A Two-Step Methodology for Dynamic Construction of a Protein Ontology

Mohamed Hachem Kermani, *Member, IAENG,* Zahia Guessoum and Zizette Boufaida

***Abstract*—Proteins are biological molecules that contribute to maintaining cell structure, catalysis of organic reactions, and regulation of the gene expression. They also 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—particularly after the conclusion of the Human Genome Project, 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. We propose a two-step methodology that first translates DNA sequences into amino acid sequences based on a multi-agent approach and then uses the results to construct a dynamic protein ontology. We also present a software application that allows scientists to use the ontology as a reference protein knowledge base for developing effective disease prevention mechanisms, personalized medicine and treatments, and other aspects of healthcare.**

***Index Terms*—DNA sequences, Protein synthesis, Protein, Amino acid sequences, DNA sources, Structured knowledge, Multi-agent system (MAS), Ontology, Personalized medicine.**

I. Introduction

PROTEINS, which consist of long or short sequences of amino acids, respectively called polypeptides and peptides, are assembled from amino acids according to the information contained in genes [1]. The protein synthesis process involves at least two steps: transcription of the DNA into a messenger RNA (mRNA) and translation of the mRNA into a protein. The first step is called transcription because the DNA is copied into the mRNA without changing the language (the language of nucleotides, i.e. adenine [A], cytosine [C], guanine [G], and thymine [T]). Transcription is performed with a polymerase RNA which binds to the DNA and synthesizes an RNA that is complementary to the DNA. For eukaryotes (humans, animals, fungi, and plants), there is an intermediate step, known as the maturation of the pre-messenger RNA (pre-mRNA), in which it undergoes excision of its introns (i.e. elimination of the parts of the gene that do not encode a polypeptide), and splicing of its exons (i.e. linking of the coding strand) to form the mRNA. Then the translation of the mRNA involves manufacturing proteins constituted by amino acid chains from the mRNA sequence [2]. Among the amino acids, twenty are most common, and so the translation of the mRNA transforms information coded with a four-character alphabet (as given above) to information coded with a twenty-character alphabet [3].

In sum, it is known that DNA is the origin of protein [4]. With the help of high-throughput sequencing technologies and efficient computational tools, DNA sequencing (i.e. determining the sequences of nucleotides in DNA) has led to a considerable volume of data that has successfully been stored and archived in DNA sources [5]. We utilize the DNA sources to dynamically construct a reference protein ontology in two steps. In the first step, a multi-agent model takes a DNA sequence as input and simulates the protein synthesis process, mapping the DNA to an amino acid sequence. In the second step, the generated amino acid sequence is automatically annotated, to structure it and then integrate it into the ontology.

The remainder of this paper is organized as follows. Section 2 provides an overview of research that is related to our approach. Section 3 presents our proposal for the dynamic construction of the protein ontology. Section 4 presents a software application and experimentation based on our methodology. Section 5 presents a discussion. Finally, Section 6 concludes the paper and suggests future research directions.

II. Related work

Advances in information and communication technologies coupled with increased knowledge about genes and proteins have opened new avenues for studying protein complexes [6]. There is a growing need to provide structured and integrated knowledge about various proteins for effective disease prevention mechanisms, personalized medicine and treatments, and other aspects of healthcare. Our two-step methodology for providing a reference protein ontology involves modeling the protein synthesis process and then structuring and integrating the proteins into the ontology. These two aspects of our methodology have been the subject of several studies.

A. Modeling the protein synthesis process

Protein synthesis is one of the more complex and dynamic biological processes. This complexity makes it difficult to explain, teach, demonstrate, and understand. Furthermore, laboratory experimentation regarding this biological process is time-consuming and requires days or weeks before the dynamic behavior or the expected results can be observed. For this reason, several computational approaches have been proposed to model this biological process. In [7] and [8], the authors developed a modeling methodology for the protein synthesis process using finite automata which represented

Manuscript received April 11, 2018; revised September 2, 2018. This work was supported by LIRE laboratory, University of Constantine2 - Abdelhamid Mehri, Constantine, Algeria.

Mohamed Hachem Kermani and Pr. Zizette Boufaida are with the department of software technologies and information systems, University of Constantine2 - Abdelhamid Mehri, Constantine, Algeria. e-mail: hachem.kermani@univ-constantine2.dz zizette.boufaida@univ-constantine2.dz

Dr. Zahia Guessoum is with LIP6 Laboratory, University of Paris6 - Pierre et Marie Curie, Paris, France. e-mail: zahia.guessoum@lip6.fr

the entire process from the initial state, DNA, to the final state, protein. In [9], the authors introduced a new modeling tool for protein synthesis that uses a data-flow diagram to represent the transformation of the input DNA to the output Protein. In [10], the same authors presented the same idea, this time using a deterministic Boolean network as a modeling tool. In [11], the author used a Petri net model to represent the protein synthesis process.

These computational approaches can assist in visualizing the protein synthesis process state by state and can explain how molecular events, reactions, and operations together produce proteins from DNA.

B. Structuring and integrating proteins

In order to integrate knowledge about proteins, it is critical to develop a structured data representation for proteomics knowledge, such as ontology. Several computational approaches have been proposed for structuring and integrating knowledge about proteins into ontologies. In [12], the authors described a protein ontology model for integrating protein databases and deduced a structured vocabulary to provide biologists and other scientists with a description of the sequences, structures, and functions of proteins. In [13], the authors proposed creating a protein ontology resource which provides a comprehensive understanding of the complex protein mechanisms. This ontology, which is available online, contains 91 concepts, 248 properties, and 99 instances. In [14], the authors proposed an ontology-based knowledge representation for protein data; the ontology is a unified terminology description integrating various protein database schemas.

We found more protein ontology integration projects [15], [16], and [6] that aim to provide a set of structured vocabularies for protein domains. However, while these projects provide a structured knowledge representation for proteomics, they are not dynamic but instead either transform static sources into static ontologies or develop static ontologies with a small number of concepts and properties.

TABLE I
COMAPRATIVE STATE OF SOME PROPOSED APPROACHES

Aspects Approches	Modeling of the protein synthesis process	Structuring and integrating proteins
[7]	X	
[15]		X
[8]	X	
[12]		X
[10]	X	
[13]		X
[9]	X	
[14]		X
[11]	X	
[16]		X
Our approach	X	X

III. OUR PROPOSAL

Proteins are assembled from amino acids according to the information contained in genes (protein synthesis). It is known that DNA, which consists of chains of nucleotides, is the origin of proteins. A way to determine the sequence of these nucleotides (DNA sequencing) was discovered in the second half of the 1970s, and since this discovery, research in molecular genetics and more specifically in DNA sequencing has seen a remarkable evolution—culminating in the Human Genome Project, which took thirteen years, three billion dollars, and a work force spanning many institutions to complete [17]. All of these DNA sequencing projects have led to a considerable mass of data, which has been stored in databases and knowledge bases (DNA sources). We propose utilizing what has been achieved in DNA sequencing to produce a reference protein ontology. We present a non-real-time agent-based system that dynamically constructs a reference protein ontology knowledge base using the existing DNA sources (Figure 1). The knowledge base can be used by experts for various aspects of healthcare.

We produce the ontology in two steps:

- Step 1: Sequence proteins automatically by modeling the protein synthesis process with an MAS.
- Step 2: Structure the sequenced proteins and dynamically integrate them into the ontology.

