Biomedical Data Science
2025, VOL. 5, NO. 3, 212-217
DOI: 10.47297/wspbdsWSP2752-630544.20250503

Protein Structure Prediction: Methods and Advancements

Zhenyu Yang

School of Science, Xi'an Jiaotong-Liverpool University, Deyang, Sichuan, 61800 ,China

ABSTRACT

Protein structure prediction is pivotal for understanding biological functions and disease mechanisms. This review systematically examines mainstream computational approaches, including homology modeling, fold recognition, ab initio prediction, and AI-assisted techniques. Traditional methods such as homology modeling construct structures through template matching, while fold recognition leverages the limited fold universe for sequence-structure alignment. Ab initio prediction explores conformational space based on physical principles but faces computational complexity constraints. Recently, deep learning models represented by AlphaFold have achieved breakthroughs by integrating evolutionary constraints and 3D coordinate generation, with experimental-grade accuracy for single-domain proteins (e. g., AlphaFold2). AlphaFold3 further extends capabilities to multi-molecular complexes. Despite becoming the preferred choice for researchers, AI methods still exhibit deviations from physiological conformations. Future advancements require fundamental innovations integrating physical principles to enhance applications in drug design and gene editing.

KEYWORDS

Protein structure prediction; Homology modeling; Fold recognition; Ab initio prediction; Artificial intelligence; AlphaFold; Deep learning

1 Introduction

Protein is a fundamental component of all cells and tissues in the human body and serves as the basis material of life. Protein monomers are formed by the dehydration and condensation of amino acid monomers. And the specific arrangement and composition of amino acids determine the diverse functions of proteins.

Proteins generally have four hierarchical levels of structure. The primary structure of protein is the one-dimensional linear sequence of amino acid residues in the polypeptide chain. The secondary structure refers to the localized spatial conformations adopted by segments of the chain, such as α-helices and β-sheets, which governed by specific patterns of hydrogen bonding. The tertiary structure describes the overall three-dimensional folding of a single polypeptide chain, formed by secondary structures and side chains, which means that the tertiary structure is developed from the secondary structure, and using secondary structure to predict the structure is a main part for nowadays ' research. Finally, the quaternary structure arises when multiple polypeptide chains (subunits), each with its own tertiary structure, assemble into a functional complex through non-covalent interactions, means that the quaternary structure is the highest structure of protein.

Since proteins typically require a specific three-dimensional structure to perform their biological functions, figuring the amino acid sequence alone is insufficient to fully determine protein function. Thus, structural errors can render proteins dysfunctional or lead to aberrant activities of body. For instance, the neuropathology of Alzheimer's disease involves neurofibrillary tangles primarily composed of abnormally folded, highly phosphorylated Tau proteins [1].

Studying protein structure is crucial for understanding protein function and its role in physiological activities. Consequently, accurately predicting protein structure is essential for research in biology and medicine.

2 Methods of protein structure prediction

Driven by the rapid advancement of computer technology, numerous computational methods for protein structure prediction have emerged in recent years. These include homology modelling, fold recognition, and ab initio prediction. Furthermore, the rise of artificial intelligence has introduced powerful deep learning methods, such as AlphaFold and RoseTTAFold, for predicting protein structures [2].

Traditional experimental methods for structure determination (e.g., X-ray crystallography, NMR) can be costly and time-consuming. Computational prediction methods serve as valuable complements to these experimental approaches.

2.1 Homology modelling methods

Homology modelling, also known as comparative modelling, predicts the structure of a target protein based on its similarity to one or more experimentally determined structures of related proteins (templates).

2.1.1 Protein sequence similarity and homology

It is important to distinguish between protein homology and similarity. Homology implies an evolutionary relationship, signifying that two protein sequences share a common ancestor. Similarity is a quantitative measure (e. g., percentage identity) or qualitative description of sequence resemblance. High sequence similarity often suggests homology, but it is not definitive; convergent evolution can produce similar sequences without shared ancestry. Sequences inferred to share a common ancestor are termed homologous. Sequence alignment is the fundamental operation for comparing similarity, establishing residue-by-residue correspondences to identify conserved and divergent regions.

