International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 13 (2021) pp. 308-318

Submitted: 21 March, 2021; Accepted: 8 September, 2021; Published: 7 October, 2021

An MVC-inspired Approach for an Intelligent Annotation of a Protein Ontology : IA-PrOnto

Mohamed Hachem Kermani[1], Zizette Boufaida[2], Sabrina Benredjem[2] and Amani Nesrine Saker[2]

[1]National Polytechnic School - Malek Bennabi, Constantine, Algeria
LIRE Laboratory, Constantine, Algeria
hachem.kermani@enp-constantine.dz
hachem.kermani@univ-constantine2.dz

[2]University of Constantine 2 - Abdelhamid Mehri, Algeria
LIRE Laboratory, Constantine, Algeria
zizette.boufaida@univ-constantine2.dz sabrina.benredjem@univ-constantine2.dz
nesrine.saker@univ-constantine2.dz

Abstract: **Current medicine has recently recognized the limits of delivering the same treatment to different patients with the same disease. Although, for a long time, clinicians have adjusted patient's treatment according to several parameters: gender, age, weight, etc.., response rate still varies from 20% to 80% for these conventional therapies. A new medicine has therefore been developed, which involves the design of specific treatments based on the individual biological characteristics of the patient (i.e. genetic and protein information). DNA sequencing was one of these new approaches that have been developed for obtaining and analyzing genetic information. While this newly available genetic information has opened new avenues for applying personalized medicine, some issues remain to be addressed. One of these issues concerns the availability of the protein information. Therefore, and in order to provide structured knowledge about proteins, our previous research has investigated the dynamic development of a Protein Ontology: PrOnto. And in this paper, we propose an intelligent annotation approach which dynamically enrich PrOnto, the annotation method was inspired by the MVC pattern where the Model is the Protein Ontology, the View is an Intelligent User Interface that semi-automatically annotates PrOnto and the Controller is an intelligent agent that automatically annotates the Protein Ontology. Moreover, the proposed approach enables the automatic prediction of the 2D and the 3D protein structures, which will allow providing all protein information needed to annotate PrOnto with more reliable knowledge.**

Keywords: Personalized Medicine, Protein Ontologies, Model View Controller Pattern, Semi-Automatic Annotation, Automatic Annotation, Intelligent User Interfaces, Intelligent Agents, 2D/3D Protein Prediction.

I. Introduction

Proteins are vital molecules that play many important roles in the human body; they contribute to the tissue growth and maintenance, the catalysis of organic reactions, the communication between cells, tissues and organs and help improve immune health. Each protein is a macromolecule consisting of a chain of smaller molecules (i.e. amino acids) called monomers. Four levels of protein structure are distinguished: primary, secondary, tertiary, and quaternary. Amino acids are assembled through peptide bonds (i.e. an amino acid group of carboxylic acid with a neighboring amino acid group $\rightarrow$ C = O — NH) and thus form the primary structure [14]. The primary structure of a protein refers to the amino acid sequence within the polypeptide chain [7]. Then comes secondary protein structure which is the three-dimensional form of local protein segments. Alpha helices and beta sheets are the two most common secondary structural elements. Secondary structure elements typically form spontaneously as an intermediate before the protein folds into the tertiary three-dimensional structure where the α-helixes and β-pleated-sheets are folded into a compact globular structure [16].

Advances in information technologies coupled with increased knowledge about genes and proteins have opened new avenues for studying protein complexes [20]. Hence, there is a growing need to provide structured and integrated knowledge about various proteins for the study of unknown proteins, the search for new drugs and the application of personalized medicine [24]. Indeed, proteins are biological molecules that play an essential role in identifying causes of diseases. Therefore, providing structured knowledge about proteins is one of the most important and frequently studied

issues in biological and medical research [13], particularly upon the completion of the Human Genome Project [8], which helped answer the question of whether there is a unique correspondence between genes and generated proteins, opening new avenues for the study of proteins. Hence, in order to create universal protein knowledge bases, it is particularly important to find them structured representations, such as ontologies [10]. For that reason, several computational approaches have been proposed to develop ontologies integrating knowledge about proteins [23]. However, these approaches are not dynamic, instead they either transform static sources into static ontologies or develop static ontologies with a small number of concepts and properties. To address these limitations, we proposed in our previous work [15] a two-step methodology that dynamically develop a protein ontology: PrOnto. And in this paper, we propose a dynamic enrichment of PrOnto by using an intelligent annotation approach. This solution was inspired by the MVC pattern where the Model is the Protein Ontology, the View is an Intelligent User Interface that semi-automatically annotates PrOnto and the Controller is an intelligent agent that automatically annotates the Protein Ontology.

The proposed intelligent annotation system combines both semi-automatic and automatic annotation methods to dynamically enrich the Protein Ontology. Unlike some existing annotation solutions that annotate static protein sources with a limited number of concepts and properties, IA-PrOnto annotates and extends the Protein Ontology automatically and continuously with more reliable knowledge. The main aim of this approach is to increase the number of proteins knowledge integrated into PrOnto in order to allow experts to use it as a reference protein knowledge base, enabling a deeper understanding of life with medical, pharmaceutical and pathological issues. The rest of this paper is organized according to the following. Section 2 provides an overview of research that is related to our approach. Section 3 presents our proposal which is the Intelligent Annotation of the Protein Ontology. Section 4 presents a software application and experimentation. Section 5 presents a discussion. Finally, Section 6 concludes the paper and suggests some directions for future research.

II. Related Work

The more we understand our genes and how they act, the more we will be able to realize the complexity and the beauty of life. Our genes determine how we look, who we are and our risk to disease. Oncology and genetic diseases are today the main disciplines of investigation in medicine. However, improved diagnosis and therapeutic solutions are also required in these fields of investigation as in other common diseases, in diabetology, rheumatology, psychiatry, etc. Recently, and based on the consideration of individual genetic, environmental and lifestyle variants of each patient, the concept of personalized medicine has emerged as a more ambitious concept, surrounding both treatment and disease prevention. Deported from the advancement of molecular genetic knowledge and the extraordinary advancement of digital technologies. Medicine is gradually moving away from the traditional model of reactive care towards a more holistic approach. Personalized medicine, consists of designing medical treatments based on the patient's individual characteristics (i.e. genetic and protein features). The main elements that play an indispensable role in personalized medicine are genetics and protein information. For this reason, research in the past few years has focused on obtaining and understanding this information held in our cells. Hence, there is a growing need to provide knowledge on several proteins to allow the study of unknown proteins, the discovery of novel proteins, the research for new drugs and the applying of personalized medicine. In order to integrate knowledge about proteins, it is critical to develop a structured data representation for protein knowledge, such as ontology. Several computational approaches have been proposed for structuring and integrating knowledge about proteins into ontologies. Our intelligent annotation approach aims to dynamically enrich an existing ontology: PrOnto, this enriching involves semi-automatically and automatically annotates the Protein Ontology. Annotation is the process of attaching metadata to ontological concepts in order to enrich the data with information that makes it easier to discover, use and manage [11]. For bio-ontologies, two main methods, including semi-automated and automated annotation are currently used:

A. Semi-Automated Annotation

The semi-automated annotation process means that the ontology will often be manually labeled and enriched by an expert (i.e. a human annotator). Several previous studies were the subject of semi-automated bio-ontologies annotation, including [28] where the authors have proposed the Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG), a system which supports the creation and extension of OBO ontologies by semi-automatically generating terms, definitions and parent–child relations from text in PubMed, the web and PDF repositories. DOG4DAG is seamlessly integrated into OBO-Edit. It generates terms by identifying statistically significant noun phrases in text. For definitions and parent–child relations it employs pattern-based web searches. DOG4DAG systematically evaluate each generation step using manually validated benchmarks. The term generation leads to high-quality terms also found in manually created ontologies. Up to 78% of definitions have been validated and up to 54% of relationships between children and ancestors can be retrieved. Furthermore, in [9] the authors developed OMIT: Semi-Automated Ontology Development for the microRNA Domain, which makes use of machine intelligence, considers miR domain-dependent and domain-independent properties/relationships, is scalable and has significantly reduced human efforts. Experiments of this approach have been conducted to thoroughly evaluate the methodology. The authors contributions can be summarized as: (i) The development and the critical improvement of OMIT has been continued strongly based on previous research results. (ii) They explored efficient and effective algorithms with which the development of ontology can be seamlessly combined with machine intelligence and carried out in a semi-automated manner. Moreover, a methodology was presented in [21] that addresses the capture of change by predicting ontology extension. This work was motivated

by the fact that these changes can be semi-automatically discovered by analyzing the ontology data and its use. It is a supervised learning based strategy that predicts the areas of the ontology that will undergo extension in a future version based on previous versions of the ontology. By pinpointing which areas of the ontology are more likely to undergo extension, this methodology can be integrated into ontology extension approaches, both manual and semi-automated to provide a focus for extension efforts and thus contributing to ease the burden of keeping the ontology up-to-date.

B. Automated Annotation

Automatic ontological annotation (also known as automatic labeling or linguistic indexing) is the process by which a system automatically assigns metadata to ontological classes. Several computational approaches have therefore been proposed for the annotation of bio-ontologies, such as TermGenie [6]: a web-based class-generation system that complements traditional ontology development tools, in which all classes added through pre-defined templates are guaranteed to have OWL equivalence axioms that will be used for automatic classification and in some cases for inter-ontology linkage. As a result, TermGenie was used to automatically generate 4715 new classes. Morever, authors in [18] have developed LION/web: a web-based ontology enrichment tool for lipidomic data analysis. This web-based interface had allowed identification of lipid-associated terms in lipidomes. LION/web has shown significant enrichment of high membrane fluidity-related terms. Furthermore, authors in [1] have proposed a method that uses machine learning and word embedding to classify terms and phrases used in biomedical Europe PMC full-text articles to refer to an ontology class. Once labels and synonyms of a class are known, they used machine learning to identify the super-classes of a class. For this purpose, they identified lexical term variants, used word embeddings to capture context information and rely on automated reasoning over ontologies to generate features and they used an artificial neural network as classifier. The effectiveness of the method was shown by identifying terms in Human Disease Ontology that refer to diseases and distinguishing between various categories of diseases. In [17] UniProt has developed a method of annotation, known as UniRule, based on expertly curated rules, which integrates related systems (RuleBase, HAMAP, PIRSR, PIRNR) developed by the members of the UniProt consortium. UniRule uses protein family signatures from InterPro, combined with taxonomic and other constraints to select sets of reviewed proteins which have common functional properties supported by experimental evidence. This annotation is propagated to unreviewed records in UniProtKB that meet the same selection criteria, most of which do not have (and are never likely to have) experimentally verified functional annotation. Furthermore, a deep learning framework called DeepGOA [29] has been proposed to predict protein functions with protein sequences and protein-protein interaction (PPI) networks. To evaluate the performance of DeepGOA, several different evaluation methods and metrics was utilized and the experimental results showed that DeepGOA outperforms DeepGO and BLAST.

III. IA-PrOnto

Proteins are biological molecules that contribute to the maintenance of cell structure, the catalysis of organic reactions and the modulation of gene expression. They also play a crucial role in determining disease causes. The availability of structured protein knowledge is therefore one of the most important and frequently studied issues in biological and medical science. For this reason, the creation of universal protein ontologies is a very challenging task, requiring both skills in the field of ontologies modeling and design. This implies that individuals of diverse backgrounds, such as genetics, philosophy and computer science, should be involved in the ontologies development process, which is always a manual, time-consuming and costly task. Thus, a dynamic development of a Protein Ontology: PrOnto which can be used by scientists for the application of personalized medicine, the discovery of new diseases and the development of new drugs has been explored in our previous research. However, the endeavor to keep PrOnto up-to-date is a challenging and expensive task that requires many experts. A major part of this work relates to the introduction of new concepts in order to enrich the protein ontology with more knowledge. To do this, we propose an intelligent annotation approach which dynamically enrich PrOnto, the annotation method was inspired by the **MVC** pattern where the **M**odel is the Protein Ontology, the **V**iew is an Intelligent User Interface that allows experts to annotate PrOnto semi-automatically and the **C**ontroller is an intelligent agent that automatically annotates the Protein Ontology from the external protein sources (i.e. UniProt [2], Gene Ontology [5]).

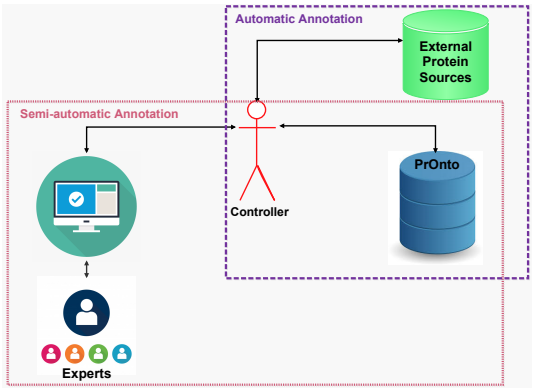

Figure. 1: The intelligent annotation of PrOnto

A. IA-PrOnto MVC Model

The intelligent annotation approach that we propose was inspired by the MVC pattern, an architectural framework that divides an application into three major conceptual components: the model, the view and the controller. Generally used for Desktop GUI applications, this pattern became popular with the advent of web apps. Today, almost all common languages support this architecture [22]. In our intelligent annotation approach, we have divided each part according to its consistency with the MVC pattern, where the Model, the central component of the framework is the Protein Ontology: PrOnto. The View, the output representation of knowledge represents an Intelligent User Interface that allows experts

to interact with the knowledge and annotate PrOnto semi-automatically. The Controller, which is an Intelligent Annotation Agent that acts as an intermediate between the Model and the View to handle both business logic and incoming requests, enabling the dynamic annotation of PrOnto.

1) The model "PrOnto"

PrOnto [15] is the model we rely on in order to accomplish our goal. This Protein Ontology includes concepts (type definitions) that are data descriptors for proteomics data and the relations between these concepts. The main features of 'PrOnto' are: (i) a hierarchical classification from generic to specific concepts (classes). (ii) an attribute list for each of the classes. (iii) a set of relationships to link the concepts. In PrOnto there are three sub-classes of 'Proteins', called generic classes that are used to define complex concepts: 'Known, Evolved and Abnormal proteins'. The 'Known protein' class includes several classes which represent the different types of proteins: Enzymatic, Immune, Transport, ...