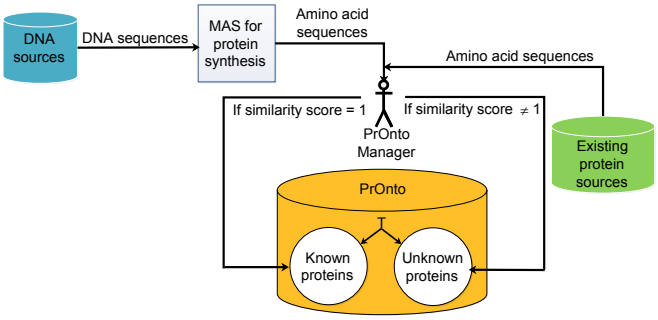

Fig. 1. The dynamic construction of the protein ontology

A. Multi-agent system modeling of protein synthesis

This step involves acquiring several DNA sequences at a time and modeling the protein synthesis process with an MAS that simulates this interactive and dynamic biological process, with several actors in each cell synthesizing (sequencing) the protein. After translating the DNA sequences into amino acid sequences, the MAS then recommends the most likely ones.

1) Description of the proposed MAS: We exploit the advantages of agents, including autonomy, information exchange, and cooperative negotiation [18], to simulate the biological process of protein synthesis in order to obtain the sequenced protein (the amino acid sequence).

- Information exchange: Acting in groups, our agents exchange information about each agents position in

the group, the type of each agent (A, C, G, or T), and which agents will allow starting or stopping the transcription and the translation.

- Autonomy: Our agents can change their state in the transcription stage and leave the process in the maturation stage.

- Cooperative negotiation: Our agents negotiate about which agents will disappear from the process in the maturation stage (non-coding nucleotides in the RNA), which agents will continue the path to the protein, and which amino acids will be formed by each triplet of agents in the translation stage. In these negotiations, the agents will face conflicts, which are solved by using references such the databases of known non-coding RNAs, to detect the non-coding agents and the genetic code, to know which amino acids can be formed by each triplet of agents, and so forth.

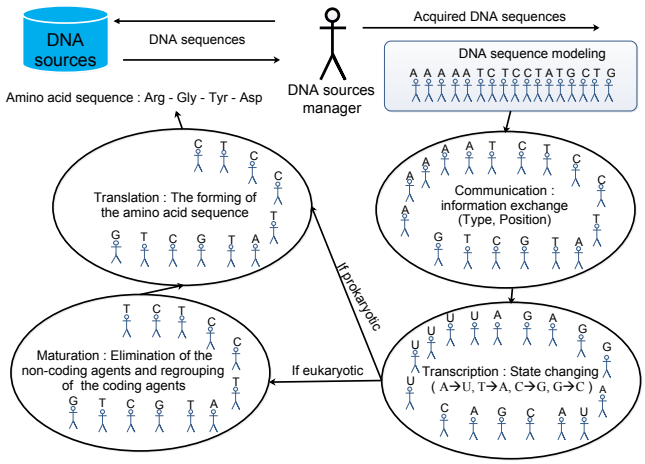

Fig. 2. General architecture of the MAS

Our proposed multi-agent system is composed of two essential components, the DNA sources manager agent and the nucleotide agents.

a) DNA sources manager agent: This agent connects to the DNA sources (e.g. GenBank [19], UniProtKB [20], and Gene Ontology [21]) to recuperate multiple DNA sequences of interest. The DNA sources manager agent first uploads a file containing DNA accession numbers and then creates a query for each accession number to be delivered to the DNA sources CGI program. In response to each query, the CGI program retrieves the corresponding record from the nucleotide database and sends it back to the DNA sources manager agent, which parses the record and extracts the DNA sequences, storing them in an internal data structure. The sequences will be modeled as groups of agents, with each agent modeling a nucleotide.

b) Nucleotide agents: The nucleotide agents represent four types of nucleotides (adenine, cytosine, guanine, and thymine). These agents play the same role that the nucleotides play in the cell; they collaborate by passing from one state to another until they reach the sequenced protein.

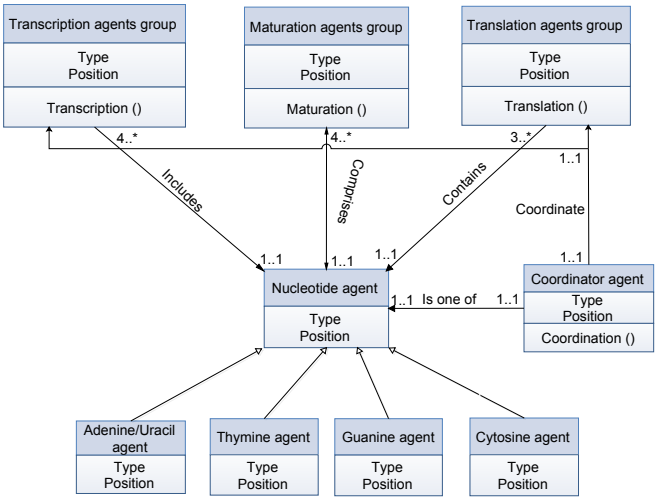

Fig. 3. Nucleotide agents

2) The behavior of the agents: Once the DNA sequences are recuperated by the DNA sources manager, each sequence is modeled as a group of nucleotide agents. A modeled DNA sequence can include up to thirty thousand agent instances, depending on the size of the DNA sequence.

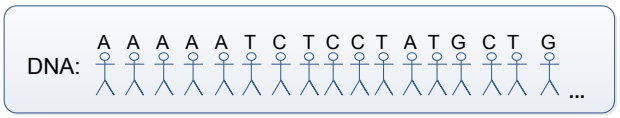

Fig. 4. Modeling of a DNA sequence as a group of agents

The group of agents forming the DNA sequence will go through several steps, as described below, until it reaches the sequenced protein.

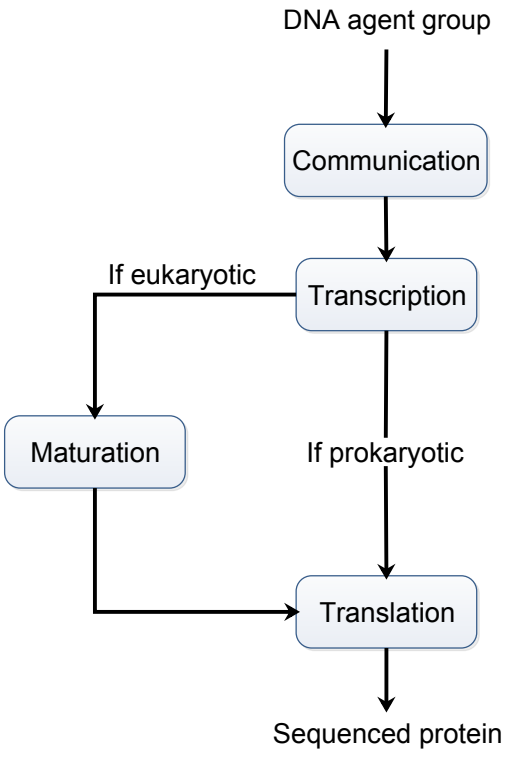

Fig. 5. The different steps of the protein synthesis

a) Communication: In this step, each agent communicates with the other agents, by sending a message including its position in the group and its type (A, C, G, or T) to the coordinator agent, which is one of the agents forming the DNA sequence. The coordinator agent initiates and coordinates the communication between the agents. This communication allows the agents forming the DNA sequence to know about each other, which facilitates the coorporative negotiation in the following steps.

b) Transcription: This step aims to transcribe the DNA sequence into a pre-mRNA sequence in which each nucleotide changes its state, with A to U, T to A, C to G, G to C. The group of agents forming the DNA sequence thus transcribes to a group forming a pre-mRNA sequence in which each agent has changed its state as indicated. Once the pre-mRNA sequence is formed, each agent communicates with the other agents by sending a message to the coordinator agent indicated its position in the group and its ***new*** state (U, A, C, or G). After this communication, the group of agents forming the pre-mRNA sequence moves to a third stage, maturation.

c) Maturation: This stage aims to mature the pre-mRNA sequence into a messenger RNA sequence. To this end, the agents forming the pre-mRNA sequence will negotiate regarding which agents should disappear and which should be maintained for the next step. If they cannot decide, they consult the non-coding RNA databases [22], [23], [24] in order to identify the non-coding agents that will disappear from the group. The retained agents form the mRNA sequence.

d) Translation: This step translates the mRNA sequence into an amino acid sequence. The agents forming the mRNA sequence cooperate and negotiate which amino acids will be formed by each triplet, based on the genetic code.