2.1.2 Template identification

The initial step involves searching for suitable template structures for the target protein sequence within a protein structure database, typically the Protein Data Bank (PDB). Search tools like FASTA, BLAST, or PSI-BLAST are commonly employed on subsets of the PDB. Selection criteria often include thresholds such as a FASTA score >10, a BLASTp E-value < 1e-5, and sequence identity covering at least 40% of the target sequence with a minimum similarity (e. g., 25%). Subsequently, a few top-scoring templates are selected. Their structures are optimally superimposed in 3D space, and their sequences are aligned to generate a structurally informed multiple sequence alignment. FASTA is generally suitable for detecting more distantly related sequences through pattern matching, while BLAST is widely used for local sequence alignments, particularly for nucleotide and amino acid sequences. The core difference lies in their underlying search algorithms and sensitivity [3].

2.1.3 Target-template sequence alignment

The alignment of the target sequence to the template sequence(s) is the most critical step in homology modelling, as it directly determines the accuracy of the final results. This alignment works by identifying structurally conserved regions (SCRs) and this process typically relies on multiple sequence alignment (MSA) techniques. Dynamic programming algorithms are commonly accepted in MSA which is an effective method for MSA when templates with strong homology are available. However, the accuracy diminishes for target sequences with low homology to known structures. In addition, since Profile provides an evolutionary map between sequences, the Profile-Profile alignment methods, generally yield more accurate results. However, when this method is insufficient to improve the accuracy of the results, only integrating multiple alignment strategies making them cover each shortage can improve the accuracy. For example, the EsyPred3D structure prediction server utilizes such an integrated approach [4]. Pairwise sequence alignment, a fundamental method, involves inserting gaps and substituting residues to maximize the similarity score between two sequences. A simple scoring scheme might be: +1 for identical residues aligned, 0 for mismatches, and -1 for gaps. Most homology search methods employ similar scoring principles or sophisticated variations based on dynamic programming.

2.1.4 Protein substitution matrices

Commonly used matrices for scoring amino acid substitutions include the BLOSUM (Blocks Substitution Matrix) and PAM (Percent Accepted Mutation) matrices.

* BLOSUM Matrix: Derived by analyzing the substitution frequencies within conserved blocks of aligned, related protein sequences. It does not explicitly use an evolutionary model but relies on observed frequencies from local alignments. Lower-numbered BLOSUM matrices (e.g., BLOSUM45) are suitable for comparing distantly related sequences, while higher-numbered matrices (e. g., BLOSUM80) are better for closely related sequences. The number (e. g., 62 in BLOSUM62) indicates the minimum percentage identity within the blocks used to construct the matrix.

* PAM Matrix: Based on an explicit model of evolutionary point mutations. One PAM unit represents an evolutionary distance where 1% of amino acids have changed. Lower-numbered PAM matrices (e.g., PAM30) are designed for closely related sequences, while higher-numbered matrices (e. g., PAM250) target more distant relationships. Compared to BLOSUM, higher-order PAM matrices can suffer from error accumulation due to matrix multiplication and may miss complex mutation patterns due to smaller underlying datasets.

2.1.5 Target-template structure alignment

Following template search and sequence alignment, the structural alignment of the target sequence to the template structure(s) is crucial for model accuracy. Structural information derived from alignments built using multiple templates generally yields higher confidence than alignments based on a single template. Key steps include:

(1) Generating a multiple sequence alignment (MSA) of the target and template sequences using tools like ClustalW or

MUSCLE.

(2) Performing structural superposition of the selected template structures based on the MSA to create a structural alignment of the templates.

(3) Obtaining high-confidence secondary structure predictions for the target sequence and refining the alignment to minimize gaps, especially within predicted secondary structure elements.

2.1.6 Profile-profile comparison method

This method enhances the detection of remote homologs by incorporating evolutionary information for both the target (query) sequence and the template sequences. It transforms position-specific amino acid frequency profiles (derived from MSAs) into new representations (e.g., log-odds profiles) and constructs a specialized scoring system for comparing these profiles. The core difference among profile-profile methods lies in how they calculate the similarity score between two profile positions (each position being a vector of amino acid frequencies). This new scoring system is then integrated with dynamic programming algorithms to align the query profile to the template profile(s). The accuracy of the final alignment can be assessed using benchmarks like HOMSTRAD [5].

2.1.7 Data smoothing techniques (context in profile methods)

Data smoothing is a technique borrowed from natural language processing (NLP) to address the "data sparsity" problem, a special condition which leads to zero probabilities in statistical models making the data looks sparse. The mechanism of this technique is to adjusts probability distributions derived from maximum likelihood estimation (MLE) on training data to ensure no event has zero probability, resulting in a smoother distribution.