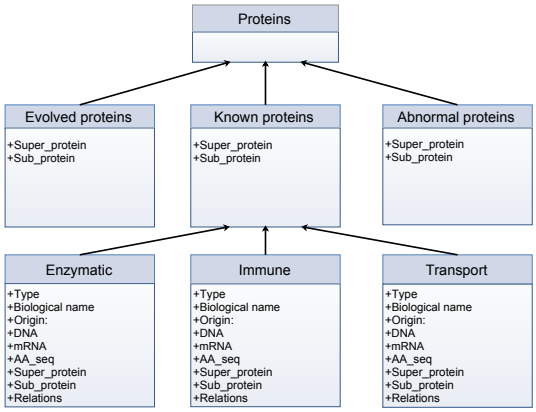

Figure. 2: The generic classes of PrOnto

Each new protein will be structured and integrated into PrOnto as a sub-class of the generic classes: 'Evolved, Abnormal, Enzymatic, Immune, Transport,... with name format 'Protein's biological name'. Example: 'Actin', 'Alpha amynase', ... Evolved and abnormal proteins sub-classes have name format like:'P0000001', 'P0000002','P000000n'. PrOnto currently comprises 15 concepts or classes, 78 attributes or properties and 83 instances. However, this Protein Ontology has been constructed in such a way that it can be constantly expanded to include a very large number of proteins.

2) The view "IUI"

We model the view in our MVC annotation architecture as an Intelligent User Interface, which is the output representation of knowledge that requires some form of artificial intelligence (i. e. autonomy, information exchange and cooperative management). Generally, an intelligent user interface means that the computer side has advanced understanding of the environment, which enables the interface to better understand the user's needs and to personalize or lead the interaction [31]. The proposed Intelligent User Interface provides experts with two features. First, it allows experts to interact with "PrOnto" by exploring or consulting the available protein knowledge. The second feature provides experts the ability to annotate current protein information with metadata and enrich PrOnto with new proteins.

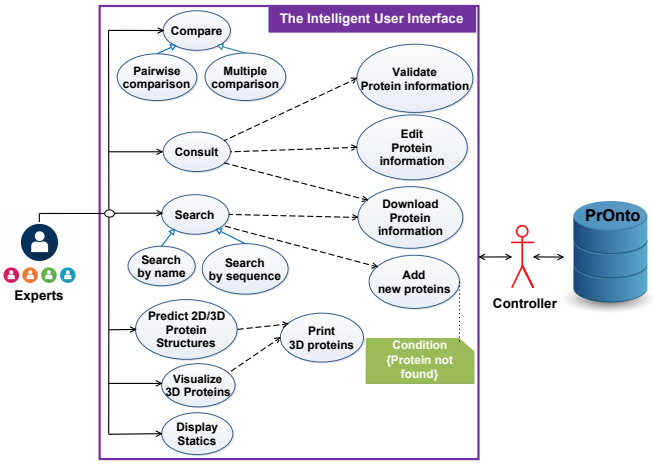

Figure. 3: IUI use cases

The proposed IUI besides transmitting user requests to the controller and leading the annotation process in order to enrich the ontology, it also proposes and constantly asks the subscribed interface experts to check, validate and enrich the current protein knowledge, although keeping track of the ontology modifications and updates to ensure the consistency of the knowledge stored in PrOnto.

3) The controller "IAA"

Which is the most important component of the MVC architecture, as it is responsible for decision-making, controlling the system logic and serving as an intermediary between the model and the view. In our annotation architecture, we exploit the advantages of agents, including autonomy, information exchange [12], and cooperative negotiation to introduce an Intelligent Annotation Agent performing the same functions as the controller to handle both the interactions between Pronto and the IUI and to dynamically enrich the Protein Ontology from external protein sources [2]. In our approach, we concentrate primarily on PrOnto evolution that can be: the population and/or the enrichment of the ontology, which are sub tasks of ontology annotation. This annotation is focused on information extracted from the existing protein sources [5]. The Intelligent Annotation Agent automatically and constantly adds complementary and missing information to a PrOnto's current protein, depending on the results of the comparison between the current protein and all the proteins found in the other existing protein sources in order to dynamically enrich Pronto. On the other side, we have a semi-automatic annotation that is achieved with the guidance of an expert who interacts directly with Pronto through the Intelligent User Interface, which enables it to validate, modify and add new proteins.

The IAA keeps track of these improvements made to the ontology each time the expert updates PrOnto. With this process, the Protein Ontology will be dynamically enriched and will include a greater number of proteins and thereby become a reference protein knowledge base that will be used for de-

veloping effective disease prevention mechanisms, personalized medicine and treatments, and other aspects of healthcare.

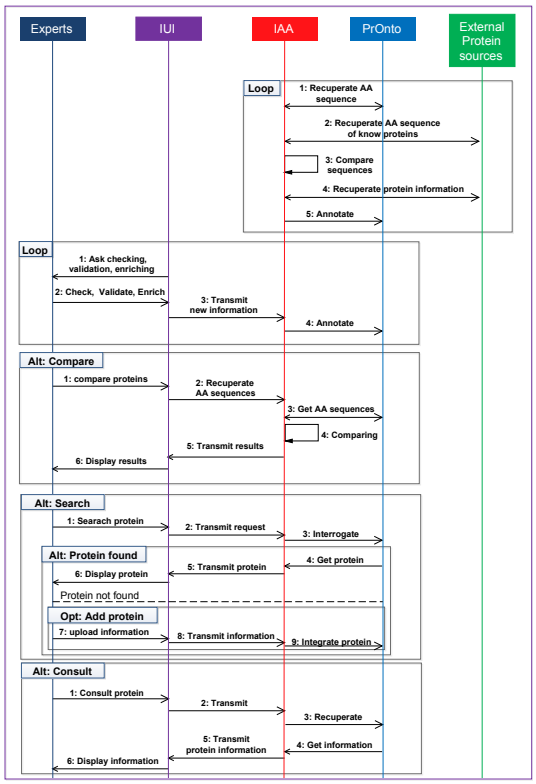

Figure. 4: Interactions of the Intelligent Annotation Agent

The Intelligent Annotation Agent acts on the basis of the automatic/semi-automatic annotation processes that we present below.