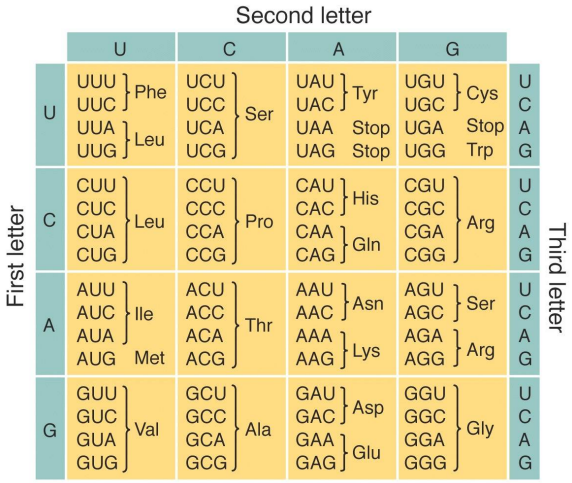

Fig. 6. The genetic code. [25]

The amino acids that are formed give us the sequenced protein (Figure 7).

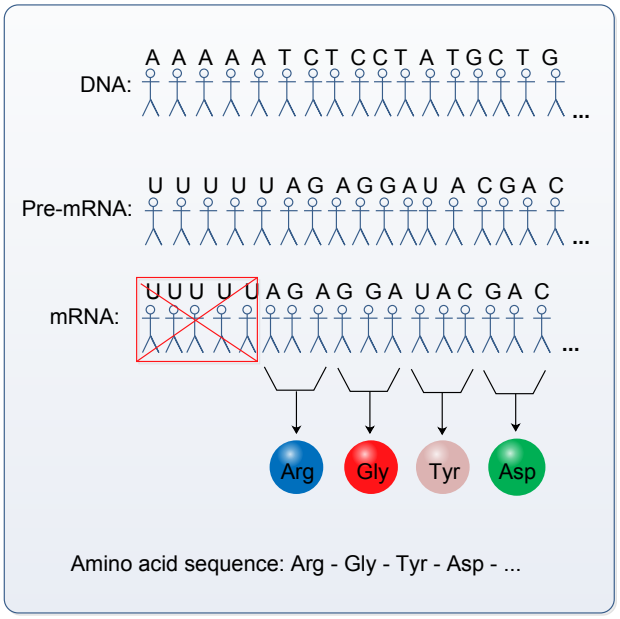

Fig. 7. Formation of the sequenced protein

The process just described is for the case in which the DNA sequence is a eukaryotic sequence. If the DNA sequence is prokaryotic, we will have the same process but without the maturation step, as shown in the state chart diagram in Figure 8:

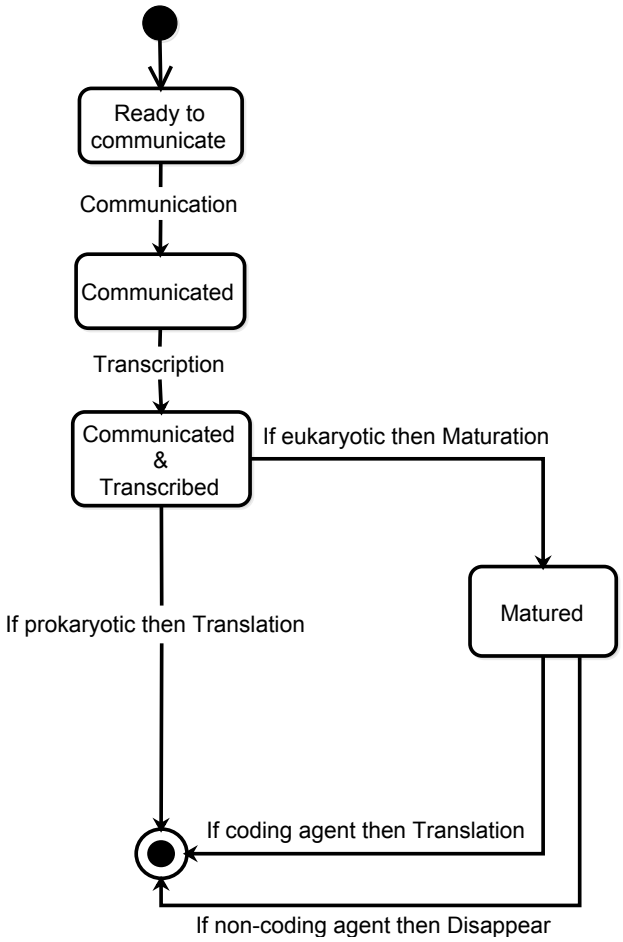

Fig. 8. The different states of the agent

3) The mechanisms of coordination: The final state of each agent is the result of the coordination between the agents; this coordination allows us to automatically produce the sequenced protein.

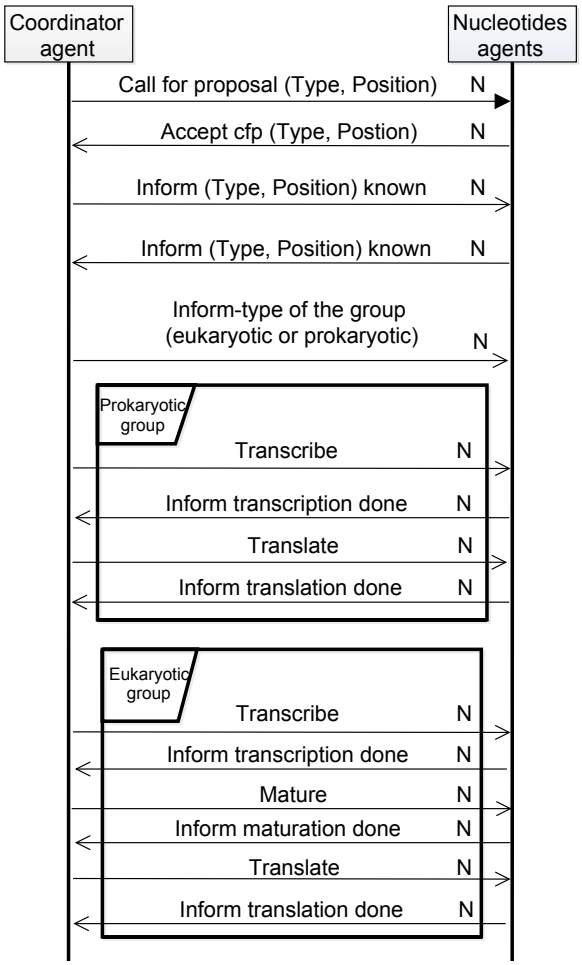

Fig. 9. The mechanisms of coordination between the agents

B. Structuring and integrating the sequenced proteins into an ontology

Proteins fall into several types (several classes), are organized as a hierarchy (super-proteins, sub-proteins), and stand in relations to other proteins (semantic relationships) [26]. For this reason, we propose using the amino acid sequences generated by the MAS to construct an ontology, PrOnto (Protein Ontology). We propose an agent, the PrOnto manager, which annotates, structures, and categorizes the generated proteins and then stores them in the ontology. PrOnto provides a reference protein knowledge base which can be exploited by scientists for a better understanding of life in order to address challenges in the medical, pharmaceutical, and pathological fields.

This step involves three substeps:

- Substep 1: Compare the generated amino acid sequences to other existing sequences in the available sources (e.g. RefSeq [27], NCBI's cdd [28], RCSB [29]) in order to annotate the generated proteins or identify them with new properties (evolved or abnormal)
- Substep 2: Formulate the generated proteins as ontological concepts.
- Substep 3: Integrate, structure, and categorize the proteins into the PrOnto ontology.

