RESEARCH PAPER

Protein Structure Prediction: From Homologue Identification to Accurate 3d Modelling Using an Ultra-Fast Search Engine

Rohit Mishra[1]*,Manoj Kumar Pal[1]

[1]*Department, Computer Science and Engineering, United University, Rawatpur, Jhalwa, Prayagra, 211012, Uttar Pradesh, India.*
ORCID ID: 0009-0005-3913-4959
[1]*Department, Computer Science and Engineering, United University, Rawatpur, Jhalwa, Prayagra, 211012, Uttar Pradesh, India..*

ABSTRACT
The Protein Data Bank (PDB) currently has three times as many proteins as it did 10 years ago. That is a significant increase. There are other datasets, such as SCOP and CATH, that exhibit a trend that is comparable to this one. Despite this, the categorisation of protein structures is still a procedure that is laborious, costly, and time-consuming. Because the amount of data is increasing at an exponential pace, the techniques of manually classifying proteins are becoming outdated. The use of precise computational and machine learning methods is a viable option that has the potential to provide a substantial boost in order to handle the growing amount of data. The introduction of lightning-fast search engines, on the other hand, has significantly sped up the process of protein structure prediction. This has made it possible to find homology, design models, and optimise them more quickly. These lightning-fast search engines represent a significant advancement in the field of modern computational biology. The declared objective of these search engines is to simplify and make the process of protein structure prediction more straightforward
Keywords: Protein Structure Prediction, 3d modelling, ultra-fast search engine, Homologue.
How to cite this article: Mishra R, Pal MK., Protein Structure Prediction: From Homologue Identification to Accurate 3d Modelling Using an Ultra-Fast Search Engine, In Vitro, and Ex Vivo Experimental Validation Supporting Adjunct Antifungal Therapy. Int J Drug Deliv Technol. 2026;16(3s): 533-543; DOI: 10.25258/ijddt.16.3s.68
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

In live cells, proteins are present everywhere. They are essential to the workings of species and their evolutionary machinery. To comprehend the molecular processes of life, it is crucial to study protein structures and activities. Millions of proteins- encoding sequences are being generated by high-throughput technologies, but none of them have been functionally characterised as of yet. The Protein Data Bank (PDB) now has three times as many proteins as it did ten years ago. You can see the similar pattern in alternative databases like SCOP and CATH. Nevertheless, protein structure categorization is still an involved, expensive, and time-consuming process. Data is growing at an exponential rate, making manual protein categorisation methods obsolete. An effective alternative that could provide a significant boost to address the expanding data load is accurate computational and machine learning techniques.

Proteins: Folding of lengthy amino acid chains in three dimensions: The intricate three-dimensional folding of lengthy amino acid chains is what gives proteins their structure. Because of its crucial role in protein function, this spatial structure is under constant evolutionary pressure to improve the inter-residue inter-actions that hold it together. Some of the computer methods that are now being used for protein categorisation make an attempt to imitate the biological processes that are responsible for determining the structure and function of proteins. The gold standard method involves searching a reference database of annotated proteins for similarities with an unknown protein. The protein under investigation is categorised according to how similar it is to the reference protein, either in terms of sequence or structure. Different approaches exist for protein sequence-based classification (e.g., Blast, ProtFun, SVM-Prot, etc.) and structure-based classification (e.g., Combinatorial Extension, Sheba, FatCat, Fragbag, etc.). These techniques are based on the premise that proteins with similar domain architectures are probably from the same family. The idea that related proteins may have descended from a common ancestor is the foundation of this categorisation scheme.

The following four procedures are typically used to forecast the three-dimensional structure of an unknown protein. Protein threading and homology modelling are two examples of backbone prediction methods that are used to anticipate a protein's backbone skeleton. In the second step, loops are incorporated to join the various portions of the backbone and create a whole backbone configuration. If the abinito folding method is employed to forecast a protein's structure, the first two phases are typically executed simultaneously. In order to create a full-atom model of the novel protein, the orientation of its side chains is next allocated. Lastly, the projected structure can be further refined using various molecular dynamic modelling approaches.

*Author for Correspondence: rohitmishra.academic@gmail.com

Protein threading problem Apologies, but the protein side-chain prediction problem and the protein threading problem are both NP-hard. To address these two issues, numerous heuristic methods and applications have been created. Here, we offer a tree-decomposition based method for breaking down protein structures into their component parts. We were able to resolve the protein threading problem and

the protein side-chain prediction challenge using this tree-decomposition method. We transform both issues into a geometric neighbourhood graph labelling problem so that they can be solved simultaneously. Based on the results of the experiments, it seems that the tree-decomposition method has the potential to successfully handle both of the found concerns. When it comes to providing a solution to a threading problem, the tree-decomposition method often outperforms the linear programming approach. In addition to this, we determine the circumstances in which the tree-decomposition technique is superior than the linear programming method. Through combining the two approaches, we will be able to reach greater levels of efficiency.

Key Challenges in Protein Structure Prediction:

Accuracy: Predicting the structure of very big or complicated proteins is still challenging, even though algorithms like AlphaFold have substantially increased accuracy.

Disordered Regions: There are several proteins with areas that are fundamentally disordered and cannot take on a stable shape. To simulate them would be a challenge.

Conformational Flexibility: Because proteins may take on a variety of shapes and sizes in response to their surroundings, it is challenging to foretell all of the potential conformations.

Computational Cost: Significant computing resources and time are required by certain approaches, particularly ab initio.

