I 3D3P: an Intelligent 3D Protein Prediction Platform

Mohamed Hachem Kermani[1] and Zizette Boufaida[2]
[1]*LIRE Laboratory, National Polytechnic School - Malek Bennabi, Constantine, Algeria*
[2]*LIRE Laboratory, University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria*
hachem.kermani@enp-constantine.dz, {hachem.kermani, zizette.boufaida}@univ-constantine2.dz

Keywords: Computational Biology, 3D Protein Structure, Protein Structure Prediction, Multiple Sequence Alignment, Machine Learning, Intelligent Platform.

Abstract: Proteins are macromolecules consisting of a chain of smaller molecules (i.e. amino acids) known as monomers. Three levels of protein structure are distinguished: primary, secondary and tertiary. Determining the three-dimensional (3D) structure of a protein when only a sequence of amino acids is given, is one of the most important and frequently studied issues in bioinformatics and computational biology. Therefore, in this paper, we propose an Intelligent 3D Protein Prediction Platform, which aims to completely determine the tertiary protein structure of a given protein primary structure (i.e. the amino acid sequence). The proposed intelligent platform is based on multiple sequence alignment and machine learning techniques to predict automatically 3D protein structures. We also present a software application and an experiment of the proposed platform, which will be used by experts for a better understanding of protein functions and activities in order to develop effective mechanisms for disease prevention, personalized medicine and treatments and other healthcare aspects.

1 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 macro-molecule consisting of a chain of amino acids, which are assembled through peptide bonds (i.e. an amino acid group of carboxylic acid with a neighboring amino acid group and thus form the primary structure [1]. 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, which 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. Some proteins, known as oligurics (i.e. made up of several polypeptide chains, each chain has a primary, secondary and tertiary structure), such as hemoglobin, reach a quaternary structure by adopting a symmetrical structure [2]. Many computational methodologies and algorithms have been proposed as a solution to the 3D Protein Structure Prediction problem, including comparative modeling methods and sequence alignment strategies, deterministic computational techniques, optimization techniques, data mining and machine learning approaches [3]. In our case we combine both sequence alignment and machine learning techniques to automatically predict 3D protein structure of a given amino acid sequence. Furthermore, the proposed intelligent platform provides experts with all information needed for a deeper understanding of proteins functions and activities. 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 3D Protein Prediction Platform. 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.

2 RELATED WORK

X-ray crystallography, which is a time-consuming and relatively expensive method, has determined most of the protein structures available in the Protein Data Bank [4]. Hence computational methods have been

developed to compute and predict protein structures based on their sequences of amino acids.

2.1 X-ray Crystallography Method

X-ray crystallography is a technique for determining the structure of molecules in three dimensions, including complex biological macromolecules such as proteins and nucleic acids. It is a powerful method at atomic resolution in elucidating the three-dimensional structure of a molecule. The X-ray crystallography technique uses diffraction patterns that are generated by irradiating a crystalline sample of the molecule of interest with X-rays, making diffraction quality crystals mandatory for this process [5].

Although this method provides a powerful tool in elucidating the three-dimensional structure, the major drawback is time. Thousands of experiments on crystallization can be performed daily in a single laboratory, each experiment is observed over time, with the normal time span being weeks to months [6]. It was for this reason that computational methods were developed to reduce time and costs.

2.2 Computational Methods

Proteins fold into one or more specific conformations to exercise their biological functions [7]. The Determination of a protein's structure can be achieved through computational techniques that automatically predict protein structures based on their amino acid sequences. The three common bioinformatics methods used to predict the protein structure are: comparative modeling, fold recognition and ab initio prediction.

2.2.1 Comparative Modeling

Also known as homology modeling, it is a technique which uses known information from one or more homologous partners to predict the structure of an unknown protein. Comparative modeling usually involves three steps: a) identifying template structures for modeling the query protein, b) aligning the template with the query sequence, and c) modeling the query structure [8]. This family of methods enables greater number of potential templates to be produced and better templates to be identified [9]. To predict the three-dimensional protein, both the template and the query can be submitted to a comparative modeling program once the better template has been identified.

2.2.2 Fold Recognition

We model the proteins in fold recognition by threading which have the same fold as the proteins of known structures. Protein threading is used for protein that is not stored in the Protein Data Bank (PDB) with its homologous protein structures [10]. Many algorithms for determining the correct threading of a sequence into a structure have been proposed. They employ some form of dynamic programming. The problem of identifying the best alignment for the complete 3D threading is very difficult (it is an NP-hard issue for some threading models) [11]. Researchers have therefore proposed many methods of optimization, such as Conditional random fields, simulated annealing, branch and bound and linear programming, in order to achieve heuristic solutions.