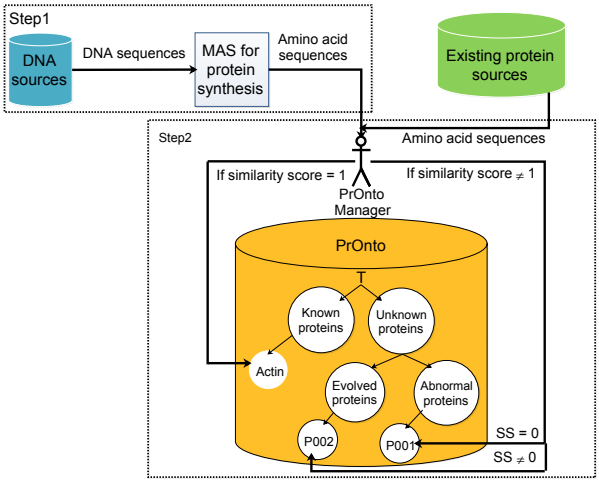

Fig. 10. Structuring and integrating the sequenced proteins

1) The process of structuring and integrating the sequenced proteins: The process of structuring and integrating that we present is a complete process in that, starting with the generated amino acid sequences, it automatically builds the PrOnto ontology.

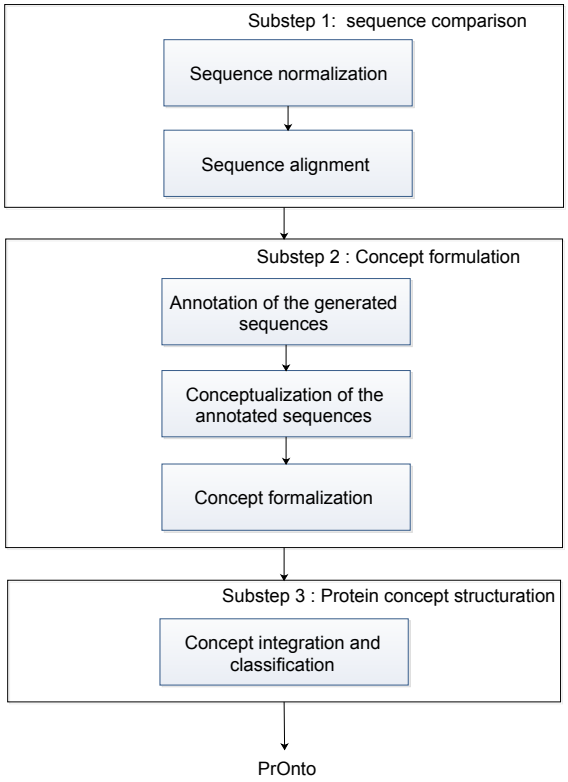

Fig. 11. The structuring and integration process

a) Substep 1: sequence comparison: In this substep, the PrOnto manager agent compares the generated amino

acid sequence with the amino acid sequences of already known proteins, which are stored in different existing protein sources.

- Sequence normalization: The amino acid sequences generated and recuperated from the existing sources are standardized to facilitate alignment:

1) The amino acids must be represented with three-letter codes (the first letter is capitalized). (Table II).

TABLE II
REPRESENTATION OF THE AMINO ACIDS

Symbol	Signification
Ala	Alanine
Cys	Cysteine
Phe	Phenylalanine
Gly	Glycine
His	Histidine
Ile	Isoleucine
Lys	Lysine
Leu	Leucine
Pro	Proline
Gln	Glutamine
Arg	Arginine
Ser	Srine

2) The spaces between the amino acids are represented by the symbol "-".

3) The amino acids of a protein sequence must be listed from left to right, with the first amino acid numbered as 1.

- Sequence alignment: The PrOnto manager agent is responsible for the alignment between the amino acid sequence generated by the MAS and the amino acid sequences recuperated from existing protein sources. We use pairwise alignment methods to compare the amino acid sequences; the pairwise alignment can be global or local. In global alignment, the entire sequences are compared, while in local alignment, one sequence is compared with a part of another sequence [30]. The alignment between the generated sequence, denoted below as X, and the recuperated sequence, denoted as Y, 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, for example:

X:	Lys	-	Glu	-	Thr	-	Lys
Y:	_	-	Glu	-	Thr	-	Thr
	0		1		1		0

The similarity score for the two sequences is calculated as follows:

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

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 examples:

Example 1:

X:	Lys	-	Glu	-	Thr	-	Lys
Y:	Lys	-	Glu	-	Thr	-	Lys
	1		1		1		1

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

Example 2:

X:	Lys	-	Glu	-	Thr	-	Lys
Y:	Thr	-	Glu	-	Thr	-	_
	0		1		1		0

$$Ss(X,Y) = \frac{\sum C,D}{NAA} = \frac{2}{4} = 0.5 \quad 50\,\% \tag{3}$$

Example 3:

X:	Lys	-	Glu	-	Thr
Y:	Thr	-	Lys	-	Glu
	0		0		0

$$Ss(X,Y) = \frac{\sum C,D}{NAA} = \frac{0}{3} = 0 \quad 0\,\% \tag{4}$$

The alignments result in one of three cases:

1) Ss = 1: The generated amino acid sequence perfectly matches a sequence of a known protein.

2) $0 < Ss < 1$: The generated amino acid sequence partially matches a known protein.

3) Ss = 0: The generated amino acid sequence does not match any known protein.

b) Substep 2: concept formulation: This substep formulates the generated amino acid sequences as ontological concepts.

- Annotation of the generated sequences: The MAS for protein synthesis generates an unknown sequenced protein, without a biological name, a protein type, relations to other proteins, or a position in the protein hierarchy. The PrOnto manager agent annotates the sequenced protein with the missing information, based on the result of the sequence comparison:

1) Ss = 1: The agent annotates the generated sequence with the information about the similar protein (its biological name, the type of protein,

its relations with other proteins, and its position in the protein hierarchy).

2) $0 < Ss < 1$: The agent annotates the generated sequence with the information about the partially similar protein and labels the generated sequence as an evolved protein.

3) Ss = 0: The agent labels the generated sequence as an abnormal protein.

- Conceptualization of the annotated sequences: In this phase, the PrOnto manager agent automatically formulates the annotated sequences as ontological concepts.

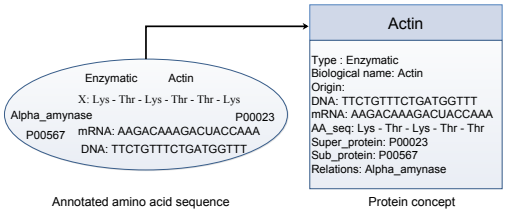

Fig. 12. Example of formulation of a protein concept

- Concept formalization: This represents the protein concepts with a formal and operational language, using the syntax of the SHIQ description logic [31].

1) Concept representation

$$\textbf{Actin} : (\forall\, Part\,of.Enzymatic) \cap (\geq 1\, In\,relation\,with.Alpha\,amynase \cap \leq 1\, In\,relation\,with.Alpha\,amynase) \cap (\exists\, Super\,protein.P00023) \cap (\exists\, Sub\,protein.P00567).$$

2) Relation representation

In relation with (Actin, Alpha amynase)
Super protein (P00023, Actin)
Sub protein (P00567, Actin)

c) Substep 3: protein concept structuring: In this substep, the PrOnto manager agent structures and classifies the protein concepts.