B. Semi-automatic annotation process

The key purpose of the semi-automatic annotation is to ensure the ontology's reliability and extension through the Intelligent User Interface, which helps experts to explore PrOnto in order to discover the available protein knowledge that can be used for the study of unknown proteins, the search for new drugs and the application of personalized medicine. The IUI also provides experts the ability to annotate existing protein information with metadata and to extend PrOnto with new proteins. This semi-automatic annotation process is triggered in two cases: First, when the IUI requests the subscribed interface experts to check, validate or enrich the Protein Ontology. In that case, all the information provided by the experts will be used to annotate the current protein or to classify and integrate a new protein. Second, once scientists choose to consult the existing proteins or to look for proteins. In that case, protein knowledge will be presented to scientists. In addition, the IUI provides scientists with the opportunity to validate the reliability of the knowledge or correct erroneous information. Unless the searched protein can not be found in the ontology, the IUI recommends that the scientist add it by providing the basic information of this protein. When the scientist can provide this information, the protein will be classified and inserted into PrOnto. If the scientist can only have some protein details, this protein will be classified and added as a protein with new properties (i.e. abnormal or evolved). The semi-automatic annotation is illustrated by the following algorithm:

Semi-automatic annotation algorithm

1: **Begin**
2: **If** (Search_App_Active) **Then** \\ *The expert has launched a protein search.*
3: **If** (Protein_Found == 'true') **Then**
4: **If** (Edit_Option) **Then**
5: Edit () ; *Modify protein information.*
6: **EndIf**
7: **If** (Validate_Option)
8: Validate () ; *Confirm protein information.*
9: **EndIf**
10: **Else** *Protein Not Found.*
11: **If** (Add_Option) **Then**
12: **If** (Protein_Info_Available == 'true') **Then**
13: Integrate_Known () ; *Integrate the protein as a known protein.*
14: **Else**
15: Integrate_Unknown () ; *Integrate the protein as Unknown protein.*
16: **EndIf**
17: **EndIf**
18: **EndIf**
19: **EndIf**
20: **If** (Consult_App_Active) **Then** *The expert explores proteins.*
21: **If** (Edit_Option) **Then**
22: Edit () ;
23: **EndIf**
24: **If** (Validate_Option)
25: Validate () ;
26: **EndIf**
27: **If** (Add_Option) **Then**
28: **If** (Protein_Info_Available == 'true') **Then**
29: Integrate_Known () ; *Integrate the protein as a known protein.*
30: **Else**
31: Integrate_Unknown () ; *Integrate the protein as Unknown protein.*
32: **EndIf**
33: **EndIf**
34: **EndIf**
35: **End**

This algorithm of annotation comprises three important procedures that allow the annotation of existing protein information with metadata and permit the extension of PrOnto with new proteins.

1) Validate ()

This procedure provides experts with the opportunity to validate and confirm the protein information, the 'Validate' function helps us ensure the reliability of the knowledge stored in the ontology.

2) Edit ()

The 'Edit' function permits experts to correct erroneous information and annotate current proteins with metadata, which can also ensure the reliability of the knowledge stored in PrOnto.

3) Integrate ()

The integration function ensures the automatic protein concepts structuring and classification into the ontology. The integration is performed through the application of a classification algorithm developed in our previous work [15].

The three functions (i.e. Validate, Edit and Integrate) generate a detailed and an automated recording of all ontology updates, by whom and on what date they were initiated. The

intention of this recording is to determine the sources of information from which the knowledge was obtained in order to control the different processes.

C. Automatic annotation process

In our approach, the automated annotation involves extracting information derived from existing protein sources [2, 5] in order to automatically and constantly integrate new proteins or to add additional and missing information to current proteins, based on the results of the protein comparison with all known proteins in other existing protein sources.

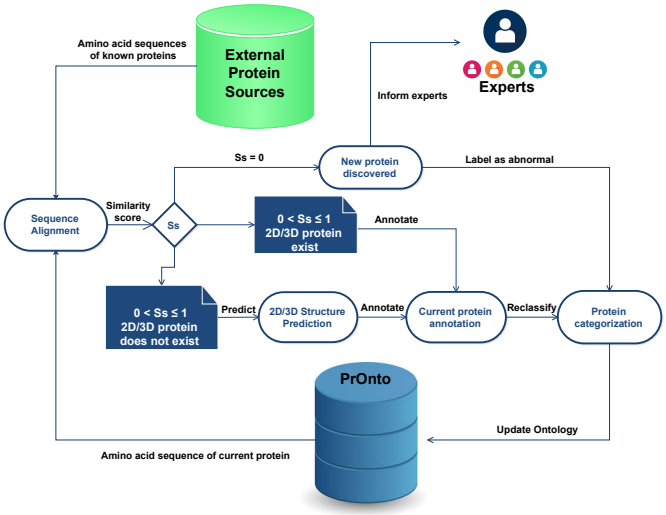

Figure. 5: The automatic annotation process

The intelligent Annotation Agent ensures the dynamic (i.e. automated and ongoing) application of this process as described in the following algorithm:

Automatic annotation algorithm

1: **Input:** X *// The AAS retrieved from PrOnto.*
2: **Input:** Seq$_n$ *// The AAS extracted from known protein sources.*
3: **Begin**
4: **Compare** (X, Seq$_n$);
5: **If** (PSS (X, Seq$_n$) = 0) **Then**
6: **Inform** (); *// Inform experts that a new protein has been identified.*
7: **Tag** (X); *// label X as an abnormal protein.*
8:**Else** *// PSS (X, Seq$_n$) ≠ 1*
9:**Recuperate** (C(Seq$_n$)); *// The corresponding protein C.*
10: **Recuperate** (PC(Seq$_n$)); *// The partial corresponding protein PC.*
11: **If** (2D.3D == 'false') **Then** *// No 2D or 3D structure.*
12: **Predict 2D/3D** (); *// Predict automatically the 2D/3D structures.*
13: **EndIf**
14: **Annotate** (X(C, PC)); *// Annotate X*
15: **Reclassify** (X); *// Reclassify X in the ontology*
16: **EndIf**
17: **End**

The automated annotation algorithm comprises three important functions:

1) Compare ()

We developed a multiple sequence alignment technique to compare a current protein contained in PrOnto with all known proteins available in the existing protein sources in order to recuperate metadata and missing information, which will be used to automatically annotate PrOnto's current proteins. The proposed sequence alignment technique consists to recuperate all protein sequences from the available sources and align them with PrOnto's current protein sequence. This alignment creates a similarity score matrix between the current protein sequence, denoted below as X and all known protein sequences, denoted as Seq1, Seq2,...Seq$_n$

	X	Seq1	Seq2	Seq3	Seq$_n$
X	Ss = 1	Ss (X, Seq1)	Ss (X, Seq2)	Ss (X, Seq3)	Ss (X, Seq$_n$)
Seq1	Ss (Seq1, X)	Ss = 1	Null	Null	Null
Seq2	Ss (Seq2, X)	Null	Ss = 1	Null	Null
Seq3	Ss (Seq3, X)	Null	Null	Ss = 1	Null
Seq$_n$	Ss (Seq$_n$, X)	Null	Null	Null	Ss = 1

The pairwise similarity score between X and Seq1, Seq2, Seq3, ...Seq$_n$ depends on the similarities and dissimilarities between the amino acids in each sequence position. A correspondence between the amino acids is counted as 1, C = 1, and a dissimilarity or a gap in the case of local alignment is counted as 0, D = 0. The pairwise similarity score is calculated as follows:

$$Ss(X, Seq_n) = \frac{\sum C, D}{NAA} \quad (1)$$

Where C and D represent the similarities and dissimilarities between the amino acids and NAA represents the number of amino acids constituting the sequence, as illustrated in the following example:

X:	Lys	-	Glu	-	Thr	-	Lys
Seq1:	Lys	-	Glu	-	Thr	-	Lys
	1		1		1		1

$$Ss(X, Seq1) = \frac{\sum C, D}{NAA} = \frac{4}{4} = 1 \quad 100\,\% \quad (2)$$

The calculation of all the pairwise similarity scores will enable to get the following similarity score matrix.