BACKGROUND

Ultra-Fast Search Engines in Protein Structure Prediction A significant amount of time and computer resources are required, according to the conventional wisdom, in order to deduce the three-dimensional (3D) structure of a protein based on its amino acid sequence. Through the facilitation of faster homology detection, model construction, and optimisation, the advent of ultra-fast search engines has considerably sped the process of protein structure prediction.

With the declared aim of simplifying and speeding up the process of protein structure prediction, these very fast search engines represent a giant leap forward in modern computational biology.

Proteome language models (pLMs) supersede genomic data. To leverage the massive quantities of data contained in the quickly growing yet unlabelled protein sequence databases, protein language models (pLMs) have emerged as a result of recent advances in natural language processing (NLP).

These models rely only on sequential patterns in the input. For instance, one may get an embedding—a representation of the protein sequence—by feeding the network a protein sequence and, in the last 16 network layers, constructing activation-based vectors. This allows one to interpret the information gained by such pLMs. Several domains of protein prediction have shown the potential benefits of this capacity, including function prediction, structure and disorder prediction, and any downstream prediction task 18 that needs numerical protein representations. Protein activity is more correlated with embedding space distance than sequence similarity when classifying proteins into families. The use of pLMs for the prediction of protein three- dimensional structures is an emerging field. Predicting future structures using pLM embeddings, as opposed to MSA evolutionary data, is simpler and takes less time. A priori calculated More than 200 million proteins have AlphaFold2 predictions avail- able, however precision is sacrificed for speed. Is it rational to compromise accuracy for efficiency?

Applications of Protein Structure Prediction:

Drug Design: The three-dimensional conformation of a target protein enables researchers to design minute molecules, or pharmaceuticals, that adhere to the protein with exact precision.

Disease Understanding: Proteins that are either misfolded or altered

are associated with a wide range of illnesses. Therapeutic treatments may be found by understanding these structures.

Biotechnology: Understanding the structure of proteins is essential for

engineering them with specific functionality or stability for industrial uses.

Objectives of the Study

To learn about the most important problems in predicting protein structures.

Researching the Use of Very Fast Search Engines for Predicting Protein Structures.

MATERIALS AND METHODS

3.1 Methods for Forecasting Protein Structure

An interdisciplinary research topic that has attracted the interest of researchers from a variety of domains, including computer science, biochemistry, physics, medicine, and mathematics, is the prediction of the structure of proteins. Biochemists and physicists investigate the fundamentals of protein folding; mathematicians, particularly statisticians, use a target sequence and a probability distribution of potential protein conformations to determine the structure that is most likely to be present; and computer scientists conceptualise the challenge of protein structure prediction as an optimisation problem, with the goal of finding the best possible solution (Schonherr et al.,

2018). Research on biological subjects has attracted an increasing number of academics from a wide range of fields during the second part of the twentieth century.

The computational process of protein structure prediction is complex and requires a wide variety of tools and techniques. Protein structure prediction approaches may be broadly classified into two groups: template-based modelling, also known as

homology modelling or comparative modelling (CM), and de novo modelling, often known as ab initio modelling (Yan et al., 2020).

Proteins with identical sequences are thought to fold into identical three- dimensional structures. The more specific terms Template Based Model and comparative model indicate that a template protein is utilised, but the template is not necessarily related to the target in terms of history or function (Yan et al., 2020). In contrast to this, HM begins the construction of the protein's three-dimensional structure from structural information of evolutionarily-related sequence(s). Discovering homologs (templates), aligning targets to those templates, creating structures, refining and validating them are all steps in TBM. In order to improve accuracy, hybrid techniques include elements from both groups (Rives et al., 2019).

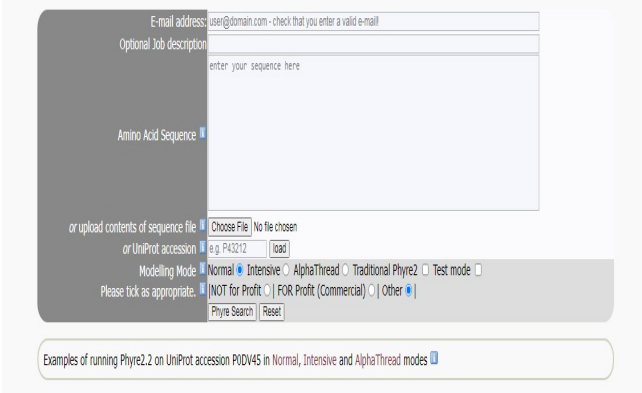
3.2 Template-Based Modelling (Homology Modelling)

Homology modelling is a theory that proposes proteins with sequences that are simi- lar are likely to have structures and functions that are comparable to one another. To begin, we need to choose a template for the structure of a protein that is highly congruent with the sequence of the protein that we are trying to create. During the succeeding step of the alignment process, the target protein and the template sequence are com- pared in order to locate areas of similarity between the two. In accordance with Houkes and Zwart (2019), the coordinates of the template protein are transferred in order to generate a three-dimensional model of the target protein. It is widely acknowledged that Phyre2, MODELLER, and SWISS-MODEL are the most well-known programs for homology modelling. However, the availability of a known structure that closely corresponds with a significant portion of the projected sequence is necessary for this protein structure model to function well. These situations are specified by comparison modelling that takes place during the building of an all-encompassing atomic model and the selection of a fold from a group of candidate templates. Following the elucidation of the experimental structure of a protein belonging to a family, it is possible to infer the structures of other proteins belonging to that family based on the degree to which they correspond to the structure that has been determined. Due to the fact that a tiny alteration in the sequence of a protein often leads in a minor adjustment to the three-dimensional structure of the protein, this is definitely a possibility. Further- more, it is important to highlight that the three-dimensional structures of proteins demonstrate a higher degree of conservation compared to the amino acid sequences that correspond to them within the same family (Kim and Chung, 2020).