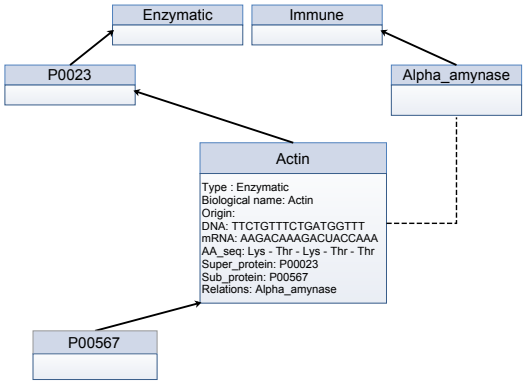

Fig. 13. Structuring the Actin concept

- Concept integration and classification: The integration is performed by applying the following classification algorithm:

Classification algorithm

1: **Input: X** // *The concept (i.e. the protein) to classify.*
2: **Begin**
3: Sps := **SPS(X)** ; // *Search the super concepts of X*
4: Spg := **SPG(X)** ; // *Search the sub concepts of X*
5: **If** ((Sps ∩ Spg) = ∅) **Then** // *X is a new protein to classify.*
6: **For each** (S ∈ Sps) **Do**
7: **Add** (X ⊆ S) ; // *Adding the subsumption link X ⊆ S.*
8: **End**
9: **For each** (G ∈ Spg) **Do**
10: **Add** (G ⊆ X) ; // *Adding the subsumption link G ⊆ X.*
11: **End**
12: **For each** (S ∈ Sps) **Do**
13: **For each** (G ∈ Spg) **Do**
14: Remove (G ⊆ X) ; // *Removing the subsumption link G ⊆ X.*
15: **End**
16: **End**
17: **Else** // *X is an existing protein.*
18: **Return** (nil);
19: **End if**
20: **End**

The proposed structuring and integration process is an iterative process. Whenever an amino acid sequence is generated by the MAS, it is used to dynamically build the PrOnto ontology, as described by the following structuring and integration algorithm:

Structuring and integration algorithm

1: **Input:** X // *The generated sequence.*
2: **Input:** Y // *The recuperated sequence.*
3: **Begin**
4: **Compare** (X, Y);
5: **If** (Similarity_score(X, Y) = 1) **Then**
6: **Recuperate** (C(Y)); // *knowledge about the corresponding protein C.*
7: **Annotate** (X(C)); // *Annotate X with the knowledge about C.*
8: **Integrate** (X); // *Integrate the generated protein X into the ontology.*
9: **Else** // *Similarity_score(X, Y) ≠ 1*
10: **If** (Similarity_score(X, Y) = 0) **Then**
11: **Recuperate** X; // *The generated protein.*
12: **Integrate** (X); // *Integrate X into the ontology as an abnormal.*
13: **Else** // 0 <Similarity_score(X, Y) <1
14: **Recuperate** (PC(Y)); // *the partial corresponding protein PC.*
15: **Annotate** (X(PC)); // *Annotate X with the knowledge about PC.*
16: **Integrate** (X); // *Integrate X into the ontology as evolved*
17: **End If**
18: **End If**
19: **End**

IV. Software Application and Experiment

In this section, we present a software application which uses our model and an experiment using the modeled system.

A. Software application

The scenario presented here is a case where the DNA sequence is a prokaryotic sequence. This software application is implemented in the Java programming language.

1) Translation interface: The interface in Figure 14 shows a fragment of an amino acid sequence generated by the MAS.

Fig. 14. A fragment of a generated amino acid sequence

2) Alignment interface: This interface shows the alignment between the generated amino acid sequence and an amino acid sequence recuperated from existing protein sources.

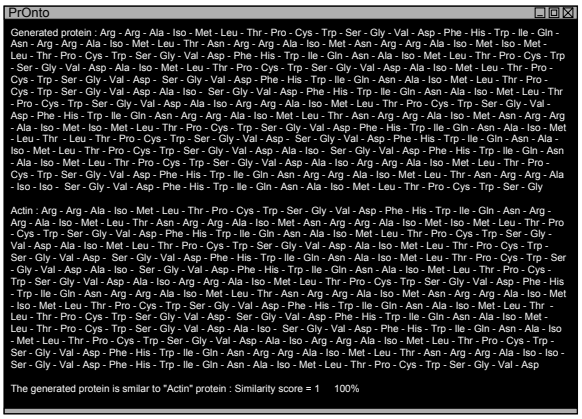

Fig. 15. Sequence alignment interface

The generated amino acid sequence will be annotated based on the results of alignment then formulated as an ontological concept, which will be structured and automatically integrated into the ontology.

3) The ontology: The sequenced proteins are dynamically structured and categorized as ontological concepts using the OWL language [32].

```
<?xml version="1.0"?><rdf:RDF
xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:daml="http://www.daml.org/2001/03/daml+oil#"    xmlns="http://www.owl-
ontologies.com/unnamed.owl#"    xmlns:dc="http://purl.org/dc/elements/1.1/"
xml:base="http://www.owl-ontologies.com/unnamed.owl"><owl:Ontology rdf:about="">
<owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
</owl:Ontology>  <owl:Class rdf:ID="Enzematic">    <rdfs:subClassOf>
<owl:Class rdf:ID="Proteins"/>    </rdfs:subClassOf>  </owl:Class>  <owl:Class
rdf:ID="Immune">    <rdfs:subClassOf rdf:resource="#Proteins"/>  </owl:Class>
<owl:Class rdf:ID="Transport">    <rdfs:subClassOf rdf:resource="#Proteins"/>
</owl:Class>  <owl:DatatypeProperty rdf:ID="RNA">    <rdfs:domain
rdf:resource="#Proteins"/>    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>  </owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="DNA">    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>    <rdfs:domain
rdf:resource="#Proteins"/>  </owl:DatatypeProperty>  <owl:DatatypeProperty
rdf:ID="Amino_acids">    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>    <rdfs:domain
rdf:resource="#Proteins"/>  </owl:DatatypeProperty>  <Enzematic
rdf:ID="Actin"><Amino_acids
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Methionine Leucine Threonine
Threonine Proline Arginine Alanine Alanine Isoleucine Arginine  </Amino_acids><RNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">AUGCUGACUACCCCGAGGGCUGCAAUUC
GU</RNA><DNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">TACGACTGATGGGGCTCCCGACGTTAAG
CA</DNA></Enzematic><Immune rdf:ID="Aldehyde_dehydrogenase"><Amino_acids
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Methionine Leucine Threonine
Threonine Proline Arginine Alanine Alanine Isoleucine  </Amino_acids><RNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">AUGCUGACUACCCCGAGGGCUGCAAUU<
/RNA><DNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">TACGACTGATGGGGCTCCCGACGTTAA<
/DNA></Immune><Transport rdf:ID="Alpha_amylase"><Amino_acids
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Leucine Threonine Threonine
Proline Arginine Alanine Alanine Isoleucine Arginine  </Amino_acids><RNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">CUGACUACCCCGAGGGCUGCAAUUCGU<
/RNA><DNA
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">GACTGATGGGGCTCCCGACGTTAAGCA<
/DNA></Transport></rdf:RDF>
```

Fig. 16. A fragment of the OWL file for the ontology

4) Interrogation and utilization of the ontology: Our ontology can provide scientists with a better understanding of life for addressing medical, pathological, and pharmaceutical issues. To allow scientists to exploit this ontology, we developed a platform supporting its interrogation. Scientists can consult the sequenced proteins, search for proteins, and compare proteins.

a) Consulting the sequenced proteins: If a scientist wishes to consult all the sequenced proteins, the platform recuperates all concepts and presents them to the scientist for use in the scientists research (Figure 17).