Sequences	Seq1	Seq2	Seq3	Seq$_n$
X	Ss (X, Seq1) = 1	Ss (X, Seq2) = 0.4	Ss (X, Seq3) = 0	Ss (X, Seq$_n$) = 0.2

Based on this similarity score matrix we can calculate the Pairs Similarity Sum as below:

$$PSS(X, Seq_n) = \sum Ss(X, Seq_n) \quad (3)$$

The multiple alignment results will be one of the following cases:

1. PSS (X, Seq$_n$) = 0 : The current protein does not match any known protein. In that case, the IAA will notify experts that an abnormal protein has been identified in PrOnto and it will be labeled as an unknown protein.

2. PSS (X, Seq$_n$) ≠ 0 : The current protein sequence matches perfectly a sequence of a known protein and/or

matches partially some known proteins. In this case, we will have two different situations based on the similarity score of each pair:

- Ss (X, Seq_n) = 1 : The current protein sequence matches perfectly a sequence of a known protein and also matches partially some known proteins. In this case, all missing information will be added to the current protein from the external protein sources including the 2D and 3D protein structures if they exist, otherwise the 2D and 3D structures will be automatically predicted based on a machine learning technique which is a work in progress that will be presented in our future work. Furthermore, relationships between the current protein and all partially similar proteins will be automatically generated.
- $0 < Ss < 1$: The current protein sequence partially matches some known proteins. The Intelligent Annotation Agent will label the current protein as a potential evolved protein and will generate automatically relationships between the current protein and all partially similar proteins.

2) Annotate ()

For both semi-automatic and automatic annotation the 'Annotate' function ensures that any missing information in current proteins will be added semi-automatically by the experts or automatically from the external protein sources. This additional information can be: any important metadata (i.e. Protein type, origin, biological name, features, functions, activities, relationships,...) or the 2D/3D protein structures as only the primary protein structure is represented in PrOnto's current proteins.

3) Reclassify ()

This function enables PrOnto's protein concepts to be re-categorized each time a protein is annotated, as described in the state chart diagram below:

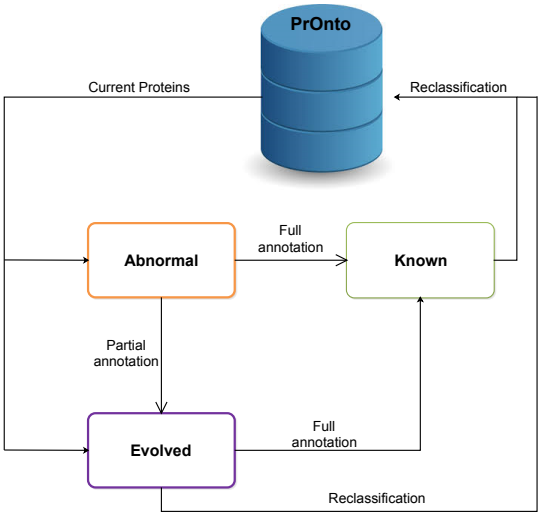

Figure. 6: The current protein re-categorization

First, assuming that the PrOnto's current abnormal or evolved protein has been annotated with the same knowledge as a known protein, this will generate an automated re-categorization from the Abnormal or Evolved class to the Known class.

Secondly, an automatic re-categorization from the abnormal class to the evolved class will then be triggered when the PrOnto's current abnormal protein has been partially annotated with the same knowledge of a known protein.

IV. Software application and experiment

We have developed a software application according to the IA-PrOnto MVC model to simulate and to experiment the proposed intelligent annotation approach. The software application has been developed as a back and front platform that permits both semi-automatic and automatic PrOnto annotation.

Figure. 7: IA-PrOnto Platform

A. Platform back-end

The platform back-end represents the behaviour of the Intelligent Annotation Agent, which automatically and continuously annotates and enriches the Protein Ontology.

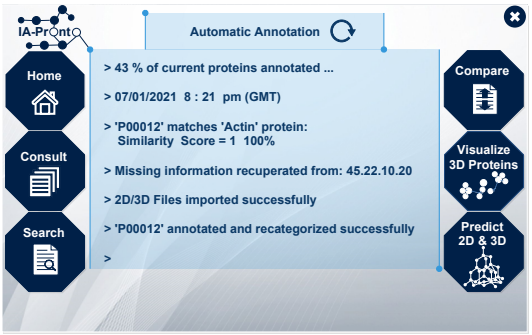

Figure. 8: Back-End of the IA-PrOnto Platform

B. Platform front-end

The proposed Intelligent User Interface is the platform's front-end, allowing experts to interact with "PrOnto" by discovering and consulting the available protein knowledge. Experts can consult proteins, search for proteins, compare proteins, predict 2D/3D protein structures and visualize 3D protein models by using the platform's front-end.

1) Consulting available proteins

As shown in Figure 9, the platform provides experts all the protein information required to be used in scientific research.

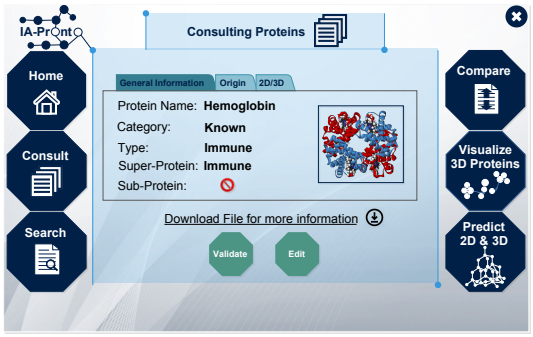

Figure. 9: Consulting interface

In addition, the platform provides experts the opportunity to validate the reliability of knowledge or to correct inaccurate information. This option allows us to ensure the semi-automatic annotation of PrOnto.

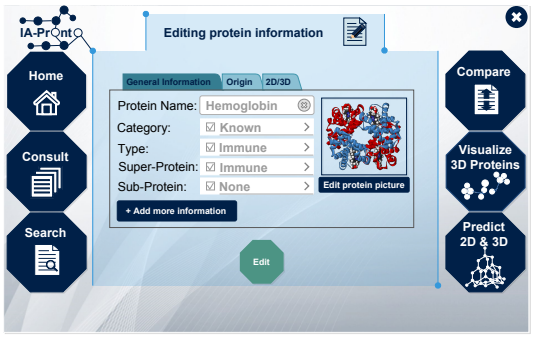

Figure. 10: Protein information editing

2) Searching proteins

The Intelligent User Interface allows experts to search for proteins by name or by the amino acid sequence.

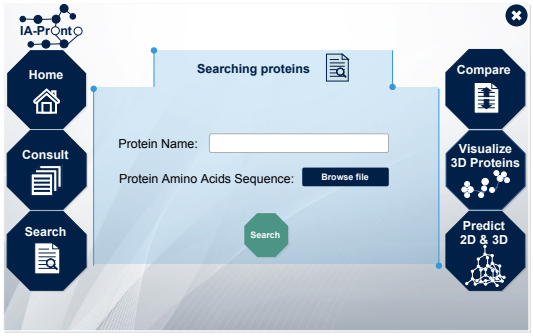

Figure. 11: Searching proteins

If the desired protein is not found in PrOnto, the IUI advises the scientist to include it by providing the protein's basic information.