3.3 Protein-protein interaction analysis

If there are protein-protein interactions involving the target protein, the homology model may help identify the interface. Analysis of surface residues and prediction of possible binding partners allow the model to guide experimental studies of protein complexes and probe protein-protein interactions (Jin et al., 2020). Functional annotations may be improved with the use of the homology model by comparing its structure to those of known proteins with comparable functions. It is possible to get valuable functional insights into the target protein by applying known functions to the modelling template. Use the homology model to check hypotheses about the target protein's function or mechanism of action. As an example, the model may be used to investigate the theoretical role and structural context of a residue that is believed to be crucial for a biological process (Jha and Saha, 2020). Virtual screening and structure- based drug development are two applications of the homology model. A few examples of its practical applications include predicting ligand-binding interactions, guiding the creation of novel compounds, and assessing the medicinal potential of ligands.

When the sequences of the target and template proteins are thirty to forty percent similar, template-based



modelling may be beneficial. The homology model gets less precise as the sequence similarity decreases, which makes it more difficult for proteins with low identity to match structural structures that are already known. Hybrid approaches and de novo modelling are two additional methodologies that may be used for the purpose of protein structure prediction in certain circumstances (Lin et al., 2021).

3.4 Bioinformatics Resources for Predicting Protein Structure

Phyre1: A Tool for Identifying Protein Homology and Analogy.

Phyre2: Protein Homology/AnalogY Recognition Engine 2 (PHER2) is a popular web-based program for the prediction and analysis of protein structures. The Söding Group of Oxford University constructed and oversees it. The Phyre2 server provides enhanced functionality and greater accuracy, succeeding the original Phyre server (Nardo et al., 2018).

3.5 Features and functionalities

In cases when a reliable template is unavailable, it employs ab initio modelling to explore potential conformations and predict stable structures. In addition, Phyre2 can identify protein folds, infer structural and functional properties, annotate proteins functionally, predict their domains, and give tools for in-depth structural analysis (Orbán-N´emeth et al., 2018; Zhou Panaitiu, 2020).

Usage Through its web-based interface, Phyre2 is usually easy to use:

Start by going to the Phyre2 website and submitting an interesting protein sequence in FASTA format.

Prediction and analysis: Phyre2 will look at the sequence and see if it can find any homologous templates using ab initio or homology modelling. Best guesses are shown to the user.

Analysis and visualisation: The user may investigate and visualise the predicted protein structures with the help of the provided tools and features.

3.6 Dataset and Validation

The assessment of our approach is conducted using a benchmark set (20), including three distinct datasets sourced from Pfam (21), Gene3d (24), and SUPERFAMILY

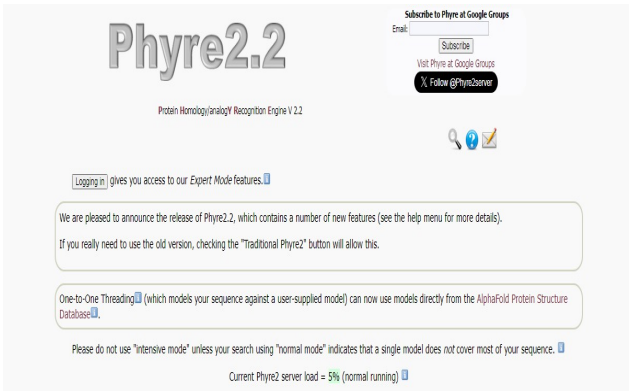


Fig. 1 PHYRE homepage (login page)

Fig. 2 Submission page

(22). In these databases, you could find proteins that contain more than one domain inside their structure. Proteins are considered to be homologs if they have the same specific sequence of domains. They are said to be nonhomologous if they do not have any domain similarities with one another. Two pieces of data constitute the benchmark. As a worldwide standard for homology, the first set, which we called “max50”, restricts domain-to-domain region lengths to 50 residues. The second group is comprised of all the others that do not have this restriction; we have called them “nomax50”. The Pfam dataset supports the” max50” benchmark with 5,245 homolog pairs. There are 5,047 pairings like this in the Gene3d dataset. The number of couples in the SUPERFAMILY is 5,656. Non-homology pairings are equally represented in each dataset as homolog pairs. The max 50 benchmark has 31,896 protein pairs in total, while the nomax50 benchmark contains 180,566 pairings. With the use of domain databases, the benchmark dataset was compiled from sixteen species. In place of structural databases that check for the inclusion of protein pairings with uncertain structures. They disagree on the homology of protein pairs with a margin of just 0.1 percent, despite the fact that the curation approach varies between databases. That all components of the benchmark dataset agree is crystal clear (20). To do this, we look at the benchmarking paper’s measures and compare them to the PROST accuracy (20).

RESULT

A comparative examination was conducted between the PROST tool and the pre- dominant alignment-based homolog identification methods. Gene3D, Pfam, and SUPERFAMILY are among the datasets used during the evaluation procedure. One technique to assess performance measures is by examining the AUC1000 score, which indicates the quality of the ranking. Previous benchmarking work (20) assigned scores of nine to CSBLAST, ten to PHIMMER, three to NCSI-BLAST, and four to FASTA. Our results were compared to those scores on subsequent benchmarking work. One further thing that we performed was examine the similarities and differences between the homology identification approach that was employed in the ESM1b research (18) and the PROST. The approach gives a representation of proteins that is of a constant size by making use of the average of the 34th layer of ESM1b embedding columns. Following that, a homology indication is executed by making use of the fundamental distance that exists between these models. This outcome is represented in the SI Appendix by the code ESM1DL34M. By a large margin, PROST is superior to other tools in its category. When comparing NCBI-BLAST with PROST on the SUPER- FAMILY dataset, for instance, the most notable increase is a 9.8 percent gain. This is an application of the comparison.