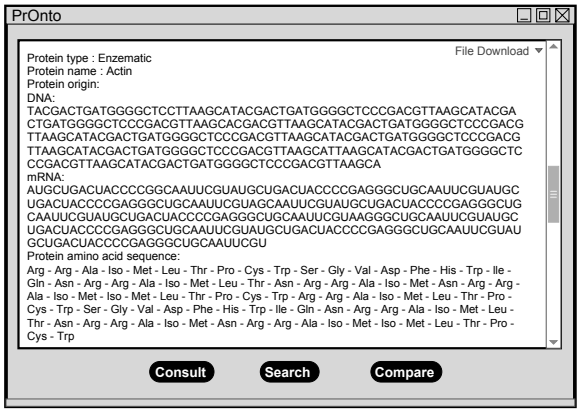

Fig. 17. Consulting the sequenced proteins

b) Searching proteins: The platform permits scientists to search proteins by their name or the amino acid sequence:

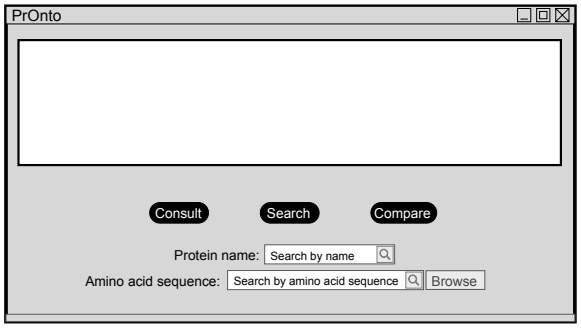

Fig. 18. Searching proteins

If the desired protein is found in the ontology, it will be presented to the scientist. In addition, the platform provides scientists with the opportunity to validate the reliability of the knowledge or correct erroneous information. This option helps us ensure the reliability of the knowledge stored in our ontology.

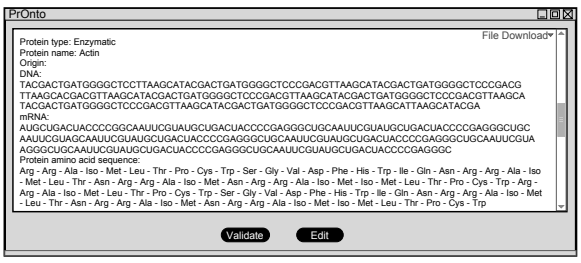

Fig. 19. Recuperation of a searched protein

If the desired protein is not found in the ontology, the platform proposes that the scientist add it by providing the essential knowledge about this protein (i.e. its name, the amino acid sequence, the mRNA sequence, the pre-mRNA sequence, and the DNA sequence). If the scientist can provide all of this information, the protein will be classified and integrated. If the scientist can only provide some knowledge about the protein, it will be classified as a protein with new properties. This option allows us to ensure the evolution of our ontology.

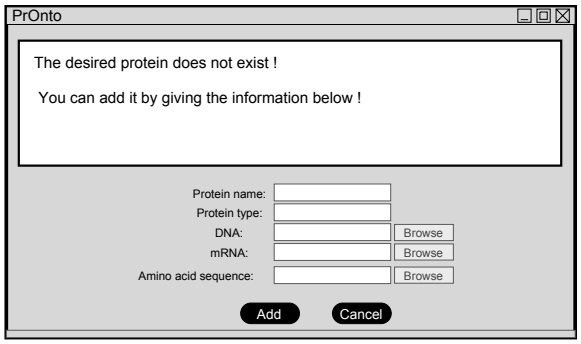

Fig. 20. Protein not found

c) Comparing proteins: The platform permits scientists to compare proteins in order to detect mutations and differences between a protein being studied and a reference protein existing in the ontology. This comparison can be:

- A comparison between a protein to be studied and one protein in the ontology. In this case we use pairwise alignment [33].

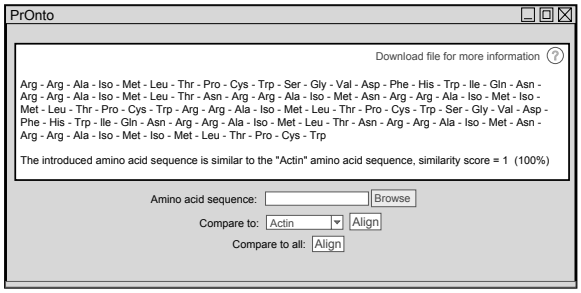

Fig. 21. Comparison between two amino acid sequences

- A comparison between a protein to be studied and all proteins in the ontology, in order to retrieve the maximum number of similar proteins. In this case we use multiple alignments [34].

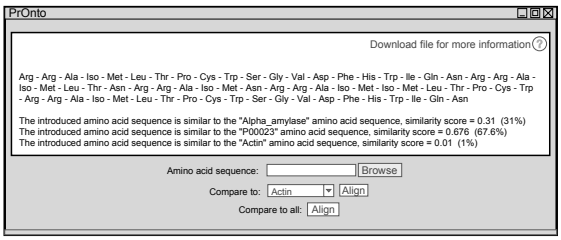

Fig. 22. Comparison between one amino acid sequence and all other sequences

B. Experiments

Our system utilizes existing DNA sources, which contain several genomes constituted by thousands of genes, which in turn are constituted by thousands of nucleotides. The genome size varies between 0.0065 MegaNucleotides and 1.4 MegaNucleotides for prokaryotes and between 1.45 MegaNucleotides and 1700 MegaNucleotides for eukaryotes. [35], [36] as shown in Table III.

TABLE III
GENOME SIZES OF SOME SPECIES

	Genomes	Size
Prokaryotes	Virus of influenza	0.0065
	Bactriophage λ	0.025
	Bactriophage T4	0.0825
	Mycoplasma pneumoniae	0.408
	Pelagibacter ubique	0.65
	Staphylococcus aureus	1.4
	Nanoarchaeum equitans	0.245
Eukaryotes	Encephalitozoon cuniculi	1.45
	Saccharomyces cerevisiae (yeast)	6
	Caenorhabditis elegans	50
	Drosophila melanogaster(fly)	59
	Arabidopsis thaliana (plant)	59.5
	Mus musculus (mouse)	1700
	Homo sapiens (Human)	1700

Based on these genome sizes, we used ten DNA sequences of different sizes and launched execution of our system ten times, in order to measure the running time.

TABLE IV
PROTEIN SEQUENCING TIME

Size of the DNA sequence (Nucleotides)	Minimum PST (Ms)	Average PST (Ms)	Maximum PST (Ms)
6500	32	54,8	85
25000	299	662,1	1178
82500	5213	13837,5	20528
245000	36607	91288,4	154962
408000	73801	206019,5	382423
600000	154128	457374,2	761754
650000	191192	601513,3	1135803
885000	356008	1129887,4	2080859
900000	357448	1182161,5	2121059
1400000	1079705	3216849,7	5708287

The results presented was obtained after launching the execution ten times on the 'OAR resource and task manager' [37], which is a batch scheduler for HPC clusters and other computing infrastructures.

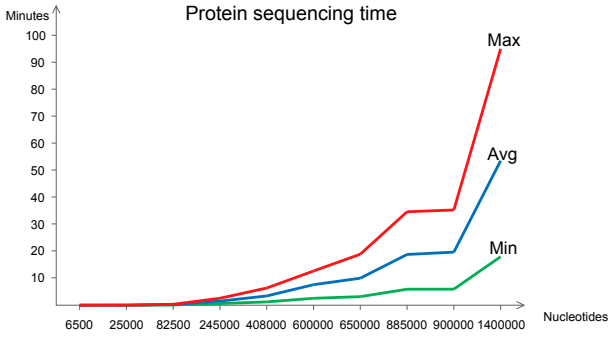

Fig. 23. Protein sequencing time

Analyzing this graph, we note that our system can sequence a single protein by exploiting an existing DNA sequence in a very short amount of time, varying between 0.0005 minutes and 95.14 minutes.

The protein sequencing time depends on the size of the DNA sequence and on the performance of the computer that is used.

Furthermore, our protein sequencing time is very short compared to the laboratory experimentation which is time-consuming and requires days or weeks before the dynamic behavior or the expected results can be observed.

