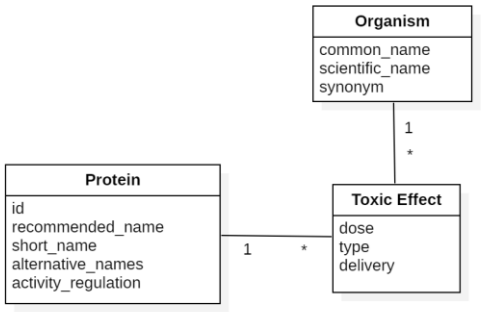

Fig. 18. Conceptual model that describes the toxic effect that proteins can cause when they are present in a certain dose or concentration.

Besides the information that have been already described in the previous sections, the UniProtKB also provides additional data associated to 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.

2.9 Additional Information

The UniProtKB database reports cross-references to other relevant sources where more specialized data can be found (sequence databases, chemistry databases, genome annotation, enzyme, and pathways, etc.). Additionally, the UniProtKB provides links to other proteins which 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.

It can also be found information about the protein entry in the UniProtKB database such as 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. The entry status, that indicates whether an entry in the database (in this case the data about the protein) has been manually annotated and reviewed by the UniProtKB curators or not. Its possible values are “Reviewed” and “Unreviewed”.
2. The annotation score, that 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 rather basic annotation and proteins with an annotation score of 5 are considered the best-annotated entries.
3. The protein existence, that 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 a 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 can be seen in Fig. 19.

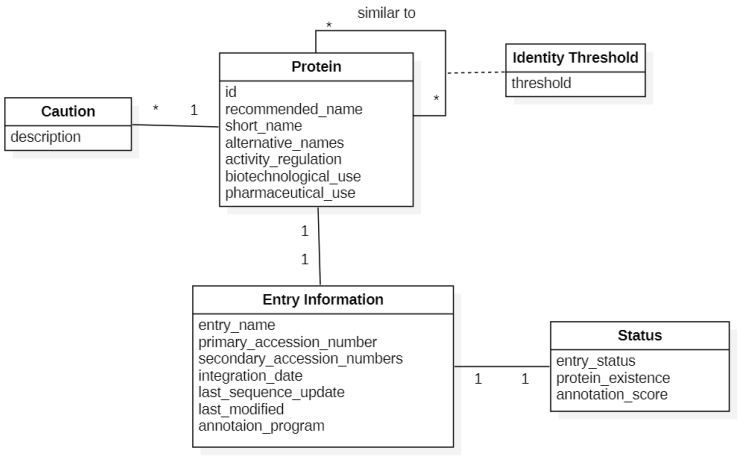

Fig. 19. Conceptual model that describes additional information about the protein entry in the database, cross-references to other sources and similarity with other proteins.

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

The evidence that supports the assertions made over the characteristics of a protein is represented as a set of “evidence tags”, that describe 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, 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 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. 20 shows how evidence and cross references are represented in the conceptual model, using the Evidence Tag, the Evidence Source and the Cross-reference classes.

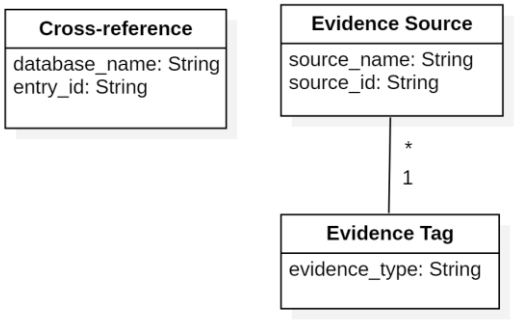

Fig. 20. Conceptual model that represents the evidence supporting protein assertions and cross-references to external sources.

Any element of the conceptual model is susceptible of 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 omitted from the figures shown along this paper. Nevertheless, in the global model all evidence tags, and cross-references are represented.

Along 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 conceptual model. The concepts have been represented as they are described in a well-know and 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.

3 Discussing the Underlying Ontological Commitment

Having a sound ontological commitment provides a shared understanding of the domain, helping to structure, share, collect and analyze data in a precise way to derive meaningful conclusions. Along this work, the different concepts about protein structure, function, and association with disease, used by the UniProtKB repository, 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 of certain types of proteins such as enzymes. Nevertheless, it is interesting to mention that the UniProtKB 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 it is very important. The specialization of proteins into different types allows to clearly determine which functions can be carried out by each type and

provides a better understanding of the domain and the information that is going to be explored. This also avoids 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, a site is defined as an "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), that contradicts the main definition provided by the documentation.

Despite the mentioned issues, the UniProtKB database can be considered a well-grounded and complete repository that represents all the relevant 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.

4 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 in 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 and 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 the user.

Along this work, an 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 represented using the 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 widely used database that provides valuable knowledge about this complex domain.

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

1. 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., known as biological context. This information is provided by specialized databases and repositories such as DIP, IntAct, and MINT.

2. Analysis of biological pathways: The chemical reactions and processes that occur within a cell determine its correct or incorrect function and therefore the healthy or unhealthy state of a living system. These reactions and processes can be very complex and detailed information about them, along with the role that the proteins have, are described in specialized databases and repositories such as KEGG[16] and Reactome[17].
3. 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)[18]. 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.

The analysis done in this work is crucial to state the need of having a sound ontological commitment in such a complex domain as genomics is. In this case, we have focused on the protein context, but the description of the genome structure and the understanding of how it works requires an 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 Polytechnic University of Valencia) [13] represents a first stone intended to build 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.

Acknowledgements

This work has been developed with the financial support of the Spanish State Research Agency and the Generalidad Valenciana under the projects TIN2016-80811-P and PROMETEO/2018/176, co-financed with ERDF.

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*

[16] KEGG: https://www.genome.jp/kegg/
[17] Reactome: https://reactome.org/
[18] Gene Ontology: http://geneontology.org/

Solutions. CB, 2019.

[3] "NIH Genetics Glossary (Protein)." [Online]. Available: https://www.genome.gov/genetics-glossary/Protein. [Accessed: 20-May-2020].

[4] R. Apweiler, "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Res.*, vol. 38, no. suppl_1, pp. D142–D148, Jan. 2010.

[5] "About UniProt." [Online]. Available: https://www.uniprot.org/help/about. [Accessed: 20-May-2020].

[6] J. de Las Rivas and C. Fontanillo, "Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks," *PLoS Comput. Biol.*, 2010.

[7] L. Salwinski, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 449D – 451, Jan. 2004.

[8] L. Licata *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D857–D861, Jan. 2012.

[9] W. R. P. Novak, "Tertiary Structure Domains, Folds, and Motifs," in *Molecular Life Sciences*, New York, NY: Springer New York, 2014, pp. 1–5.

[10] "What are proteins and what do they do?," *Genetics Home Reference*. [Online]. Available: https://ghr.nlm.nih.gov/primer/howgeneswork/protein. [Accessed: 20-May-2020].

[11] P. Dönnes and A. Höglund, "Predicting Protein Subcellular Localization: Past, Present, and Future," *Genomics. Proteomics Bioinformatics*, vol. 2, no. 4, pp. 209–215, Nov. 2004.

[12] 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.

[13] J. F. Reyes Román, Ó. Pastor, J. C. Casamayor, and F. Valverde, "Applying conceptual modeling to better understand the human genome," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.