The results that were acquired from the benchmarking dataset using the most popular tools are shown in Table 1. According to the area under the receiver operating characteristic curve (ROC) for the first thousand false positives (AUC1000), the PROST approach is the most effective one. Given that PROST has an AUC1000 score that is 4.8% higher and an AUC that is 2% higher than the instrument that is closest to it, CS-BLAST, it is evident that PROST is the preferred option. Com- pared to the widely used NCBI-BLAST tool, PROST has a 3.1% higher AUC and a 6.3% higher AUC 1000 rating. This is a significant improvement. PROST surpasses CS-BLAST in benchmarking using the Gene3D database, with an AUC score that is 2.6% higher and an AUC 1000 score that is 4.7% higher. This can be seen in the comparison between the two. In conclusion, when it comes to the SUPERFAMILY dataset, PROST outperforms CS-BLAST by a margin of 2% AUC and PHMMER by a margin of 5.2% AUC1000 (PHMMER score is greater than CS-BLAST score). Table S2 in the SI appendix has the AUC and AUC 1000 values for all of the popular tools. You may get these numbers by looking at the table. A depiction of the ROC curves for the techniques that were assessed on the max 50 dataset can be seen in Figure 2. When evaluated on the Pram, Gene3D, and SUPERFAMILY datasets, PROST, which employs a reduced minimum optimised data format, outperformed various alignment- based homolog identification techniques that are routinely used. This demonstrates that PROST is a trustworthy instrument for homology detection in general. They consistently pave the way for the finding of considerable putative homologs, which is the case for sequences with lower levels of sequencing identity.

4.1 Data Analysis

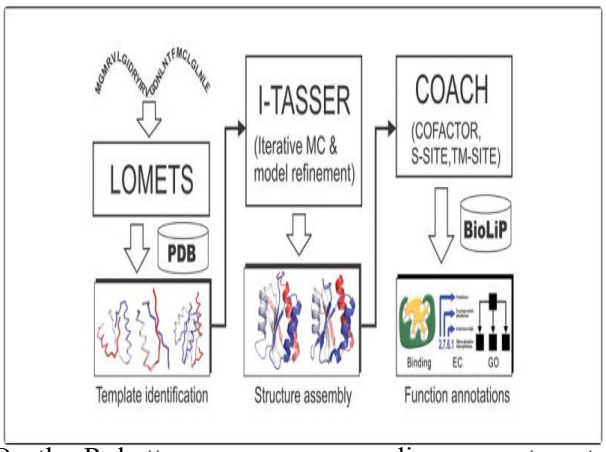
TASSER: Iterative Threading Assembly Refinement Workflow & principles of I-TASSER I-TASSER is a hierarchical system for structure-based function annotation and automated protein structure prediction. I-TASSER initially creates full-length atomic structural models from numerous threading alignments and iterative structural assembly simulations, followed by atomic level structure refinement, starting with the amino acid sequence of the target proteins (Xu et al., 2018a). Based on sequence and structure profile comparisons, after that, databases of known protein activities are used to deduce the protein's biological functions, such as its ligand-binding sites, enzyme commission number, and gene ontology keywords (Zhou et al., 2019). There is no cost to use either the web server or the standalone version of I- TASSER. Here we go over the steps to take when utilising the I-TASSER protocol to make predictions about a protein's structure and function, how to decipher those predictions, and some other ways to improve the accuracy of your I-TASSER models when working with targets that are very similar to each other or have several domains (Cheung & Yu, 2018).

Fig. 3 An example of an I-TASSER submission form

I-Tasser protocol for protein structure & function prediction. The specifics of the I-TASSER protocol have been explained in earlier articles. I-TASSER starts

with the amino acid sequence and uses LOMETS, a meta-threading technique made up of several separate threading programs, to search the PDB database for homologous structure templates (or super-secondary structural segments in the lack of such templates). Reassembling the continuously aligned fragment structures from the LOMETS templates and super-secondary structure segments creates the topology of the full-length models. Based on replica-exchange Monte Carlo simulations, ab initio folding is used to build the structures of the unaligned areas from scratch (Milanetti et al., 2018). SPICKER identifies the conformations with the lowest free energy by combining the trajectory data from Monte Carlo simulations. The structural models proceed through a second stage of structure reassembly, starting with the SPICKER clusters, according to Tong et al. (2018). Then, full-atomic simulations using FG-MD and ModRefinerused are used to refine the low free-energy conformations.

Fig. 4 Protocol of I-Tasser



On the Robetta server, you may discover automated tools that can examine and predict protein structures. Parsing sequences provided to the server into feasible regions for structure prediction generates structural models (Rodrigues et al., 2019). The next step is to use methods like comparative modelling or de novo structure pre- diction. If BLAST, PSI-BLAST, FFAS03, or 3D-Jury finds a strong match to a protein with a known structure, it will be used as a foundation for comparative modelling. To create structural predictions when a match is not found, the de novo Rosetta fragment insertion method is used. For RosettaNMR de novo structure determination, experimental nuclear magnetic resonance (NMR) constraints data may be supplied with a query sequence. One such capability that is currently available is the ability to predict the impact of mutations on protein-protein interactions by computational interface alanine scanning. The service will soon include Rosetta protein design and protein-protein docking (Mao et al., 2019). The Baker Lab at the University of Washington

created the Robetta protein structure prediction pipeline, which is now renowned across the globe. For the purpose of protein structure prediction using amino acid sequences, Robetta employs a set of cutting-edge algorithms and approaches. It has been extensively used by structural biologists to generate reliable three-dimensional protein models (Zhao et al., 2021a). Developing a model to anticipate Robetta protein structures

Modelling with templates (homology modelling)

Robetta begins by searching the plethora of experimentally validated protein structures in the Protein Data Bank (PDB) for proteins that share sequences with the target protein.