2.2.3 Ab-initio Prediction

The ab-initio method is a technique that attempts to predict protein structures based solely on information about sequences and without using templates. Ab-initio modelling is often referred to as de-novo modeling [12]. The fundamental procedure followed by the protein structure prediction ab-initio method begins with the primary amino acid sequence, which is searched for the various conformations which lead to the prediction of native folds [13]. After recognition and prediction of the folds, the model assessment is carried out to verify the quality of the predicted structure.

Numerous methods of predicting protein structures, including X-ray crystallography, and computational methods are currently being used [14]. Each method has advantages and disadvantages, but they all have the same goal of building a consistent 3D protein model that can be useful for a detailed understanding of protein and enzyme function.

3 I 3D3P

Proteins, consisting of long or short amino acid sequences, respectively called polypeptides and peptides, are assembled from amino acids based on the information contained in the genes[15]. Protein synthesis is the process in which cells produce proteins by determining a protein's various structures: primary, secondary, and tertiary. The proposed Intelligent 3D Protein Prediction Platform (I 3D3P) aims to determine the three-dimensional protein structure from a given amino acid sequence, based on a multi-

ple sequence alignment technique or a machine learning method.

First the platform compare the amino acid sequence introduced by the user with all amino acid sequences of already known proteins existing on the available protein sources, then based on the results of this comparison, the platform predict the 3D protein structure as illustrated in Figure 1.

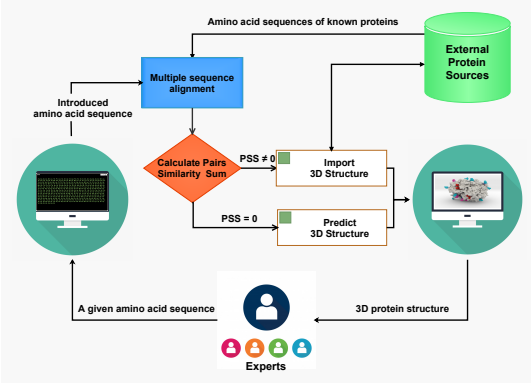

Figure 1: The 3D protein prediction process.

The proposed platform enables to predict the 3D protein structure from a given amino acid sequence based on two techniques, 3D structures importing from the available protein sources or the automatic prediction of 3D structures.

3.1 Importing 3D Structures

This technique consists of comparing the given amino acid sequence with all sequences of the already known proteins available in different protein sources in order to import the corresponding 3D protein structure. Therefore, we developed a multiple sequence alignment technique to compare the given sequence with all known proteins available in the existing protein sources.

The proposed sequence alignment technique consists to recuperate all protein sequences from the available sources and align them with the given sequence. This alignment creates a similarity score matrix between the given amino acid 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 examples:

Example 1:

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

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

Example 2:

X:	Lys	-	Glu	-	Thr	-	Lys
Seq2:	Thr	-	Glu	-	Thr	-	_
	0		1		1		0

$$Ss(X, Seq2) = \frac{\sum C, D}{NAA} = \frac{2}{4} = 0.5 \quad 50\,\% \quad (3)$$

Example 3:

X:	Lys	-	Glu	-	Thr
Seq3:	Thr	-	Lys	-	Glu
	0		0		0

$$Ss(X, Seq3) = \frac{\sum C, D}{NAA} = \frac{0}{3} = 0 \quad 0\,\% \quad (4)$$

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) = 0.5	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:

$$PSS(X, Seq_n) = \sum Ss(X, Seq_n) \quad (5)$$

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

1) PSS AX, Seq$_n$) $\neq$0 : The given sequence matches perfectly a sequence of a known protein and/or matches partially some known proteins. In this case, we will have two different situations based on the similarity score of each pair:
   - Ss (X, Seq$_n$) = 1 : The given sequence matches perfectly a sequence of a known protein. In this case, the platform will import the 3D protein structures in order to provide it to experts.
   - $0 < Ss < 1$: The given sequence partially matches some known proteins. The platform will import all the partially similar 3D protein structures and display them for experts.
2) PSS (X, Seq$_n$) = 0 : The given sequence does not match any known protein. In that case, the Intelligent Platform will automatically predict the 3D protein structure based on a machine learning technique.

3.2 The Automatic Prediction of 3D Structures

The inference of a protein's three-dimensional structure from its amino acid sequence remains an extremely difficult and unsolved task. A prediction for proteins consists of assigning regions of the amino acid sequence to be probable alpha helices, beta strands. Different methods for predicting 3D structures have been developed. One of the first algorithms was the Chou-Fasman method [16, 17], which relies mainly on the probability parameters determined from the relative frequencies of the appearance of each amino acid in each type of secondary structure [18].

In addition, overtime computational prediction methods were developed which are based on techniques of sequence alignment and methods of machine / deep learning. In our case, we propose a machine learning technique in order to automatically predict 3D protein structures. This technique is a work in progress and will be presented in our future work.