In N-gram language models, the MLE for the conditional probability of a word (w_i)given its context ($w_{i=n+1}^{i=1}$)is:

$$[p_{\mathrm{MLE}}\left(w_i|w_{i=n+1}^{i=1}\right)=\frac{c\left(w_{i=n+1}^{i}\right)}{\sum_{w} c\left(w_{i=n+1}^{i}\right)}]$$

where ($\mathrm{c}\left(\mathrm{w}_{\mathrm{i=n+1}}^{\mathrm{i}}\right)$) is the count of the n-gram ($w_{i=n+1}\cdots w_i$) in the training corpus (T).

Data sparsity occurs when ($\mathrm{c}\left(w_{i-n+1}^{i}\right)=0$), assigning zero probability. This problem worsens with model complexity. Simply increasing corpus size has limited effectiveness.

Smoothing performance is often evaluated using cross-entropy (H_m) or perplexity(PP_m) on a held-out test set:

$$[H_m=-\frac{1}{N_T}\sum_{i=1}^{|T|}\log_2 p_m\left(t_i\right)]$$

$$[\mathrm{P}P_m=2^{H_m}]$$

where (N_T) is the total word count in test text (T), ($|T|$) is the number of sentences, (p_m) is the probability assigned by the smoothed model (m), and (t_i) is the (i)-th test sentence. Lower cross-entropy/perplexity indicates better smoothing.

Common techniques relevant to profile methods include:

Additive (Laplace/Lidstone) Smoothing:Adds a constant (δ) ($(0<\delta\leq 1)$) to all n-gram counts (including unseen ones).

$$[p_{\mathrm{add}}\left(w_i|w_{i=n+1}^{i=1}\right)=\frac{c\left(w_{i=n+1}^{i}\right)+\delta}{\sum_{w} c\left(w_{i=n+1}^{i}\right)+\delta|V|}]$$

where ($|V|$) is the vocabulary size.

Good-Turing Smoothing:Estimates the frequency (r^*) of n-grams seen (r) times using the frequency of n-grams seen (r + 1) times:

$$[r^*=(r+1)\frac{n_{r+1}}{n_r}]$$

where (n_r) is the number of n-grams occurring (r) times. The probability for an n-gram occurring (r) times is:

$$[P_{\mathrm{GT}}\left(w_{i=n+1}^{i}\right)=\frac{r^*}{\sum_{k=0}^{\infty} r^*}]$$

(Adjustments are needed for unseen events).

In protein structure prediction within profile-profile methods, data smoothing is applied under the assumption that the 20 amino acids are independent. The frequency profile (a 20-dimensional vector per position) is smoothed using techniques like Additive or Good-Turing to generate a desired frequency profile for subsequent scoring, setting the model's parameter space size to 20.

2.2 Fold recognition (threading) method

Fold recognition predicts the structure of a target protein by matching its sequence against a library of known protein folds (structural templates), identifying the best-fitting fold(s). This method leverages the observation that the number of distinct protein folds in nature is limited. It bypasses the direct prediction of secondary structure, which historically had limited accuracy (~65%), and focuses on matching sequence to tertiary fold.

2.2.1 Principle and development

The concept stems from the recognition of a limited "fold space." Finkelstein and Ptitsyn (1987) proposed stereochemical constraints limit fold diversity. Chothia (1992) estimated fewer than 1000 folds, and Wang Zhixin (1998) proposed a more precise estimate of around 654 folds [6]. Fold recognition uses known fold templates to identify the probable fold type of a target sequence.

2.2.2 Key methodologies evolved from

(1) 3D-Profile Method (Bowie et al., 1991): Classifies local structural environments in known proteins (e.g., based on solvent accessibility, polarity, secondary structure). It calculates the propensity of each amino acid for each environment type, creating a 3D-1D scoring table. A 3D-profile is generated for each template fold. The target sequence is scored against all template profiles in the library; the best-matching profile(s) predict the fold.

(2) Threading Method (Jones et al., 1992): Establishes a fold library (sub-database). The target sequence is "threaded" through each template structure in the library in all possible alignments. A scoring function evaluates the compatibility of the target sequence with each template structure. The template(s) with the highest score(s) represent the predicted fold.

With established fold libraries, fold recognition has matured into a robust prediction method and transitioned from theoretical research to practical application.