When it finds appropriate templates, Robetta creates an initial model by matching the target sequence to the structure of the templates.

Afterwards, the pipeline optimises the structure's form and refines the model using molecular dynamics simulations.

Folding from scratch

If a good template isn't available, Robetta will utilise ab initio folding methods to create a new structural model of the protein.

Different conformations of the protein's backbone and side chains are examine using ab initio methods to determine the lowest-energy conformation.

Prediction of side chain positions

Robetta uses energy-based and machine learning techniques to forecast where the side chains will be located in the protein structure.

Getting a realistic protein structure requires accurate side chain prediction.

Table 1 Performance comparison across Pfam, Gene3D,

and SUPERFAMILY datasets

Method	Pfam		Gene3D		SUPER FAMILY	
	AUC	AUC 1000	AUC	AUC 1000	AUC	AUC 1000
Prost	0.03	97.2	98.9	95.7	98.5	95.5
CSBLAST	97.0	0.78	92.4	98.3	91.0	98.1
PHMMER	98.4	0.56	92.3	98.2	90.4	95.9
NCBI-BLAST	95.9	0.46	90.9	94.4	87.9	93.7
ESM1bL34M	98.0	89.8	91.9	80.1	92.0	81.1
FASTA	94.8	88.8	93.2	85.2	91.9	83.4

Table 1 displays the findings obtained from the benchmarking dataset by the most common tools. The area under the receiver operating characteristic (ROC) curve for the first thousand false positives is a measure of performance, and PROST is the best method according to this statistic. The AUC and AUC1000 scores of PROST are 2% and 4.8% higher, respectively, than those of the nearest instrument, CS-BLAST. When compared to the popular NCBI-BLAST tool, PROST has a 3.1% better AUC and a 6.3% higher AUC1000 rating. When tested with the Gene3D database, PROST outperforms CS-BLAST by 2.6% in terms of AUC and 4.7% in terms of AUC1000. For the SUPERFAMILY dataset, PROST achieves a 2% higher

AUC than CS-BLAST and a 5.2% higher AUC than PHMMER.

Applications & strengths of Robetta: Predicting protein structures is an important and often used activity in biochemistry and molecular biology. You need to know how proteins are structured in three dimensions in order to deduce their activities, roles, and interactions in all sorts of biological processes. According to Song et al. (2018), Robetta is a well-known platform in the field of protein structure prediction known for its remarkable accuracy and versatility. Robetta has several potential applications in the fields of biochemistry and biology. For functional annotation, it is absolutely necessary to begin with. Robetta fills the void created by the absence of experimentally proven structures in several newly sequenced proteins by providing accurate predictions of their structures. By outlining a framework for annotating the activities of these proteins, these theories provide light on their involvement in both healthy and sick cellular activity. The drug development method makes heavy use of Robetta. Based on the platform's accurate predictions of protein structures, researchers may identify potential therapeutic targets and develop drugs that can interact with these targets (Wang et al., 2019). Knowing the properties and form of a protein's active site is crucial for developing successful therapeutics, which is why this is particularly helpful in the area of structure-based drug design. The exceptional accuracy, longevity, and versatility of Robetta make it stand out. It employs a number of state-of-the-art computational methods, including ab initio modelling and homology modelling, to improve its prediction capabilities. The most current advances in bioinformatics and structural biology are also included into Robetta via ongoing updates. This commitment to development ensures that Robetta users have access to cutting- edge prediction methodologies, making it a trustworthy and current resource in the continually expanding area of protein structure prediction. Scientists with varying levels of computational expertise may use it with ease thanks to its intuitive interface, which in turn encourages broad adoption and collaboration across disciplines (Zheng et al., 2019).

Consequently, Robetta is an excellent choice for this purpose since it is both accurate and dependable. Predicting the structures of proteins is an important and versatile method in biology. Its uses in structural biology, drug development, and functional annotation, along with its accuracy, adaptability, and constant progress make it an essential tool for researchers trying to solve the puzzles of protein structure and function (Larke et al., 2021). Robetta's merits lay in its ability to integrate numerous approaches, including homology modeling and ab initio folding, to build accurate protein structures. It is flexible and adaptable to a variety of protein targets because to the mix of template-based and de novo prediction techniques. The predictions' accuracy can, however, differ based on variables like sequence similarity, template accessibility, and structural complexity, much like with all other methods for structure prediction (Wojtowicz et al., 2020).