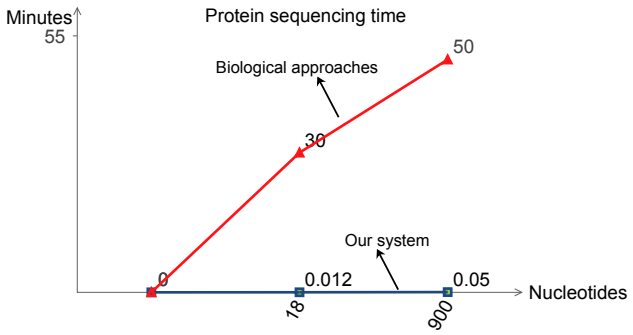

Fig. 24. Comparaison between the protein sequencing time

As shown on the graph, our system sequence a protein based on a DNA sequence of [18 - 900 nucleotides] in [0.012 - 0.05 minutes]. Unlike the biological approaches, which use the Edman method [38] that may be implemented either manually or through the use of automatic tools (i.e. protein sequencers [39]), the protein sequencing time of these biological approches varies between [30 - 50 minutes] for a DNA sequence of [18 - 900 nucleotides].

Once the amino acid sequence is generated it will be compared the to other existing sequences in the available sources (e.g. RefSeq [27], NCBI's cdd [28], RCSB [29]) in order to annotate the generated proteins or identify them with new properties (evolved or abnormal).

The experiments at this level shows that among the ten DNA sequences used, 3 generated amino acid sequences matche perfectly known proteins, 4 generated sequences matche partially known proteins and 3 others do not matche any known protein.

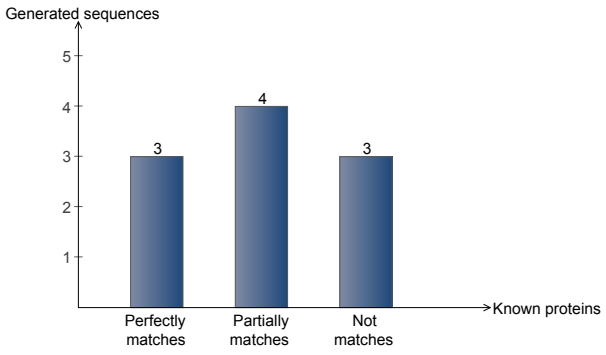

Fig. 25. The results of the sequence alignment

These results show that 70 % of the generated amino acid sequences were annotated with the missed information based on the known proteins.

The annotated amino acid sequence will be then formulated as ontological concepts, which will be structured and automatically integrated into the ontology "PrOnto".

"PrOnto" provides a reference protein knowledge base which can be exploited by scientists for a better understanding of life in order to address challenges in the medical, pharmaceutical, and pathological fields.

It includes concepts (type definitions), which are data descriptors for proteomics data and the relations among these concepts. The Key features of "PrOnto" are:

- a hierarchical classification of concepts (classes) from general to specific.
- a list of attributes for each class.
- a set of relations between classes to link concepts in ontology.

The Main Class of "PrOnto" is 'Proteins', there are 3 subclasses 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 differents types of proteins : 'Enzymatic, Immune, Transport, ...

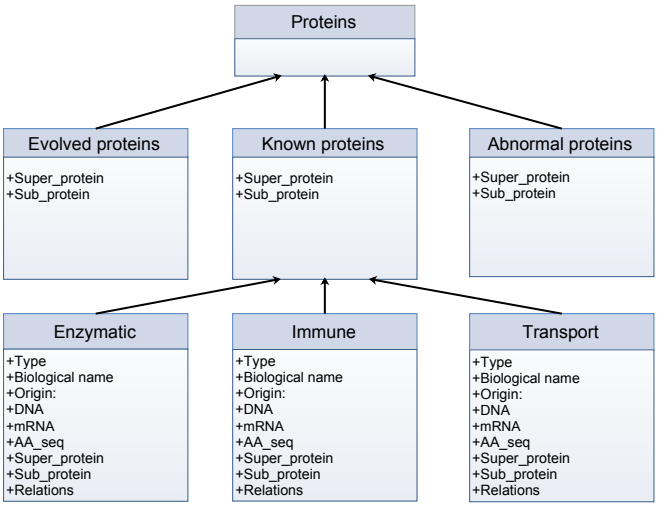

Fig. 26. The generic classes of the ontology

Each generated protein through our system will be structured and integrated into the ontology 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'.

The 'PrOnto' ontology currently contains 15 concepts or classes, 78 attributes or properties and 83 instances.

TABLE V
PROTEIN ONTOLOGIES COMPARAISON

	Concepts	Attributes	Instances
[13]	91	248	99
[12]	91	246	79
PrOnto	**15**	**78**	**83**
[6], [16]	>1000		
[20], [21]	>10000		

This number of proteins forming our ontology is still not enough yet compared to the existing protein ontologies (UniProt [20] and Gene Ontology [21]) which contains thousands of proteins.

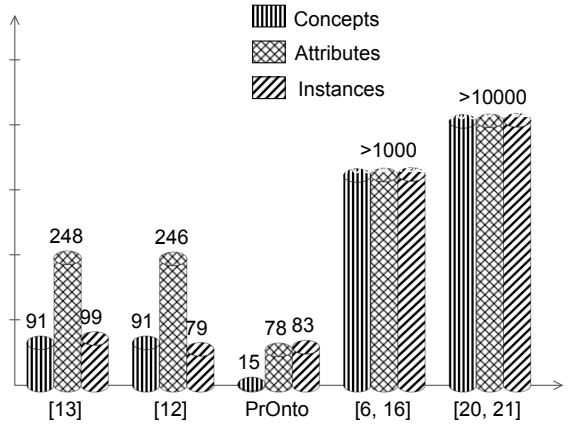

Fig. 27. Existing protein ontologies comparaison

PrOnto will continue to be enriched dynamically by the MAS as long as life exists and will contain a vast number of proteins.

V. DISCUSSION

The proposed two-step methodology for providing a reference protein ontology involves modeling the protein synthesis process and then structuring and integrating the proteins into the ontology.

The exisiting computational protein synthesis approches are theoretical and do not provide real simulations, do not produce real sequenced proteins, and do not address structuring and integrating the sequenced proteins.

Moreover, existing protein ontology integration projects, while it provide a structured knowledge representation for proteomics, they are not dynamic but instead 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 a two-step methodology that first translates DNA sequences into amino acid sequences via a multi-agent approach and then organizes the results into a dynamic protein hierarchy. We also presented a software application that implements this methodology. We modeled the protein synthesis process with a multi-agent system (MAS) in order to provide a real simulation of this dynamic, interactive biological process, that is, the MAS dynamically sequences proteins.

We provided a method for structuring, categorizing, and integrating the synthesized (sequenced) proteins into an ontology to provide a reference protein knowledge base.

Unlike some existing approaches to protein structuring and integration [12], [13], [14], [15], [16], our ontology will continue to be enriched dynamically by the MAS as long as life exists and will contain a vast number of proteins.

TABLE VI
COMPARATIVE STUDY

	Modeling of the protein synthesis process			Structuring and integrating proteins	
	Process simulation	Proteins producing	Rapidity	Static	Dynamic
[7]	**X**				
[15]				**X**	
[8]	**X**				
[12]				**X**	
[10]	**X**				
[13]				**X**	
[9]	**X**				
[14]				**X**	
[11]	**X**				
[16]				**X**	
PrOnto	**X**	**X**	**X**		**X**

VI. CONCLUSION

Medicine is gradually moving away from the traditional model of reactive sick care and towards personalized medicine, which involves designing medical treatments based on the patients individual characteristics (i.e. it tailors treatment for each patient). The main factors that play a critical role in personalizing medicine are the patients genetic and protein information. For this reason, research in recent years has focused on obtaining and understanding this information, which is contained in cells. DNA sequencing is one of the new tools that have been developed for obtaining and analyzing genetic information and making this information widely available. Although 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 second main factor that plays an essential role in personalized medicine, namely, the availability of the protein information. For this reason, we have proposed a two-step methodology for dynamic construction of a protein ontology which will be dynamically enriched as long as the DNA sources exist. This ontology provides a reference protein knowledge base that can be used for effective disease prevention mechanisms, personalized medicine and treatments, and other aspects of healthcare.