4 SOFTWARE APPLICATION AND EXPERIMENT

In this section, we present a software application and an experiment of the proposed platform. To illustrate our proposed I3D3P we developed the software application with Java programming language (see Figure 2).

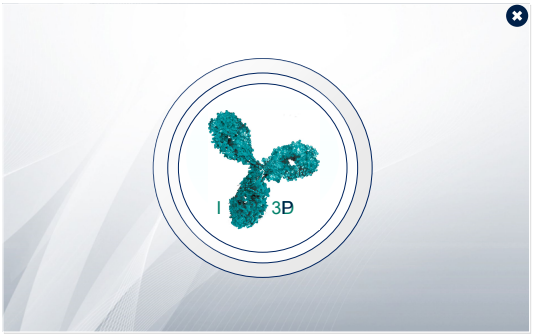

Figure 2: The Intelligent 3D Protein Prediction Platform.

I3D3P is an intelligent platform enabling experts to introduce a given amino acid sequence in order to get its 3D structure. Users can browse the file of an amino acid sequence as presented in Figure 3.

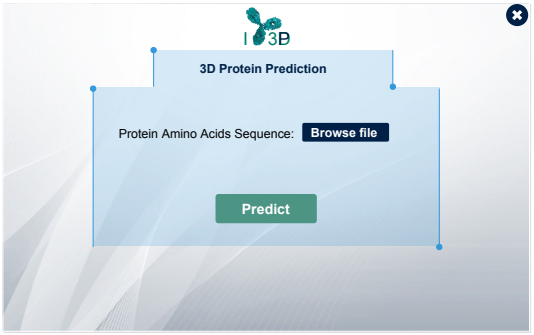

Figure 3: Browsing a given amino acid sequence.

According to our proposed intelligent platform, the prediction of the 3D protein structure depends on the results of the multiple sequence alignment. Meanwhile, the software application illustrates the case where the given amino acid sequence perfectly matches an amino acid sequence of a known protein.

Figure 4 shows that the given amino acid sequence is similar to the "Glycated Hemoglobin" amino acid sequence. In this case the 3D file will be imported from the external protein source, then displayed to allow the 3D protein visualizing, which will enable experts getting all information needed for a better understanding of protein functions and activities (Figure 5).

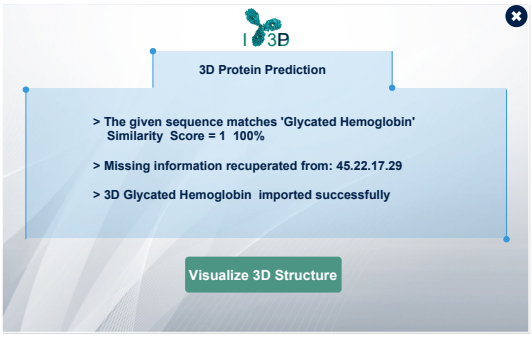

Figure 4: Importing 3D protein structure.

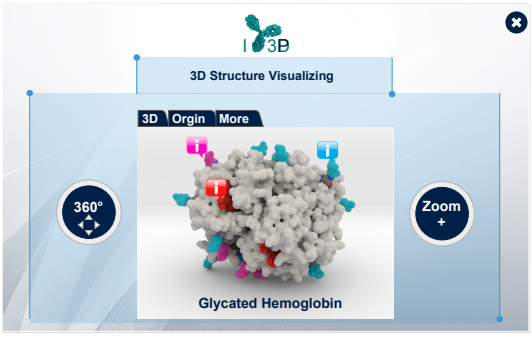

Figure 5: 3D protein visualizing.

5 DISCUSSION

Each method used to determine protein structures, including X-ray crystallography, and computational methods has advantages and disadvantages. The major drawback of X-ray crystallography is that method only aims to determine the 3D structure. In addition, laboratory experimentation with this method is time-consuming and requires days or weeks before the dynamic behavior or the expected results can be observed [19]. Instead, computational methods were developed with the aim of reducing time-consuming and getting faster results[20]. Therefore, we propose an Intelligent 3D Protein Prediction Platform, which is a computational tool that aims to determine 3D protein structures in a faster way. The proposed intelligent platform is based on computational methods by combining sequence alignment and machine learning techniques. The I3D3P will enable information on protein structures to be obtained altogether, which will allow a better understanding of protein functions and activities.