4.2 Modelling of Protein-Protein Interaction

4.2.1 Importance of Protein-Protein Interactions in Biological Processes

Proteins, the unsung heroes of biological systems, sustain life by carrying out a myriad of critical functions. Having said that, they almost never work alone. The most common and significant of these intricate molecular dances is the protein-protein inter- action (PPI), in which proteins play a central role. Interactions like this are important to many biological processes, including gene regulation, cell signalling, and enzyme function (Rigoldi et al., 2018). It is crucial to grasp the significance of PPIs in order to fathom the complexities of life's inner workings. The signalling and communication processes inside cells rely on PPIs. Hormones and neurotransmitters are examples of external signals that cells must respond to in order to adapt to a changing environment. Protein phosphoinositide intermediates (PPIs) allow proteins to transmit and receive signals that convey vital information inside the cell. For example, according to Bergenholm et al. (2018), a cascade of PPIs impacts behaviour, mood, and other physiological processes once neurotransmitters like serotonin bind to neural receptors. The chemical processes necessary for life are catalysed by enzymes. The correct functioning of many enzymes depends on the exact interactions between the many proteins sub- units that make them up. By coordinating the timing and direction of these subunits' combinations, PPIs facilitate reactions that would not occur in an optimal energy state without them. Coordination of protein activities in enzyme complexes is crucial for many biological processes, including cellular respiration, metabolic pathways, DNA replication, and others (Seath et al., 2021). By forming interactions with other proteins, PPIs are able to alter their function. This regulation meticulously regulates the quantity and timing of gene expression, which impacts cell fate, differentiation, and reactions to external stimuli (Li et al., 2018). By coordinating the timing and direction of these subunits' combinations, PPIs facilitate reactions that would not occur in an optimal energy state without them. According to Ali et al. (2019), enzyme complexes play a crucial role in a variety of biological activities, including metabolic pathways, cellular respiration, and DNA replication. PPIs play a crucial role in gene regulation. One example is the regulation of gene expression by proteins known as transcription factors, which attach to specific areas of DNA. By forming interactions with other proteins, PPIs are able to alter their function. This regulation influences cell fate, differentiation, and reactivity to environmental cues by precisely controlling the time and quantity of gene expression. Understanding the importance of PPIs to biological processes, researchers are focussing increasingly on them as potential therapeutic targets (Liu et al., 2020). By targeting certain PPIs with tiny chemicals or biologics, critical disease-related pathways may be boosted or inhibited. More precise and efficient therapies may be possible with this approach, which holds particular promise for complex illnesses where single-target approaches may not be enough. The basic building blocks of biological complexity are interactions between proteins.

Their relevance is multifaceted, spanning from fundamental biological processes to intricate disease mechanisms and the creation of new medications. While our understanding of PPIs continues to grow, our capacity to decipher complex biological systems and

develop novel approaches to modern health problems also grows (Gouthami et al., 2022). Therefore, PPI research is leading the way in modern biology, providing answers to some of the biggest questions about the most intricate dance in the universe.

4.2.2 Applications & significance

There are several uses for MD simulations in the research of PPIs. They provide atomistic insights into the function of water molecules in PPIs as well as binding routes and processes. Additionally, MD simulations reveal transitory intermediate states, clarify the structural dynamics of protein complexes, and provide information for mutagenesis investigations. Additionally, by estimating binding affinities, suggesting possible drug candidates, and supporting the logical design of new therapies, they serve a crucial role in drug development (Fang et al., 2019). Our capacity to describe and comprehend atomic-level protein-protein interactions has been completely trans- formed by molecular dynamics simulations. The multiple methods used in MD, such as improved sampling, QM/MM simulations, atomistic to coarse-grained simulations, and free energy calculations, provide a broad range of tools for researching PPIs at various timelines and degrees of detail. MD simulations continue to be at the fore- front of computational biology as methods and computer resources improve, allowing researchers to better understand the complexities of protein-protein interactions and their function in the molecular machinery of life (Sejdiu & Tieleman, 2021).

4.3.3 Machine Learning Techniques for Predictive Modelling in Bioinformatics

Machine learning algorithms have developed into powerful predictive modelling tools in the field of bioinformatics, offering a means of extracting insightful information from complex biological data. Bioinformatics, which involves the computer analysis of biological data, has significantly improved with the use of machine learning methods. These techniques are particularly useful for tasks such as understanding gene regulation, identifying disease markers, and predicting protein structures (Kumari et al., 2015). During this session, we will look at a variety of machine learning methods that are used to bioinformatics predictive modelling. One of the primary applications of machine learning in bioinformatics is sequence analysis. Machine learning algorithms may be trained on large datasets of DNA, RNA, or protein sequences to find functional elements, motifs, and patterns. For example, sequence-based classifiers can predict if a certain DNA sequence encodes a particular protein or includes a particular regulatory region. Because these models depend on information extracted from the sequences, such as the makeup of amino acids or nucleotides, they may be used to annotate unidentified sequences (Sartor et al., 2019). For structural bioinformatics, machine learning techniques are very beneficial, particularly for protein structure pre- diction. With methods like AlphaFold, it is possible to predict protein 3D structures with a remarkable degree of precision. These models combine deep learning with cur- rent protein structural information to improve our understanding of protein function, connections, and medication discovery. Machine learning is also essential to functional annotation. By analysing gene expression data, machine learning algorithms may categorise genes based on their roles in biological processes or disease pathways. This leads to the discovery of possible treatment targets or disease biomarkers. Additionally, to provide a thorough knowledge of gene function, machine learning may integrate data from other sources, including as transcriptomic, proteomic, and genomic data (Li et al., 2020).

Predicting protein-protein interactions is another area in which machine learning shines. Understanding these relationships is necessary to comprehend signalling pathways and biological activities. Machine learning algorithms may be trained on experimental data or characteristics derived from protein sequences and structures to predict likely protein interactions (Mucaki et al., 2019). Understanding this is essential for breaking down complex biological networks. In drug development, machine learning speeds up the process of identifying possible therapeutic candidates. Virtual screening uses machine learning techniques to prioritise compounds with high binding affinities. The computer screening of chemicals against medicinal targets is known as virtual screening. Machine learning also has a lot to offer customised medicine. Through the analysis of particular patient data, such as genomes and medical records, predictive models may help in treatment strategy customisation. The use of machine learning in bioinformatics is still problematic, nevertheless. Concerns about data quality, model interpretability, and ethical quandaries are a few of the issues that need careful consideration. As biological data continues to grow in volume and complexity, researchers are now working to build robust machine learning techniques that can handle large amounts of data.