Furthermore, our software application and experimentation with the modeled system demonstrate, in addition to the feasibility of our approach, other benefits. These include:

- The opportunity to exploit the DNA sequencing results, either by using existing DNA sources or by connecting

to DNA sources and retrieving DNA sequences that can be used to sequence proteins under study.

- The use of agent and ontology techniques to model a biological system, by modeling a multi-agent system for protein sequencing and structuring and integrating sequenced proteins into an ontology.

- The automation of protein sequencing by modeling the process of protein synthesis, simulating the dynamic behavior of the protein synthesis, and sequencing the proteins in a fast, continuous way.

- The structuring, categorization, and dynamic integration of the sequenced proteins into an ontology which provides a reference protein knowledge base which can detect proteins with new properties and also provides an evolutionary knowledge base (ontology) which can be automatically and continuously enriched.

However, our proposal is not without limitations. We plan to address these in the future by optimizing the performance of the modeled system with a parallel solution, which will improve the protein sequencing time, and providing knowledge about other protein structures (secondary, tertiary, and quaternary), as our approach presently treats only the primary structure (amino acids).

ACKNOWLEDGMENT

The authors would like to thank Merrie Bergmann for doing a great job editing the manuscript.

REFERENCES

[1] P. C. Winter, G. I. Hickey, H. L. Fletcher *et al.*, *Instant notes in genetics*. BIOS Scientific Publishers Ltd, 1998.

[2] N. A. Campbell, L. G. Mitchell, J. B. Reece, and M. R. Taylor, *Biology: concepts & connections*. Benjamin/Cummings, 2000, no. QH308. 2 C35 1996.

[3] O. Gaci, "A study of the protein folding dynamic," *IAENG International Journal of Computer Science*, vol. 37, no. 2, pp. 185–194, 2010.

[4] G. Parker, *Creation: Facts of Life: How Real Science Reveals the Hand of God*. New Leaf Publishing Group, 2006.

[5] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-throughput sequencing technologies," *Molecular cell*, vol. 58, no. 4, pp. 586–597, 2015.

[6] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, H. J. Drabkin, P. DEustachio, A. V. Evsikov, H. Huang *et al.*, "The protein ontology: a structured representation of protein forms and complexes," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D539–D545, 2010.

[7] J. W. Yeol, I. Barjis, and Y. Ryu, "Modeling of system biology: from dna to protein by automata networks," in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*. IEEE, 2005, pp. 523–528.

[8] T.-Y. Lin and A. H. Shah, "Stochastic finite automata for the translation of dna to protein," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1060–1067.

[9] J. W. Yeol, W. Samarrai, I. Barjis, and Y. Ryu, "Data-flow diagram representation of biological process: Dna, rna, and protein," in *Bioengineering Conference, 2005. Proceedings of the IEEE 31st Annual Northeast*. IEEE, 2005, pp. 106–107.

[10] ——, "Deterministic boolean networks (dbn) modeling of molecular biology: Dna replication," in *Bioengineering Conference, 2005. Proceedings of the IEEE 31st Annual Northeast*. IEEE, 2005, pp. 120–122.

[11] I. Barjis, W. Samarrai, and I. Augustin, *Modeling of DNA Transcription and Gene Regulation Using Petri Nets*. Idea Group Inc, 2006, pp. 548–550.

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

[13] A. S. Sidhu, T. S. Dillon, and E. Chang, "Creating a protein ontology resource," in *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*. IEEE, 2005, pp. 220–221.

[14] A. S. Sidhu, T. S. Dillon, E. Chang, and B. Sidhu, "Ontology-based knowledge representation for protein data," in *Industrial Informatics, 2005. INDIN'05. 2005 3rd IEEE International Conference on*. IEEE, 2005, pp. 535–539.

[15] M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, and P. Veltri, "Algorithms and databases in bioinformatics: Towards a proteomic ontology," in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol. 1. IEEE, 2005, pp. 322–328.

[16] D. A. Natale, C. N. Arighi, W. C. Barker, J. Blake, T.-C. Chang, Z. Hu, H. Liu, B. Smith, and C. H. Wu, "Framework for a protein ontology," in *BMC bioinformatics*, vol. 8, no. 9. BioMed Central, 2007, p. S1.

[17] J. M. Heather and B. Chain, "The sequence of sequencers: the history of sequencing dna," *Genomics*, vol. 107, no. 1, pp. 1–8, 2016.

[18] J. Qin, Q. Ma, Y. Shi, and L. Wang, "Recent advances in consensus of multi-agent systems: A brief survey," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 4972–4983, 2017.

[19] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[20] U. Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2016.

[21] G. O. Consortium *et al.*, "Creating the gene ontology resource: design and implementation," *Genome research*, vol. 11, no. 8, pp. 1425–1433, 2001.

[22] Y. Zhao, H. Li, S. Fang, Y. Kang, W. Wu, Y. Hao, Z. Li, D. Bu, N. Sun, M. Q. Zhang *et al.*, "Noncode 2016: an informative and valuable data source of long non-coding rnas," *Nucleic acids research*, vol. 44, no. D1, pp. D203–D208, 2015.

[23] C. Xie, J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, W. Zhu, W. Wu, R. Chen, and Y. Zhao, "Noncodev4: exploring the world of long non-coding rna genes," *Nucleic acids research*, vol. 42, no. D1, pp. D98–D103, 2013.

[24] X. C. Quek, D. W. Thomson, J. L. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss, and M. E. Dinger, "lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas," *Nucleic acids research*, vol. 43, no. D1, pp. D168–D173, 2014.

[25] A. A. Turanov, A. V. Lobanov, D. E. Fomenko, H. G. Morrison, M. L. Sogin, L. A. Klobutcher, D. L. Hatfield, and V. N. Gladyshev, "Genetic code supports targeted insertion of two amino acids by one codon," *Science*, vol. 323, no. 5911, pp. 259–261, 2009.

[26] E. Galeota and M. Pelizzola, "Ontology-based annotations and semantic relations in large-scale (epi) genomics data," *Briefings in bioinformatics*, vol. 18, no. 3, pp. 403–412, 2017.

[27] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D501–D504, 2005.

[28] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz *et al.*, "Cdd: Ncbi's conserved domain database," *Nucleic acids research*, vol. 43, no. D1, pp. D222–D226, 2014.

[29] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng *et al.*, "The rcsb protein data bank: integrative view of protein, gene and 3d structural information," *Nucleic acids research*, p. gkw1000, 2016.

[30] K. Tamura and T. Ichimura, "Classifying of time series using local sequence alignment and its performance evaluation." *IAENG International Journal of Computer Science*, vol. 44, no. 4, pp. 462–470, 2017.

[31] F. Baader, I. Horrocks, C. Lutz, and U. Sattler, *Introduction to Description Logic*. Cambridge University Press, 2017.

[32] B. Motik, B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz, "Owl 2 web ontology language profiles (2012)," *W3C Recommendation*, 2016.

[33] R. Giegerich and D. Wheeler, "Pairwise sequence alignment," *Bio-Computing Hypertext Coursebook*, vol. 2, pp. 1–6, 1996.

[34] D. Higgins, "Multiple sequence alignment," in *Genetic Databases*. Elsevier, 1997, pp. 165–183.

[35] J. Dolezel, "Nuclear dna content and genome size of trout and human," *Cytometry*, vol. 51, pp. 127–128, 2003.

[36] T. R. Gregory, "Synergy between sequence and size in large-scale genomics," *Nature Reviews Genetics*, vol. 6, no. 9, p. 699, 2005.

[37] OAR resource and task manager, "Oar," 2016.

[38] P. Edman *et al.*, "Method for determination of the amino acid sequence in peptides," *Acta chem. scand*, vol. 4, no. 7, pp. 283–293, 1950.
[39] Shimadzu sequencers, "Shimadzu, inc," 2016.