2.3 Ab initio prediction

Ab initio prediction aims to predict protein structure solely from the amino acid sequence and physical principles, without relying on explicit templates thus, it is the most ideal method to predict protein structure. However, currently this method faces significant challenges in accuracy and computational cost it is far from satisfaction. Broadly speaking ab initio prediction encompasses predicting secondary structure, supersecondary motifs, tertiary structure, and protein structural class, most of them are based on templates, secondary structures and other information to complete predictions. Compared to homology modelling, ab initio methods are essential for proteins lacking detectable homologs (<25% sequence identity), making them crucial for protein design and novel protein studies.

2.3.1 General steps involve

(1) Energy Function: Defining a potential energy function capable of distinguishing the native conformation from non-native ones.

(2) Conformational Search: Exploring the vast conformational space to find low-energy states. Search strategies include:

Systematic Search (e.g., Lattice Models): Systematically explores conformational space. Computationally intractable for large proteins due to exponential scaling with chain length.

Stochastic Search (e.g., Genetic Algorithms, Monte Carlo, Simulated Annealing): Samples conformational space more efficiently than systematic methods but offers no guarantee of finding the global minimum within feasible time [7]. These methods navigate the search using the energy function and move sets.

A major challenge for ab initio prediction is simulating the protein folding pathway via molecular dynamics based on atomic physics, which requires immense computational power. Thus, it is impractical for predicting natural folds accurately for most proteins [8]. Consequently, ab initio methods are primarily applicable to small proteins (<100-150 residues) nowadays.

2.3.2 Fragment-based assembly

Fragment-based assembly is among the most successful ab initio approaches. It relies on the observation that short sequence fragments (e.g., 3-9 residues) tend to adopt local structures similar to those observed in known proteins, even without global sequence homology, thus dividing the target sequence into overlapping fragments basing on these fragments can restrict the size of conformational space needed to search. Notable methods like ROSETTA and TASSER are built upon this principle [7].

2.4 Artificial intelligence-assisted protein structure prediction

2.4.1 Deep learning background

The rapid advancement of computer science has enabled significant progress in machine learning. Specifically, deep learning—an approach developed within machine learning—aims to achieve unstructured data learning autonomously, whereas traditional machine learning typically requires explicit commands. Consequently, deep learning needs to minimize human intervention as much as possible. Nowadays this deep learning has been widely adopted in fields such as image processing and speech recognition

2.4.2 Alphafold breakthrough

Following the success of AlphaGo, DeepMind applied AI to biology, specifically protein structure prediction. The resulting AI system, AlphaFold, achieved unprecedented accuracy:

AlphaFold2: Its prediction pipeline involves:

(1) MSA Construction: Searches sequence databases for homologs of the target, building a deep Multiple Sequence Alignment (MSA) to capture evolutionary constraints.

(2) Pair Representation: Computes a matrix representing the likelihood of residue pairs being in spatial proximity, often derived from correlated mutations within the MSA and structural database queries.

(3) Structure Module: A sophisticated deep neural network (Evoformer + Structure Module) processes the MSA and pair representations through multiple iterations (recycling) to generate atomic-level 3D coordinates, satisfying both evolutionary and spatial constraints.

Trained on tens of thousands of structures from the PDB, AlphaFold2 achieved accuracy comparable to medium-resolution experimental methods for single-domain proteins, marking a major breakthrough in a decades-old problem [9]. It dramatically reduces the cost of obtaining 3D models but has limitations, such as predicting disordered protein regions poorly [10]. Its speed (hours) makes it invaluable for drug discovery and research into disease mechanisms (e.g., Parkinson's, Alzheimer's) [11].

AlphaFold3 (Launched May 2024): Significantly enhances accuracy and expands capabilities beyond single proteins to predict complexes involving proteins, nucleic acids (DNA/RNA), small molecules, ions, and antibody-antigen interactions, outperforming other specialized tools [12].

3 Summary and outlook

Predominant computational methods for protein structure prediction include homology modelling, fold recognition, and ab initio prediction. However, the advent of AI, exemplified by platforms like AlphaFold, has dramatically shifted the landscape, becoming the preferred choice for researchers due to its remarkable accuracy and speed. AI-predicted structures are already enabling applications such as designing proteins that bind specific targets (e.g., heart drug digoxin, enzyme cofactor heme, light-capturing bilin) using tools derived from these methods (e. g., RFdiffusionAA based on RoseTTAFold All-Atom) [13]. Deep learning-based prediction holds immense potential for drug design, gene editing, and synthetic biology [7].