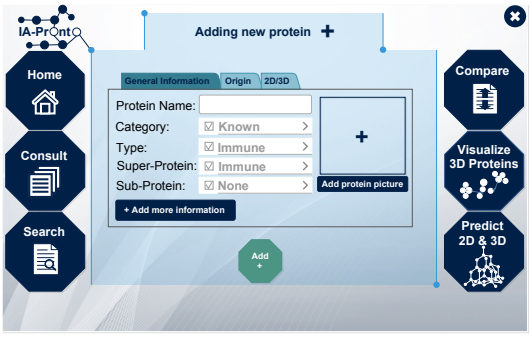

Figure. 12: Adding a new protein

This option allows us to ensure the evolution of the Protein Ontology.

3) Comparing proteins

Scientists can use the platform to compare proteins in order to find mutations and variations between a protein under study and a reference protein in PrOnto.

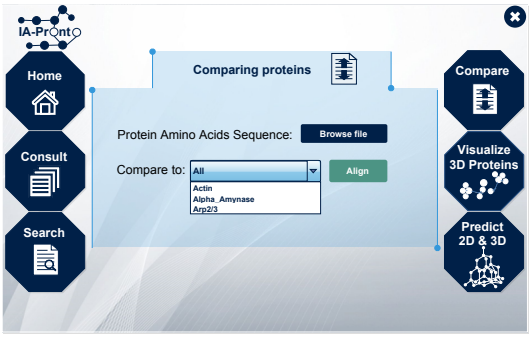

Figure. 13: Comparing proteins

4) Visualizing 3D protein Models

The platform proposes a 3D visualisation that allows all protein knowledge to be provided.

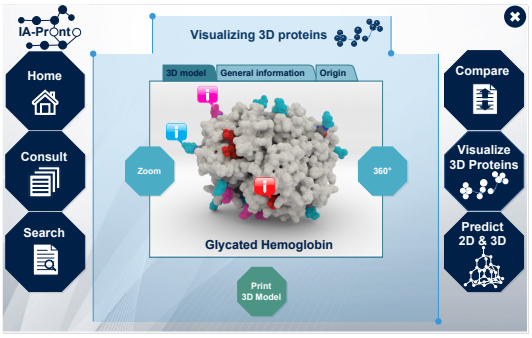

Figure. 14: 3D protein visualization

C. *Experiment*

We engaged the help of multiple experts who agreed to use the platform in order to evaluate our approach. The experiment lasted four months and gave us the following results:

	Protein Knowledge			**Protein Categories**		
	Instances	Attributes	Relations	Known	Evolved	Abnormal
PrOnto_V0	**83**	**78**	**86**	**13**	**17**	**53**
Month1	+316	+188	+336	+280	+20	-8
Month2	+180	+204	+208	+156	+12	+15
Month3	+132	+88	+152	+120	+12	-13
Month4	+32	+60	+52	+68	+4	-6
PrOnto_V1	**743**	**618**	**834**	**637**	**65**	**41**

Table 1: Experiment results1

These results indicate how the number of proteins Knowledge (i.e. instances, attributes, relationships) increased remarkably over the four months that our platform was in use, resulting in a higher number of known and evolved proteins integrated into PrOnto than before the platform was used (Figure 15).

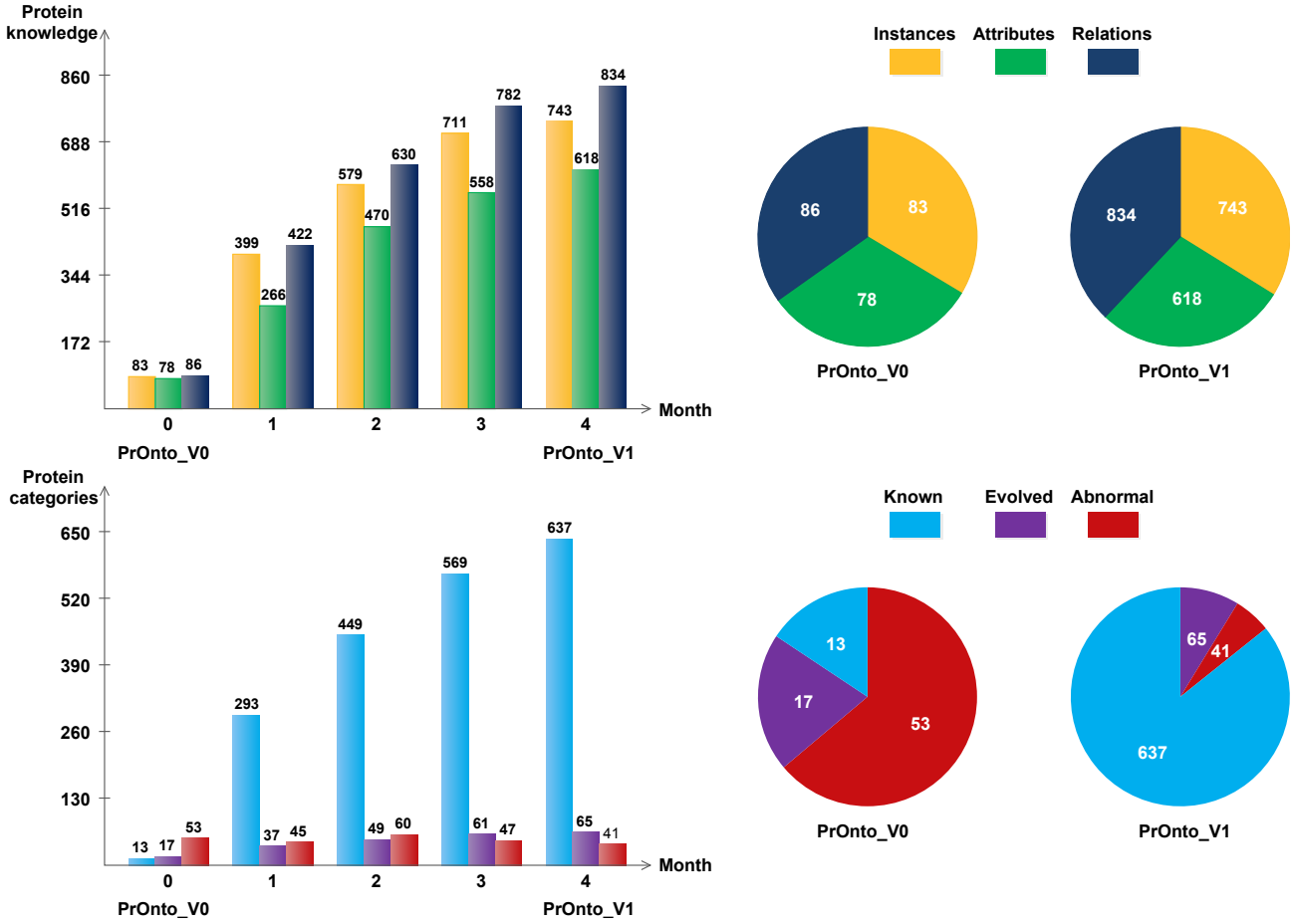

Figure. 15: Results1 analysis

In addition, the four-month experiment has shown that 413 proteins were annotated and additional metadata including secondary and tertiary protein structures have been added to the existing proteins.