CONCLUSION

With applications ranging from protein structure prediction to complex interaction modelling and function linkage elucidation, bioinformatics tools have emerged as essential resources in the area of protein research. These tools, developed by combining the introduction of innovations such as AlphaFold, which show the remarkable precision of protein structure prediction, has revolutionised structural biology. These discoveries not only provide insight into the intricate structure of proteins, but they also provide new opportunities for the creation of novel treatments and the hunt for novel medications. In the context of protein-protein interaction modelling, bioinformatics tools have proven crucial in unravelling the complex web of biological processes. By providing crucial insights into the laws regulating biological systems, these models aid in the discovery of

novel treatment targets and the elucidation of disease pro- cesses. The prediction and comprehension of protein function relationships have also been greatly enhanced by bioinformatics. Researchers may navigate the complex fields of proteomics and genomics with the help of bioinformatics tools, whether they are identifying the roles of newly sequenced proteins in biological settings or annotating them. the area of bioinformatics continues to grow and expand. With the advent of personalised medicine, the speed at which drugs are being developed, and the pursuit of a deeper comprehension of the molecular basis of life, these methods remain at the forefront of scientific progress. In the next years, bioinformatics will continue to help researchers and biochemists better understand proteins and their functions. It is a path marked by ingenuity, collaboration, and an unwavering commitment to deepening our knowledge of the intricate realm of proteins, one data point at a time.

REFERENCE

1. Ali, A. M., Atmaj, J., Van Oosterwijk, N., Groves, M. R., & D¨omling, A. (2019). Stapled peptides inhibitors: A new window for target drug discovery.
2. Computational and Structural Biotechnology Journal, 17, 263-281
3. [2] Bajpai, A. K., Davuluri, S., Tiwary, K., Narayanan, S., Oguru, S., Basavaraju, K., Dayalan, D., Thirumurugan, K., & Acharya, K. K. (2019). How helpful are the protein-protein interaction databases and which ones? bioRxiv.
4. [3] Ban, X., Lahiri, P., Dhoble, A. S., Li, D., Gu, Z., Li, C., Cheng, L., Hong, Y., Li, Z., & Kaustubh, B. (2019). Evolutionary stability of salt bridges hints its contribution to stability of proteins.
5. [4] Nielsen, J. (2018). Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. mSystems,
6. [5] Chen, M., Lin, X., Lu, W., Schafer, N. P., Onuchic, J. N., & Wolynes, P. G. (2018). Template-guided protein structure prediction and refinement using optimized folding landscape force fields. Journal of Chemical Theory and Computation, 14(11), 6102-6116.
7. [6] Gemovic, B., Sumonja, N., Davidovic, R., Perovic, V., & Veljkovic, N. (2019). Mapping of protein-protein interactions: Web-based resources for revealing inter- actomes. Current Medicinal Chemistry, 26(21), 3890-3891
8. [7] Scop: a structural classification of proteins database. Nucleic Acids Research.
9. [8] Pal, M.K., Lahiri, T., Tanwar, G., Kumar, R.: An improved protein structure evaluation using a semi-empirically derived structure property. BMC Structural Biology 18(1), Article 16 (2018). https://doi.org/10.1186/s12900-018-0097-0.
10. [9] Ihm, Y. (2004). A threading approach to protein structure prediction: studies on tnf-like molecules, rev proteins, and protein kinases..
11. [10] Jahan, M. S., Khan, H. U., Akbar, S., Farooq, M. U., Gul, S., and Amjad, A. (2021). Bidirectional language modeling: A systematic literature review. Scientific Programming
12. [11] Jumper, J. M., Evans, R. O., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Z˘´ıdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R. D., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D. L., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021).
13. Highly accurate protein structure prediction with alphafold. Nature
14. [12] Pal, M.K., Lahiri, T., Kumar, R.: ProtPCV: A fixed dimensional numerical representation of protein sequence to significantly reduce sequence search time. Interdisciplinary Sciences 12(3), 276–287 (2020). https://doi.org/10.1007/s12539- 020-00380-w.
15. [13] Kandathil, S. M., Greener, J. G., Lau, A. M., and Jones, D. T. (2020). Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments. bioRxiv.
16. [14] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. arXiv: Learning.
17. [15] Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B. O., and Marcatili, P. (2018). Netsurfp-2.0: improved prediction of protein structural features by integrated deep learning. Proteins.
18. [16] Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. empirical methods in natural language processing
19. [17] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020).
20. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.
21. [18] Lee, J. H., Yadollahpour, P., Watkins, A., Frey, N. C., Leaver-Fay, A., Ra, S., Cho, K., Gligorijevic, V., Regev, A., Bonneau, R., Design, P., and Genen- tech (2022). Equifold: Protein structure prediction with a novel