6 CONCLUSION

Protein structure prediction is the inference of a protein's three-dimensional structure from its amino acid sequence, – i.e., the prediction of its folding and tertiary structure from its primary structure. Determining a protein's 3D structure gives us a lot of information about how it operates, which we can use to control or modify its function, predict what molecules attach to it and understand diverse biological interactions. Therefore, protein structure prediction has been an important open research problem for more than 50 years. As a result, several approaches and techniques, including X-ray crystallography and computational methods, have been proposed as 3D protein structure prediction solutions. Our proposed platform is based on computational methods that combine multiple sequence alignment and machine learning techniques to predict 3D protein structures from a given amino acid sequence. I3D3P enabled the identification of 3D protein structures, allowing all protein information to be available at the three different structures. These information will be used to gain a better understanding of the proteins' functions and activities in order to develop effective mechanisms for disease prevention, personalised medicine and treatments, and other aspects of healthcare. However, our proposal has some drawbacks. We intend to address these issues in the future by presenting a machine learning method for 3D structure prediction, as our platform currently relies solely on multiple sequence alignment to predict 3D protein structures.

ACKNOWLEDGEMENTS

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] D. T. Haynie and B. Xue, "Superdomains in the pro-tein structure hierarchy: The case of ptp-c2," Protein Science, vol. 24, no. 5, pp. 874–882, 2015.

[2] I. Kumari, P. Sandhu, M. Ahmed, and Y. Akhter, "Molecular dynamics simulations, challenges and op-portunities: a biologist's prospective," Current Pro-tein and Peptide Science, vol. 18, no. 11, pp. 1163–1179, 2017.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A.Žıdek, A. Potapenko et al., "Highly accurate protein structure prediction with alphafold," Nature, vol. 596, no. 7873, pp. 583–589, 2021.

[4] D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman, and S. K. Burley, "The rcsb pdb "molecule of

the month": inspiring a molecular view of biology," PLoS biology, vol. 13, no. 5, 2015.

[5] J. A. Brito and M. Archer, "X-ray crystallography," in Practical Approaches to Biological Inorganic Chemistry. Elsevier, 2013, pp. 217-255.

[6] M. H. Kermani, Z. Guessoum, and Z. Boufaida, "A two-step methodology for dynamic construction of a protein ontology." IAENG International Journal of Computer Science, vol. 46, no. 1, 2019.

[7] Y. Zhang, "I-tasser server for protein 3d structure prediction," BMC bioinformatics, vol. 9, no. 1, pp. 1-8, 2008.

[8] S. D. Lam, S. Das, I. Sillitoe, and C. Orengo, "An overview of comparative modelling and resources dedicated to large-scale modelling of genome se-quences," Acta Crystallographica Section D: Struc-tural Biology, vol. 73, no. 8, pp. 628-640, 2017.

[9] B. John and A. Sali, "Comparative protein structure modeling by iterative alignment, model building and model assessment," Nucleic acids research, vol. 31, no. 14, pp. 3982-3992, 2003.

[10] L. A. Kelley, "Fold recognition," in From Protein Structure to Function with Bioinformatics. Springer, 2009, pp. 27-55.

[11] T. Jo, J. Hou, J. Eickholt, and J. Cheng, "Improving protein fold recognition by deep learning networks," Scientific reports, vol. 5, p. 17573, 2015.

[12] D. Xu, L. Jaroszewski, Z. Li, and A. Godzik, "Aida: ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction," Bioinformatics, vol. 31, no. 13, pp. 2098-2105, 2015.

[13] R. Townshend, R. Bedi, P. Suriana, and R. Dror, "End-to-end learning on 3d protein structure for interface prediction," Advances in Neural Information Process-ing Systems, vol. 32, pp. 15 642-15 651, 2019.

[14] M. H. Kermani and Z. Boufaida, "A2pf: An automatic protein production framework," in International Conference on Intelligent Systems Design and Appli-cations. Springer, 2020, pp. 80-91.

[15] M. H. Kermani, Z. Boufaida, S. Benredjem, and A. N. Saker, "An mvc-inspired approach for an intelligent annotation of a protein ontology : Iapronto," Inter-national Journal of Computer Information Systems and Industrial Management Applications, vol. 13, no. 2021, pp. 308-318, 2021.

[16] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," Biochemistry, vol. 13, no. 2, pp. 222–245, 1974.

[17] P. Y. Chou and G. D. Fasman, "Empirical predictions of protein conforma-tion," Annual review of biochemistry, vol. 47, no. 1, pp. 251-276, 1978.

[18] P. Y. Chou, "Prediction of the secondary structure of proteins from their amino acid sequence," Advances in enzymology and related areas of molecular biology, vol. 47, pp. 45-148, 1978.

[19] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis, "K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks," Journal of chemical information and mod-eling, vol. 58, no. 2, pp. 287-296, 2018.

[20] M. H. Kermani and Z. Boufaida, "A state of art on bi-ological systems modeling," in 2016 IEEE Intl Con-ference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Sympo-sium on Distributed Computing and Applications for Business Engineering (DCABES). IEEE, 2016, pp. 712-715.