	Protein annotated		Protein structures		
	Semi Automatic	Automatic	Primary	2D	3D
PrOnto_V0	**-**	**-**	**83**	**0**	**0**
Month1	+50	+31	+316	+68	+56
Month2	+59	+24	+180	+96	+52
Month3	+60	+57	+132	+112	+68
Month4	+66	+66	+32	+96	+80
PrOnto_V1	**235**	**178**	**743**	**372**	**256**

Table 2: Experiment results2

According to these results we can notice that 43% of the 413 annotated proteins were automatically annotated by the Intelligent Annotation Agent, while the others were semi-automatically annotated under expert guidance. We can also see that the intelligent annotation process has enabled the three protein structures (primary, secondary, and tertiary) to be fully identified, which allows all protein information to be available at the three different structures. These information will be used to better understand the functions and the activities of the proteins to develop effective mechanisms for disease prevention, personalized medicine and treatments, and other healthcare aspects.

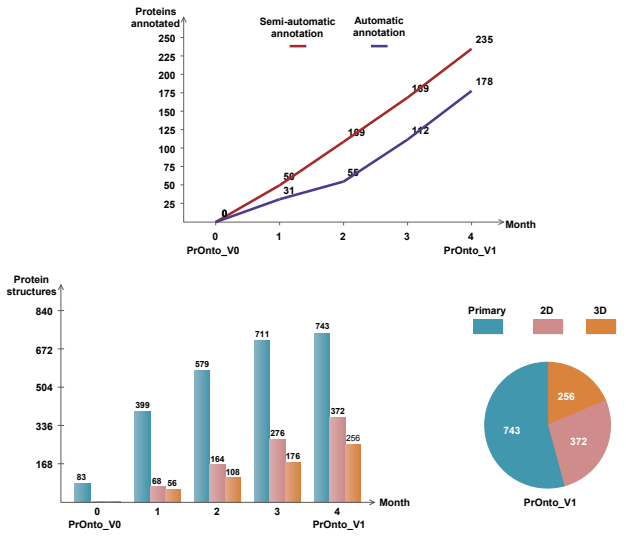

Figure. 16: Results2 analysis

In this experiment, we were able to evaluate the proposed intelligent annotation approach, which dynamically annotates and enriches the Protein Ontology. Moreover, the findings of the experiment showed how the proposed solution can extend the ontology in a few months and thereby becomes a reference protein knowledge base that integrates a vast amount of structured and reliable information to enable a better understanding of protein functions and activities, allowing the analysis of unknown proteins, the discovery of new therapies, and the application of personalised medicine.

V. Discussion

Protein knowledge are diverse, evolving and rising at a faster rate, making it more difficult for researchers to keep up with current discoveries. As a result, ontologies have played an important role in addressing this issue because they provide a structured model for representing knowledge, enabling its collection, dissemination, and computational study [19]. Keeping a Protein ontology up to date is a time-consuming and costly process that requires the involvement of many experts. A large portion of this work is dedicated to the inclusion of new proteins and the annotation of existing proteins with additional metadata. Extending and annotating bio-ontologies is a challenging task that requires both modelling and design abilities. This means that people with a multitude backgrounds, such as biology, philosophy and computer science should be involved in the annotation process, which is often a manual, time-consuming and expensive process. These challenges caused the development of computational approaches able to support semi-automatic and automatic annotation of bio-ontologies.

	Semi-automatic Annotation	Automatic Annotation
[28]	**X**	
[6]		**X**
[9]	**X**	
[21]	**X**	
[18]		**X**
[1]		**X**
[17]		**X**
[27]	**X**	
[4]		**X**
[26]	**X**	
[29]		**X**
[25]		**X**
[3]		**X**
[30]		**X**
IA-PrOnto	**X**	**X**

Table 3: Comparative study

The proposed MVC-inspired approach for Intelligent Annotation of PrOnto combines both semi-automatic and automatic annotation methods in order to better guide the annotation process and ensure the reliability and the evolution of the protein knowledge. Unlike some existing annotation solutions [28, 9, 6, 1] that annotate semi-automatically or automatically static protein sources with a limited number of concepts and properties, IA-PrOnto annotates and extends the Protein Ontology automatically and continuously with more reliable knowledge. Furthermore, and as demonstrated in the four-month experiment the intelligent annotation approach has remarkably increased the number of protein knowledge integrated in PrOnto. This increase has affected the amount of protein instances, attributes, relationships and structures, causing a reclassification of protein categories (i.e. known, evolved, abnormal), which has permit to provide all protein knowledge needed for developing effective disease prevention mechanisms, personalized medicine and treatments and other aspects of healthcare.

There are actually 743 instances, 618 attributes or properties, and 834 relationships in PrOnto. In contrast to the exist-

ing protein ontologies (UniProt [2] and Gene Ontology [5]), which contain thousands of proteins, the number of proteins in PrOnto is still insufficient.

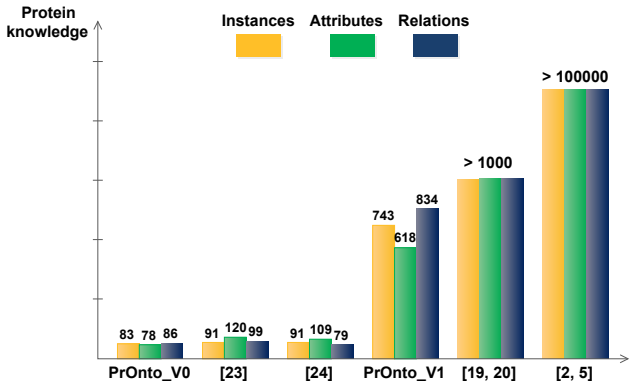

Figure. 17: Comparison with the existing protein ontologies

PrOnto will continue to be dynamically enriched and expanded as long as life remains and it will contain a large variety of proteins.

VI. Conclusion

Genetic and protein information are crucial for further understanding life and addressing problems in medical, pharmaceutical and pathological fields. For that reason, in recent years, researchers have concentrated their efforts on acquiring and comprehending these information stored in cells. Indeed, the ability to sequence the genetic code of various organisms, ranging from simple bacteria and viruses to the genetic code of humans has made genetic information widely available. Despite the fact that this newly available genetic material has opened up new avenues for a deeper understanding of life, there are still certain issues to be resolved. One of these is the availability of protein information. Therefore, to make protein information available, it is necessary to develop a structured data representation, such as protein ontologies. Thus, PrOnto was created to provide a reference protein knowledge base that can be used for disease prevention, personalised medicine and therapies, as well as other aspects of healthcare. However, in order to keep PrOnto updated, we have proposed a dynamic (i.e. automatic and continuous) annotation solution. Indeed, the MVC pattern's features (such as simultaneous development, high cohesion, low coupling, and simplicity of modification) inspired us to propose an MVC approach for the intelligent annotation of PrOnto. The main aim of this approach was to annotate and enrich the Protein Ontology with a large number of additional metadata, allowing PrOnto to be updated dynamically. Furthermore, the software application and experiment has demonstrated the dynamic extension of the ontology with 660 more proteins and 748 additional relations between the proteins. Moreover, the proposed annotation solution has showed the involvement of the Intelligent User Interface and the Intelligent Annotation Agent in better understanding the user's needs and personalizing or guiding the annotation process. Our proposal, however, is not without drawbacks. We intend to address these issues in the future by proposing a machine learning method for secondary and tertiary structure prediction as our approach is currently based solely on sequence alignment technique to automatically predict the 2D/3D protein structures.