coarse-grained structure representation.
22. [19] Li, J. and Xu, J. (2021). Study of real-valued distance prediction for protein structure prediction with deep learning. Bioinformatics.
23. [20] Li, Z., Liu, X., Chen, W., Shen, F., Bi, H., Ke, G., and Zhang, L. (2022). Uni-fold:
24. An open-source platform for developing protein folding models beyond alphafold
25. [21] Lin, Z., Lanchantin, J., and Qi, Y. (2016). Must-cnn: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. national conference on artificial intelligence.
26. [22] Luo, J., Cai, Y., Wu, J., Cai, H., Yang, X., and Lin, Z. (2020). Self-supervised representation learning of protein tertiary structures (ptsrep): Protein engineering
27. as a case study. bioRxiv.
28. [23] Ma, B. (2015). Novor: real-time peptide de novo sequencing software. Journal of the American Society for Mass Spectrometry.
29. [24] Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B. O., and Marcatili, P. (2018). Netsurfp-2.0: improved prediction of protein structural features by integrated deep learning. Proteins.
30. [25] Ma, B. and Johnson, R. (2012). De novo sequencing and homology searching. Molecular & Cellular Proteomics.
31. [26] Ma, C., Dai, G., and Zhou, J. (2021). Short-term traffic flow prediction for urban road sections based on time series analysis and lstm bilstm method. IEEE Transactions on Intelligent Transportation Systems.
32. [27] Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., and Brock, O. (2015). Rbo aleph: leveraging novel information sources for protein structure prediction
33. [28] Nucleic Acids Research. Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser,
34. J. S., and Naik, N. (2021). Deep neural language modeling enables functional protein generation across families. bioRxiv
35. [29] Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. (2020). Progen: Language modeling for protein generation.
36. [30] Magnan, C. and Baldi, P. (2014). Sspro/accpro 5: almost perfect prediction of pro- tein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics.
37. [31] Olenyi, T. a. B., Michael and Mirdita, Milot and Steinegger, Martin and Rost, Burkhard. Rostclust – 12 Protein Redundancy Reduction (School of Computation, Information, and Technology, Technical 13 University of Munich., 2022). 14
38. [32] 123456789 Olenyi, T. a. B., Michael and Mirdita, Milot and Steinegger, Martin and Rost, Burkhard. Rostclust – 12 Protein Redundancy Reduction (School of Computation, Information, and Technology, Technical 13 University of Munich., 2022). 14
39. [33] Pereira, J. et al. High-accuracy protein structure prediction in CASP14. Proteins 89, 1687-1699 (2021). 18 https://doi.org:10.1002/prot.26171 19 [33.] Pereira, J. et al. High-accuracy protein structure prediction in CASP14. Proteins 89, 1687-1699
40.
41. (2021). 18 https://doi.org:10.1002/prot.26171 19
42. [34] Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text- to-Text Transformer. arXiv 20 (2020). 21
43. [35] OSteinegger, M., Mirdita, M. & S¨oding, J. Protein-level assembly increases pro- tein sequence recovery from 22 metagenomic samples manyfold. Nature Methods 16, 603-606 (2019). https://doi.org:10.1038/s41592- 23 019-0437-4 24
44. [36] Milesi, A. Accelerating SE(3)-Transformers Training Using an NVIDIA Open-
45. Source Model 28 Implementation. (2021). . 30
46. [37] Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular 31 modeling algorithms using Rosetta. Bioinformatics 26, 689-691 (2010). 32 https://doi.org:10.1093/bioinformatics/btq007 33
47. [38] Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. 34 Nucleic Acids Research 33, 2302-2309 (2005).
48. https://doi.org:10.1093/nar/gki524 35
49. [39] Harris, C. R. et al. Array programming with NumPy. Nature 585, 357-362 (2020). 36 https://doi.org:10.1038/s41586-020-2649-2
50. [40] Chan, Y. H., Venev, S. V., Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three 26 orthologous TIM barrels originates from sequence and structure constraints. Nature Communications 8, 27 14614 (2017).
51. https://doi.org:10.1038/ncomms14614 28
52. [41] Suiter, C. C. et al. Massively parallel variant characterization iden- tifies NUDT15 alleles associated 29 with thiopurine toxicity. Proceed- ings of the National Academy of Sciences 117, 5394-5401 (2020). 30 https://doi.org:doi:10.1073/pnas.1915680117
53. [42] Hunter, J. D. Matplotlib: A 2D Graphics

Environment. Computing in Science & Engineering 9, 90-95 32 (2007). https://doi.org:10.1109/MCSE.2007.55 33 36 Schr¨odinger, L. & DeLano, W. The PyMOL Molecular Graphics System, 34 (2021).
54. [43] Tomar, S. Converting video formats with FFmpeg. Linux Journal 2006, 10 (2006). 36 38 Williams, S. G. & Lovell, S. C. The Effect of Sequence Evolution on Pro- tein Structural Divergence. 37 Molecular Biology and Evolution 26, 1055-1065 (2009).https://doi.org:10.1093/molbev/msp020
55. [44] Lo Conte, L. et al. SCOP: a structural classification of proteins database. Nucleic Acids Res 28, 257-259 39 (2000). https://doi.org:10.1093/nar/28.1.257
56. [45] van Kempen, M. et al. Foldseek: fast and accurate pro- tein structure search. bioRxiv, 41 2022.2002.2007.479398 (2022). https://doi.org:10.1101/2022.02.07.479398
57. [46] Steinegger, M. & S¨oding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of 43 massive data sets. Nature Biotechnology 35, 1026-1028 (2017). https://doi.org:10.1038/nbt.3988
58. [47] Baek, M. et al. Accurate prediction of protein structures and interac- tions using a three-track neural 45 network. Science 373, 871-876 (2021). https://doi.org:doi:10.1126/science.abj8754
59. [48] Callaway, E. AlphaFold's new rival? Meta AI predicts shape of 600 million pro- teins. Nature 611, 211-212 47 (2022). 48 44 Liu, J., Montelione, G. T. & Rost,
60. B. Novel leverage of structural genomics. Nature Biotechnology 25, 849- 49 851
61. (2007).
62. 1. [49] Mishra. R. et. al. "Hybrid-ProtDeep: A Protein Structure Prediction" https://doi.org/10.1007/978-981-96-6303-3_36.