Despite significant progress, challenges remain. Models introduced after AlphaFold2 (including AlphaFold3) often refine existing concepts (e.g., evolutionary information extraction, incorporation of physical constraints, multi-state/multi-molecule prediction) rather than introducing radically new paradigms, leading to incremental rather than revolutionary accuracy gains for core protein structure prediction [14]. A persistent issue is the deviation of predicted models from the true, functional native state under physiological conditions, limiting their value in certain applications like detailed drug docking or understanding conformational dynamics [14-15]. This indicates that protein structure prediction, while transformed by AI, faces new bottlenecks requiring fundamental innovations.

In general, commonly used protein structure prediction methods include homology modeling, folding recognition, and de novo prediction. Among these, AI-driven platforms like AlphaFold have become researchers' primary choice due to advancements in computer science. For example, RFdiffusionAA—a model fine-tuned from RoseTTAFold All-Atom (derived from AlphaFold)—validated proteins binding to the heart disease drug digoxin, the enzymatic cofactor heme, and the light-trapping molecule bilin [13]. Furthermore, deep learning-based protein structure prediction is applicable to drug design, gene editing, and synthetic biology [8].

However, although AlphaFold2 has released new prediction models, its technical framework remains fundamentally unchanged from its predecessor, primarily refining evolutionary information extraction. Consequently, its accuracy has plateaued, highlighting persistent prediction bottlenecks. Additionally, AI models like AlphaFold2 exhibit deviations between predicted structures and natural conformations, limiting their utility in specific applications [14-15].

Nevertheless, compared to traditional computational methods, AI approaches remain researchers' preferred choice due to their unparalleled accuracy. Future protein structure prediction systems may likely integrate physical principles to enhance accuracy, mirroring how AlphaFold3 incorporates physical factors to rank target antibody structures during prediction.

References

[1] Wang F, Li HJ, Li HY. Exploration of methods for protein structure prediction[J]. Modern Information Technology, 2022, 6(18): 122-125.

[2] Liu ZN, Lai HS, Song LY. An overview of protein structure prediction [J]. Chinese Journal of Medical Physics, 2020, 37(9): 1203-1207.

[3] Zhang, C. et al. (2024) 'The Historical Evolution and Significance of Multiple Sequence Alignment in Molecular Structure and Function Prediction', Biomolecules, 14(12), p. 1531.

[4] ZENG Bingjia, CAO Yicheng, DU Zhengping, et al. Dynamics of key steps in homology modelling[J]. Journal of Biology, 2008, 25(2): 7-10.

[5] Chen HM, Zhou JX. Research on protein structure prediction method based on homology modelling[J]. Henan Science, 2009, 27(9): 1108-1110.

[6] Yin C.C.. Research progress of protein structure prediction methods[J]. Computer Engineering and Applications, 2004, 40(20): 54-57.

[7] ZHOU Jianhong, AI Guanhua, FANG Huisheng, et al. Progress of *ab initio* protein structure prediction methods[J]. Bioinformatics, 2011, 9(1): 1-5.

[8] LI Ming, SU Xianzhong, YU Min, et al. Progress in protein structure prediction[J]. Biotechnology, 2009, 19(3): 87-90.

[9] WANG Tianyao, LI Jianfeng. Application and insights of deep learning in protein structure prediction[J]. Journal of Polymer Science, 2022, 53 (6): 581-591.

[10] Guo Beiyi, Guo Xiaoqiang. AlphaFold and protein structure prediction[J]. Science (Shanghai), 2024, 76(5): 39-44.

[11] Guo Fuhu. Nobel Prize in Chemistry 2024: A new era of computational protein design and protein structure prediction[J]. Journal of Northwest Normal University (Natural Science Edition), 2024, 60(6): 119-121.

[12] Wu Y. AlphaFold 3 will aid drug discovery[J]. Nature Magazine, 2024, 46(3): 184.

[13] Zhao L. Generalized biomolecular modelling and design with RoseTTAFold All-Atom[J]. Journal of Guangdong Pharmaceutical University, 2024, 40(2): 90.

[14] Zhou, Yaoqi. Post-AlphaFold 2 era: Progress continues, but a new revolution is still far away[J]. World Scientific, 2024(8): 15.

[15] ZHANG Jinlong, ZHAO Kelong, LIU Dong, et al. A protein model refinement method based on multi-structure deep learning[J]. Small Microcomputer Systems, 2024, 45(7): 1577-1584.