Acknowledgments

The authors acknowledge support from the General Directorate of Scientific Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research, Algeria.

References

[1] Althubaiti, S., Kafkas, Ş., Abdelhakim, M., Hoehndorf, R.: Combining lexical and context features for automatic ontology extension. Journal of biomedical semantics **11**(1), 1–13 (2020)

[2] Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., et al.: Uniprot: the universal protein knowledgebase in 2021. Nucleic Acids Research (2020)

[3] Bouziane, H., Chouarfia, A.: Use of chou's 5-steps rule to predict the subcellular localization of gram-negative and gram-positive bacterial proteins by multi-label learning based on gene ontology annotation and profile alignment. Journal of integrative bioinformatics **18**(1), 51–79 (2021)

[4] Bukhari, A.C., Nagy, M.L., Krauthammer, M., Ciccarese, P., Baker, C.J.: Bim: an open ontology for the annotation of biomedical images. In: ICBO (2015)

[5] Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., et al.: The gene ontology resource: enriching a gold mine. Nucleic Acids Research **49**(D1), D325–D334 (2021)

[6] Dietze, H., Berardini, T.Z., Foulger, R.E., Hill, D.P., Lomax, J., Osumi-Sutherland, D., Roncaglia, P., Mungall, C.J.: Termgenie–a web-application for pattern-based ontology class generation. Journal of biomedical semantics **5**(1), 1–13 (2014)

[7] Haynie, D.T., Xue, B.: Superdomains in the protein structure hierarchy: The case of ptp-c2. Protein Science **24**(5), 874–882 (2015)

[8] Heather, J.M., Chain, B.: The sequence of sequencers: the history of sequencing dna. Genomics **107**(1), 1–8 (2016)

[9] Huang, J., Dang, J., Borchert, G.M., Eilbeck, K., Zhang, H., Xiong, M., Jiang, W., Wu, H., Blake, J.A., Natale, D.A., et al.: Omit: dynamic, semi-automated ontology development for the microrna domain. PLoS One **9**(7), e100855 (2014)

[10] Izhar, T.A.T., Apduhan, B.O.: Cloud based enterprise global ontology for information enterprise: A proposed

framework. International Journal of Computer Information Systems and Industrial Management Applications **10**, 1–7 (2018)

[11] Izhar, T.A.T., Torabi, T., Bhatti, M.I.: Using ontology to incorporate social media data and organizational data for efficient decision-making. International Journal of Computer Information Systems and Industrial Management Applications **9**(2017), 9–22 (2017)

[12] Kermani, M.H., Boufaida, Z.: A modeling of a multi-agent system for the protein synthesis. In: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA). pp. 1–7. IEEE (2015)

[13] Kermani, M.H., Boufaida, Z.: A state of art on biological systems modeling. In: 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES). pp. 712–715. IEEE (2016)

[14] Kermani, M.H., Boufaida, Z.: A2pf: An automatic protein production framework. In: International Conference on Intelligent Systems Design and Applications. pp. 80–91. Springer (2020)

[15] Kermani, M.H., Guessoum, Z., Boufaida, Z.: A two-step methodology for dynamic construction of a protein ontology. IAENG International Journal of Computer Science **46**(1) (2019)

[16] Kumari, I., Sandhu, P., Ahmed, M., Akhter, Y.: Molecular dynamics simulations, challenges and opportunities: a biologist's prospective. Current Protein and Peptide Science **18**(11), 1163–1179 (2017)

[17] MacDougall, A., Volynkin, V., Saidi, R., Poggioli, D., Zellner, H., Hatton-Ellis, E., Joshi, V., O'Donovan, C., Orchard, S., Auchincloss, A.H., et al.: Unirule: a unified rule resource for automatic annotation in the uniprot knowledgebase. Bioinformatics **36**(17), 4643–4648 (2020)

[18] Molenaar, M.R., Jeucken, A., Wassenaar, T.A., van de Lest, C.H., Brouwers, J.F., Helms, J.B.: Lion/web: A web-based ontology enrichment tool for lipidomic data analysis. GigaScience **8**(6), giz061 (2019)

[19] Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J., Chang, T.C., Hu, Z., Liu, H., Smith, B., Wu, C.H.: Framework for a protein ontology. In: BMC bioinformatics. vol. 8, p. S1. BioMed Central (2007)

[20] Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J.A., Bult, C.J., Caudy, M., Drabkin, H.J., D'Eustachio, P., Evsikov, A.V., Huang, H., et al.: The protein ontology: a structured representation of protein forms and complexes. Nucleic acids research **39**(suppl_1), D539–D545 (2010)

[21] Pesquita, C., Couto, F.M.: Predicting the extension of biomedical ontologies. PLoS Comput Biol **8**(9), e1002630 (2012)

[22] Sharan, K.: Model-view-controller pattern. In: Learn JavaFX 8, pp. 419–434. Springer (2015)

[23] Sidhu, A.S., Dillon, T.S., Chang, E.: Creating a protein ontology resource. In: Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE. pp. 220–221. IEEE (2005)

[24] Sidhu, A.S., Dillon, T.S., Chang, E., Sidhu, B.S.: Protein ontology: vocabulary for protein data. In: Information Technology and Applications, 2005. ICITA 2005. Third International Conference on. vol. 1, pp. 465–469. IEEE (2005)

[25] Spetale, F.E., Murillo, J., Villanova, G.V., Bulacio, P., Tapia, E.: Fgga-lnc: automatic gene ontology annotation of lncrna sequences based on secondary structures. Interface Focus **11**(4), 20200064 (2021)

[26] Tao, C., Song, D., Sharma, D., Chute, C.G.: Semantator: Semantic annotator for converting biomedical text to linked data. Journal of biomedical informatics **46**(5), 882–893 (2013)

[27] Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio, S., Jonquet, C.: Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes. BMC bioinformatics **19**(1), 1–26 (2018)

[28] Wächter, T., Schroeder, M.: Semi-automated ontology generation within obo-edit. Bioinformatics **26**(12), i88–i96 (2010)

[29] Zhang, F., Song, H., Zeng, M., Wu, F.X., Li, Y., Pan, Y., Li, M.: A deep learning framework for gene ontology annotations with sequence-and network-based information. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020)

[30] Zhang, Y.H., Zeng, T., Chen, L., Huang, T., Cai, Y.D.: Determining protein–protein functional associations by functional rules based on gene ontology and kegg pathway. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics **1869**(6), 140621 (2021)

[31] Zubkova, T., Tagirova, L.: Intelligent user interface design of application programs. In: Journal of Physics: Conference Series. vol. 1278, p. 012026. IOP Publishing (2019)

Author Biography

Mohamed Hachem Kermani received his Ph.D in Computer sciences from the University of Constantine 2 - Abdelhamid Mehri, Algeria in 2019. He is currently Associate Professor at the National Polytechnic School - Malek Bennabi, Constantine, Algeria. His research areas are Information Systems, Multi-Agents Systems, Ontologies Development and Bioinformatics.